Distribution Agreement

Signature:

____
Rohit Allena

____
Date

THREE ESSAYS ON ESTIMATION UNCERTAINTY

By

Rohit Allena
Doctor of Philosophy

Business

Jay A. Shanken, Ph.D.
Committee Chair

Tarun Chordia, Ph.D.
Committee Co-chair

William Mann, Ph.D.
Committee Member

Jegadeesh Narasimhan, Ph.D.
Committee Member

Donald Lee, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

THREE ESSAYS ON ESTIMATION UNCERTAINTY

By

Rohit Allena
Master in Statistics, Indian Statistical Institute, 2014
Bachelor in Statistics, Indian Statistical Institute, 2012

Dissertation Chair: Jay A. Shanken, Ph.D., Carnegie-Mellon University, 1983
Dissertation Co-chair: Tarun Chordia, Ph.D., University of California Los Angeles, 1993

An abstract of a dissertation submitted
to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Business

2021

Abstract

THREE ESSAYS ON ESTIMATION UNCERTAINTY

By

Rohit Allena

The dissertation consists of three essays on estimation uncertainty, showing why and how considering estimation uncertainty is important in answering three fundamental asset pricing and market microstructure questions.

The first essay (Confident Risk Premia: Economics and Econometrics of Machine Learning Uncertainties) quantifies ex-ante **parameter uncertainty** of expected stock return predictions from neural networks by deriving their standard errors or confidence intervals. Considering ex-ante standard errors, the paper provides: 1) improved trading strategies known as Confident-high-low portfolios (in contrast to traditional high-low strategies), and 2) ex-post out-of-sample (OOS) inferences by generalizing Diebold-Mariano t-tests to statistically compare OOS returns and Sharpe ratios of any two trading strategies.

The second essay (Comparing Asset Pricing Models with Non-traded Factors and Principal Components) develops a Bayesian methodology to compare asset pricing models containing non-traded factors and principal components. Existing comparison procedures are inadequate when models include such factors due to **estimation uncertainties** in mimicking portfolios and return covariances. Furthermore, regressions of test assets on such factors are interdependent, rendering comparisons with recently proposed priors sensitive to subsets of the test assets. Thus, the paper derives novel, non-informative priors that deliver invariant inferences. The paper finds that macroeconomic factor models dominate several recent benchmark models with traded factors and principal components.

The third essay (True Liquidity and Fundamental Prices: US Tick Size Pilot) is joint work with Tarun Chordia. This paper develops a big-data methodology to estimate fundamental prices and true liquidity measures, explicitly considering the rounding specification (**estimation uncertainty**) due to the minimum tick size. Evaluation of the tick size pilot (TSP), which increased the tick size for some randomly chosen stocks, requires estimating the impact of rounding. True liquidity measures capture the TSP-driven decreased inventory costs of market-makers, whereas traditional measures without the rounding adjustment cannot. We find that the TSP increases market-maker profits, but does not improve liquidity and price efficiency. This result contrasts with existing empirical studies but is consistent with recent theoretical studies that account for rounding.

THREE ESSAYS ON ESTIMATION UNCERTAINTY

By

Rohit Allena
Master in Statistics, Indian Statistical Institute, 2014
Bachelor in Statistics, Indian Statistical Institute, 2012

Dissertation Chair: Jay A. Shanken, Ph.D., Carnegie-Mellon University, 1983
Dissertation Co-chair: Tarun Chordia, Ph.D., University of California Los Angeles, 1993

A dissertation submitted
to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Business

2021

Dedication

To my family and well-wishers.

Acknowledgements

I am grateful to Jay Shanken (committee chair) and Tarun Chordia (committee co-chair) for their mentorship and invaluable suggestions. Their deep research and excellent teaching of asset pricing concepts are pivotal in developing my dissertation. In addition, I thank my other committee members: William Mann, Donald Lee, and, in particular, Jegadeesh Narasimhan, for posing challenging questions, which significantly improved my dissertation papers and helped me in the 2021 job market.

I would also like to thank the PhD administration staff at Emory: Putnam Katherine and in particular Allison Gilmore, and my PhD colleagues: Chandra Sekhar Mangipudi, Shikha Jaiswal, Badrinath Kottimukkalur, Dexin Zhou, Cong Roman Wang, Joonki Noh, Ai He, Abinash Pati, Jaemin Lee, Jinoug Jeung, and Jianzhang Lin, for helpful discussions and support throughout my Ph.D. journey.

Also, I benefited from seminar participants at the business schools of Boston College, Boston University, Copenhagen, HEC Paris, HKUST, Indian School of Business, National University of Singapore, Tulane University, Universities of Florida, Georgia, Houston, Yale University, Western (European and Northern) Finance Association annual meetings, the University of Chicago SoFiE summer school, and SoFiE annual meetings. Special thanks to Bryan Kelly and Jonathan Lewellen for their insightful feedback.

Contents

1 Introduction **2**

A Contribution . . . 9

B Paper Overview . . . 10

2 Risk Premium Predictions and Predictive Standard Errors **10**

A Basics of model-based risk premium predictions . . . 11

B Risk Premium Predictions, Standard Errors and Investment Portfolios . . . 13

3 NN-based Risk Premia and Standard Errors **16**

A Neural Networks . . . 17

B Parameter Estimation, Regularization, and Dropout . . . 19

C Standard Errors of Risk Premium Predictions based on Neural Networks . . . 23

D Dropout Neural Networks and Bayesian Interpretation . . . 25

E Frequentist Justification for Standard Errors . . . 30

4 Ex-ante Estimation Uncertainty and Ex-post OOS Inferences **31**

A Out-of-Sample Comparisons with the Diebold and Mariano (2002) Tests . . . 31

B Violation of Covariance Stationarity: Empirical Evidence . . . 33

C Bootstrap Tests for Out-of-Sample Comparisons . . . 34

D Performance of the Methodology: Monte Carlo Evidence . . . 37

5 Empirical Results **37**

A Data, Definitions, and Replication Study . . . 37

B Ex-ante Confidence and Ex-post Out-of-Sample-R^2 . . . 40

C Portfolio Construction . . . 42

D Economic Gains from Confident-HL Portfolios . . . 44

E Reassessing NN-3 and Lewellen Model Comparisons Using Bootstrap Tests . . . 47

F Time-Series Variation in Ex-ante Standard Errors . . . 50

G Cross-sectional Variation in Ex-ante Confidence . . . 51

6 Conclusions **52**

A Appendix: Proofs **53**

1 Proof of Proposition-1: . . . 53

2 Proof of Proposition-2 . . . 55

3 Proof of Proposition-3 . . . 56

B Appendix: Simulations and Testing the Diebold and Mariano (2002) Assumption 58

1 Validity of Standard Errors: Monte Carlo Evidence . . . 58

2 Tests of Covariance Stationarity . . . 59

3 Performance of this paper's OOS Comparison Method: Monte Carlo Evidence . . . 60

C Internet Appendix 83

C1 Internet Appendix: Simulation Results and Robustness Checks . . . 83

C2 Internet Appendix: Simulation Details . . . 86

C3 Why Confidence-levels are Better Measures of Precision Relative to Inverse Standard Errors . . . 88

Comparing Asset Pricing Models with Non-traded Factors and Principal Components

Contents

7 Introduction **91**

8 Asset Pricing Models with Non-Traded Factors **99**

9 General Test of a K-Factor Model **101**

9a Prior specification 103

9b Bayes Factor 108

10 Comparing Asset Pricing Models with Non-Traded Factors **112**

10a Priors for Comparing Asset Pricing Models 115

11 Comparing Models with Principal Components **121**

11a Priors for Comparing Asset Pricing Models with Principal Components 123

12 Marginal Likelihoods and Out-of-Sample Predictions **124**

13 Simulation Evidence **126**

14 Empirical Comparison of Prominent Asset Pricing Models **130**

14a Models with Traded vs Non-Traded Factors 131

14b Are these Non-Traded Factors Spurious? 134

14c Traded versus Non-Traded versus Principal Components 135

15 Conclusion **138**

16 Appendix **140**

16a Bayes Factors for Absolute Tests 140

16b Invariance to the Choice of Test Assets and Scaling of Mimicking Portfolio Weights . 143

16c Bayes Factors for Model Comparisons 144

16d Model Comparisons with Principal Components 147

Contents

17 Introduction **162**

18 Model **170**

19 Methodology: Variational Bayesian Inference **175**

19a Family of Densities for Approximation . . . 178

19b Estimating the Optimal Density Function . . . 180

19c Derivations of Updates . . . 181

20 Performance of the Proposed Methodology **182**

20a Accuracy of the methodology: Monte-Carlo Evidence . . . 182

21 Data **184**

22 Quoted Spreads, True Spreads and Market-Makers' Profits **184**

22a True Spreads and Its Components . . . 190

23 Effective Spreads **192**

24 Price Discovery **193**

24a Proportion of Price Discovery through Trading and New Information . . . 193

24b Speed of Price Discovery . . . 196

25 Conclusion **198**

A Appendix **199**

B Internet Appendix **218**

List of Tables

1 Calibration of the Confidence Intervals: Monte Carlo Evidence 58
2 Violation of Diebold and Mariano (2002) conditions : Non-Stationarities due to Estimation Uncertainty 59
3 Long-short Portfolios' Performance on Subsamples with Different Levels of Ex-ante Confidence 70
4 Performance of Confident and Low-Confident Long-Short Portfolios: All Stocks . . . 71
5 Statistical Comparison of Long-Short Portfolios: All Stocks 72
6 Performance of Confident and Low-Confident Long-Short Portfolios: Non-Microcap Stocks 73
7 Statistical Comparison of Long-Short Portfolios: Non-Microcap Stocks 74
8 Transaction Costs and Higher-Moment Adjusted Performance of Confident-HL Portfolios 75
9 Statistical Comparison of Long-Short Portfolios: NN-3 versus Lewellen (2015) 78
10 Aggregate Standard Errors of NN-3-based Risk Premia 80
11 Cross-sectional Characteristics of Confidence-sorted Deciles 81
12 Characteristics Distributions of Stocks in the Decile Containing the Most Confident Risk Premium Predictions 82
A Performance of High-Low and Confident High-Low Portfolios: Simulation Evidence . 83
B Performance of Various Long-Short Portfolios: Inverse Standard Errors as Precision Measures 84
C Comparing Confident-HL Portfolios with Double-sorted HL Portfolios 85

List of Tables

13 Performance of the Proposed Methodology: Simulation Evidence 153
14 Performance of the Proposed Methodology: Simulation Evidence 154
15 Traded vs Non-Traded: Model Probabilities with 52 Test Assets of Kozak, Nagel, and Santosh (2019) 155
16 Traded vs Non-Traded: Model Probabilities with 52 Anomalies + 10 Industry portfolios156
17 Correlations of Macroeconomic Factors with the Cross-Section of Stock Returns . . . 157
18 Traded vs Non-Traded : Model Probabilities with Spurious Factors 158
19 Traded vs Non-Traded : Model Probabilities with Spurious Factors 159
20 Traded versus Non-Traded versus Principal Components: Posterior Probabilities with 52 Anomalies 160
21 Traded vs Non-Traded vs Principal Components: Posterior Probabilities with 52 Anomalies + 10 Industry portfolios 161

List of Tables

31 Performance of the Methodology: Monte Carlo Evidence . . . 209
32 Quoted Spreads, True Spreads and Market-Maker Profits . . . 210
33 Aggregated Dollar Value of Quoted Spreads, True Spreads and Market-Maker Profits 211
34 Realized Market-Maker Profits . . . 212
35 Components of Bid-Ask Spread . . . 213
36 Inventory Risks of Aggregate Market-Makers . . . 214
37 Effective Spreads with Fundamental Prices, mid-Quotes, and Weighted Mid-quotes . 215
38 Proportion of Price Discovery through Market Orders, Limit Orders and New Information . . . 216
39 Speed of Price Discovery with Fundamental Prices . . . 217
40 Market-maker Profits and Depths . . . 218

List of Figures

1 Example of a 1-layer Neural Network 17
2 NN-1 with Dropout Regularization 21
3 Test Sizes of OOS Comparison Methodologies 60
4 Power Curves of OOS Comparison Methodologies 61
5 Out-of-Sample (OOS) Performance of Equal-weighted Deciles Based on NN-3 Predictions. 66
6 Out-of-Sample (OOS) Performance of Value-weighted Deciles Based on NN-3 Predictions. 66
7 Ex-ante Confidence and Ex-post OOS-R^2: NN-3-based Predictions and Standard Errors 68
8 Ex-ante Confidence and Ex-post OOS-R^2: Lewellen-based Predictions and Standard Errors 69
9 Comparing predictive performance of NN-3 with the benchmark Lewellen (2015) model 76
10 Comparing predictive performance of NN-3 with the benchmark Lewellen (2015) model 77
11 Time-Series Variation in Standard Errors of NN-based Risk Premia 79

List of Figures

12 ELBO (y-axis) vs Number of Epochs (x-axis) 208

Confident Risk Premia: Economics and Econometrics of Machine Learning Uncertainties

Rohit Allena [*]
Goizueta Business School
Emory University

March 31, 2021

Abstract

This paper derives ex-ante standard errors of risk premium predictions from neural networks (NNs). Considering standard errors, I provide improved investment strategies and ex-post out-of-sample (OOS) statistical inferences relative to existing literature. The equal-weighted (value-weighted) confident high-low strategy that takes long-short positions exclusively on stocks that have precise risk premia earns an OOS average monthly return of 3.61% (2.21%). In contrast, the conventional high-low portfolio yields 2.52% (1.48%). Existing OOS inferences do not account for ex-ante estimation uncertainty and thus are not adequate to statistically compare the OOS returns, Sharpe ratios and mean squared errors of competing trading strategies and return prediction models (e.g., linear, NN and random forest). I develop a bootstrap procedure that delivers robust OOS inferences. The bootstrap tests reveal that large OOS return and Sharpe ratio differences between NN and benchmark linear models' traditional high-low portfolios are statistically insignificant. However, the NN-based confident high-low portfolios significantly outperform all competing strategies. Economically, standard errors reflect time-varying market uncertainty and spike after financial shocks. In the cross-section, the level and precision of risk premia are correlated, thus NN-based investments deliver more gains in the long positions.

Keywords: Machine Learning, Neural Networks, Standard Errors, Risk Premium, Novel Investment Strategies, Robust Out-of-Sample Inferences, Average Return Comparisons, Sharpe Ratio Comparisons, Machine Learning Uncertainties

[*] I am grateful to Jay Shanken (committee chair) and Tarun Chordia (committee co-chair) for their mentorship and invaluable suggestions. I benefited from discussions with my other committee members and faculty at Emory University, including William Mann, Jegadeesh Narasimhan, Gonzalo Maturano, Christoph Herpfer, and Donald Lee, as well as seminar participants at the business schools of Boston College, Boston University, Copenhagen, HEC Paris, HKUST, Indian School of Business, National University of Singapore, Tulane University, Universities of Florida, Georgia, Houston, and Yale University. Special thanks to Bryan Kelly and Jonathan Lewellen for their insightful feedback. All errors are mine.

Address: 1300 Clifton Road, Atlanta, 30322, E-mail Address: rohit.allena@emory.edu

1. Introduction

Modern empirical asset pricing literature applies machine learning (ML) models to estimate asset risk premia (i.e., expected returns in excess of the risk free rate), as these models can accommodate non-linear relations amongst a high-dimensional set of predictors. In an influential work, Gu, Kelly, and Xiu (2020) (GKX) examine various ML models, such as neural networks (NNs) and random forests, to predict individual stock's monthly risk premia. They argue that NNs statistically outperform the benchmark linear models examined by Lewellen (2015) (henceforth Lewellen) in predicting stock risk premia.[1]

However, the burgeoning ML literature has not ascertained the ex-ante precision (i.e., standard errors and confidence intervals) of risk premium predictions from NNs. Fama and French (1997) and Pástor and Stambaugh (1999) show that expected return estimates from traditional factor-based models are unavoidably imprecise due to uncertainty about unknown parameters, including asset exposures to factors (betas) and factor premia (gammas). Consequently, they argue that factor-based risk premium measurements are not suitable for making cost-of-equity capital decisions. Given that NNs entail a massive number of parameters, determining the precision of NN-based risk premia is important.

This paper develops a novel and easy-to-implement procedure to estimate predictive standard errors of NN-based risk premium predictions at both the stock-level and portfolio-level (e.g., industry portfolios). These *ex-ante* measures capture estimation uncertainty related to risk premium predictions. Whereas standard errors of traditional, linear, factor-based and characteristics-based risk premium estimates are available in the literature, those of highly complex, NN-based risk premia are not. I tackle this challenge by adapting the NNs of GKX to simultaneously deliver risk premium predictions and their standard errors every period. The predictive standard errors resemble classical bootstrap-based estimators but are available in real-time with no additional computation costs. The obtained standard errors are then theoretically justified, and empirically

[1] Bianchi, Büchner, and Tamoni (2020) and Bali, Goyal, Huang, Jiang, and Wen (2020) employ NNs to estimate bond and corporate bond risk premia, respectively.

validated using Monte-Carlo simulations.

Importantly, I present novel insights demonstrating why and how ex-ante standard errors must be explicitly considered to address two core asset pricing problems that appear in virtually every study in the burgeoning ML literature: (i) forming long-short trading portfolios using NN-based or any ML-based risk premium predictions and (ii) statistically evaluating the *ex-post* out-of-sample (OOS) performance of any model-based risk premia and corresponding trading strategies.[2] Considering ex-ante standard errors in answering both of these questions is of fundamental importance and has not been established in the literature.

Ex-ante standard errors provide investment gains. Many researchers (e.g. GKX and Avramov, Cheng, and Metzker (2020)) sort stocks into deciles based solely on their return predictions, and they take long-short positions on the extreme predicted-return deciles. This paper provides substantial enhancements to these conventional high-low (HL) investment strategies by exploiting the cross-sectional variation in the ex-ante precision of risk premia. I introduce novel "Confident-HL" trading portfolios that exclusively take long-short positions on a subset of stocks in the extreme predicted-return deciles that have more confident risk premia (i.e., high absolute ratios of risk premium predictions and their standard errors, or absolute t-ratios).[3] These strategies deliberately exclude stocks with relatively imprecise risk premium estimates and thus deliver large OOS average return and Sharpe ratio improvements.

Ex-ante standard errors impact ex-post OOS statistical inferences. To compare the ex-post OOS performance of these HL trading strategies or, any competing return prediction models or associated investment portfolios, researchers use two approaches: (i) reporting *point estimates* of models' OOS R^2s (OOS-R^2s) and investment portfolios' OOS average returns and Sharpe ratios (e.g., Chen, Pelger, and Zhu (2020)) or (ii) conducting simple t-tests motivated by Diebold and Mariano (2002) (henceforth DM) (e.g., GKX, Bianchi et al. (2020), Avramov et al. (2020), and

[2] The standard errors also impact the cost of capital decision-making with NN-based risk premia. In the spirit of Fama and French (1997) and Pástor and Stambaugh (1999), Allena (2020b) separately addresses this question.

[3] I measure the precision of risk premium predictions using their confidence-levels (i.e., absolute t-stats). See section C.C3 for an analytical motivation. Alternatively, I also present results using the inverse standard errors as proxies for the precision, and my conclusions are the same.

Bali et al. (2020)).[4] I show that these ex-post OOS inferences are inadequate because they do not account for ex-ante standard errors (i.e., estimation uncertainty).[5]

This paper presents a bootstrap procedure, robust to ex-ante estimation uncertainty, for valid statistical comparisons of any two portfolios' ex-post OOS returns and Sharpe ratios. Likewise, the method also compares the predictive performance of any two competing return prediction models (e.g., linear, random forests and NNs). Simulations suggest that whereas the 5%-level bootstrap tests yield accurate sizes close to 5%, the DM tests deliver distorted sizes between 13% and 42%, depending on the degree of estimation uncertainty.

Importantly, the bootstrap tests reveal that existing inferences with the DM tests over-reject the benchmark Lewellen model in favor of NNs. I find that the difference between both models' conventional HL portfolios' OOS returns and Sharpe ratios are either moderately significant or statistically insignificant. However, NNs exceptionally outperform on subsamples of stocks that have *confident* NN-based risk premia. Likewise, NN-based Confident-HL portfolios, which exclude stocks with relatively imprecise risk premia, statistically outperform all other competing strategies. Thus, considering ex-ante standard errors of NN-based risk premia is necessary for both real-time trading strategies and ex-post OOS inferences. Although this paper focuses primarily on NNs because of their predominance, I emphasize that the arguments hold for all ML-based risk premia.

I begin by showing that ex-ante standard errors of NN-based, or any ML-based, risk premium predictions predict their (future) squared forecast errors and thus yield large economic gains.[6] For example, when the standard errors of specific stock risk premium predictions are large, so are their squared forecast errors. This result is due to the "bias-variance" tradeoff. Expected squared forecast errors equal the sum of ex-ante "variances" and squared "biases". Whereas bias represents model misspecification, variance quantifies estimation uncertainty. Because predictions from ML models entail flexible functions involving many parameters, variances rather than biases predominantly determine their squared forecast errors. As a consequence, I establish that the Confident-HL

[4] Using simulations, Allena (2020a) shows that inferences based only on OOS point estimates are highly misleading.

[5] Diebold (2015) and GKX emphasize that the DM tests are not suitable for comparing model-based forecasts with estimation uncertainty. GKX acknowledge this limitation and conduct the DM tests. I illustrate why and how to account for parameter uncertainty to obtain accurately sized tests.

[6] Forecast errors equal the differences between true and predicted risk premia.

portfolios that deliberately drop stocks with imprecise risk premia earn superior expected returns.

A simple example provides the central intuition. Consider two stocks A and B with risk premia μ_A and μ_B, respectively. Let $\hat{\mu}_A$ and $\hat{\mu}_B$ be their risk premium predictions, which are normal, uncorrelated and unbiased, with the measurement error variance σ^2. The unbiased assumption suits ML-based predictions. Then the expected OOS return of the HL strategy that takes a long (short) position on the stock with the highest (lowest) risk premium prediction equals

$$E(HL) = (\mu_A - \mu_B) P(\hat{\mu}_A > \hat{\mu}_B) + (\mu_B - \mu_A) P(\hat{\mu}_B > \hat{\mu}_A) = (\mu_A - \mu_B) \left[2\Phi\left(\frac{\mu_A - \mu_B}{\sqrt{2}\sigma} \right) - 1 \right], \quad (1)$$

where $P(.)$, $\Phi(.)$ denote the probability and standard normal distribution measures, respectively. (1) indicates that the expected HL return monotonically decreases with the variance of risk premium predictions. In other words, between any two sets of stocks with the same levels of risk premia, the HL strategy formed from more precise predictions yields higher OOS expected returns. Intuitively, besides the level of risk premium predictions, the precision helps better determine the cross-sectional ranking among stocks and thus generates higher HL expected returns.[7]

Consistent with this intuition, the empirical section documents enormous economic gains from the Confident-HL portfolios. In particular, I consider a 3-layer NN (NN-3) examined by GKX to predict a large sample of U.S stock returns between 1987 and 2016. The conventional equal-weighted (EW) and value-weighted (VW) HL portfolios formed using NN-3-based risk premia earn ex-post OOS average monthly returns of 2.52% and 1.48%, with annualized Sharpe ratios of 1.5 and 0.9, respectively. However, the EW (VW) Confident-HL portfolio formed from a small subset of stocks confidently predicted by NN-3 delivers corresponding measures of 3.61% (2.21%) and 1.75 (1.09), respectively. Thus, dropping imprecise predictions enhances the OOS average returns by 43% (49%) and Sharpe ratio by 16% (21%). In contrast, measures of the EW (VW) "Low-Confident" portfolio that instead takes long-short positions on the subset of stocks with the most imprecise risk premia are relatively much lower, 2.35% (1.31%) and 1.18 (0.55), respectively.

The Confident-HL portfolio's impressive performance hinges on the theoretical result showing

[7] Mathematically, the prediction uncertainty induces downward bias to the maximum possible expected HL return that can be obtained when true risk premia are known. This result follows from Jensen's inequality (see section 2).

that NN-based predictions' ex-ante standard errors predict their ex-post squared forecast errors. Consistent with this result, I find that the ex-ante confidence and ex-post OOS-R^2 of NN-based predictions are monotonically related. The bottom decile containing the stocks with the most imprecise ex-ante return predictions attain an OOS-R^2 of 0.81%. In contrast, the top decile of stocks confidently predicted by NN-3 delivers a dramatic 2.21% OOS-R^2, an increase of 170%.

Notably, Confident-HL portfolios based on simple models involving a few parameters (e.g., Lewellen) are less likely to deliver impressive gains. Biases rather than variances predominantly determine expected forecast errors of simple models. Consistent with this result, I find that the Confident-HL portfolios formed using the Lewellen model's risk premium predictions and standard errors do not yield economic gains. Unfortunately, it is not possible to construct "Low-Bias-HL" portfolios (analogous to "Confident-HL" portfolios) for simple models using ex-ante biases (rather than standard errors) because true risk premia are unknown.

To assess whether the documented NN-3-based Confident HL portfolios' OOS gains *statistically* outperform other strategies, I first show that the existing DM tests are inadequate because they do account for ex-ante standard errors. Although ex-ante estimation uncertainty impacting ex-post OOS inferences seems instinctively puzzling, a simple example demonstrates the main intuition.

Consider comparing OOS returns of any two model-based HL portfolios. These portfolios could be expressed as different weighted sums of excess returns, depending on which stocks comprise the portfolios' long and short legs. Every period, the weights are estimated using all past data. The DM t-test thus equals the ratio of the HL return differentials' time-series average to its standard error estimate. DM show that this test yields valid asymptotic inferences only under the assumption that the return differential series is covariance stationary. However, the precision of the portfolios' estimated weights increases over time as more data are available. Thus, the HL return differentials exhibit time-varying second moments, breaking down the DM assumption.

Consistent with this intuition, I empirically establish that all model-based HL returns violate the DM assumption. The covariance-stationarity tests of Pagan and Schwert (1990) lends support to non-stationarities in the HL returns, suggesting that the DM tests are inadequate. To conduct valid OOS inferences, I develop a bootstrap procedure that is robust to non-stationarities induced by estimation uncertainty. The method builds on the block bootstrap procedure of Kunsch (1989), which provides asymptotically valid inferences in the presence of non-stationarities (Gonçalves and White (2002, 2005)).

The bootstrap tests suggest that the differences between NN-3 and Lewellen-based conventional HL strategies' OOS returns and Sharpe ratios are either statistically insignificant or moderately significant. For example, a seemingly large 0.72% (0.37%) difference between the EW (VW) NN-3-based and Lewellen-based HL portfolios' average monthly OOS returns are statistically insignificant at the 1% (10%) level.[8]

However, the NN-3-based Confident-HL strategy statistically outperforms all other competing strategies, including NN-3-based conventional HL portfolios, as well as Lewellen-based HL and Confident-HL portfolios. Moreover, the relative performance of NN-3 over Lewellen increases monotonically with the precision of NN-3-based risk premia. For example, the average monthly return difference between NN-3 and Lewellen VW HL portfolios formed using the stocks most confidently predicted by NN-3 is a highly significant 0.82%. In contrast, the difference is a significantly negative -1.2% on the subset of stocks most imprecisely predicted by NN-3. These results demonstrate that besides risk premium predictions, ex-ante standard errors are crucial for constructing desirable NN-based investment portfolios.

Avramov et al. (2020) argue that investments based on NN-3 predictions primarily extract gains from microcaps (i.e., stocks with market capital smaller than the 20^{th} NYSE size percentile) and deliver insignificant OOS returns on non-microcaps. However, I find that the Confident-HL portfolios yield significant economic gains even on non-microcaps. For example, the EW (VW) Confident-HL portfolio yields an average OOS monthly return of 2.25% (2.07%), whereas the HL strategy delivers 1.66% (1.42%). The Confident-HL portfolios' performance is robust to transaction costs, traditional factor model risk exposures and higher-moment risks that penalize losses more than rewarding gains.

[8] My results do not directly compare with GKX for one main reason, among others. Lewellen (2015) advocates three benchmark linear models with either three, seven, or fifteen characteristics. Whereas GKX use the model with three predictors, I examine the model with fifteen that Lewellen showed to exhibit superior return forecasting ability. Nevertheless, the conclusion that the DM tests over-reject any of Lewellen's models in favor of NNs remains valid.

To ensure that the Confident-HL strategies' superior performance is not driven by inadvertently taking long (short) positions on the stocks that have higher (lower) risk premium predictions, I construct several matching strategies. These portfolios resemble the conventional HL strategies but are matched to have the same "predicted-return" averages as those of the Confident-HL portfolios. Whereas the EW-Confident HL portfolio yields a 3.61% monthly OOS return, the matching HL strategy makes 3.07%. This result, consistent with the previously described example, reiterates that for the same levels of risk premia, trading strategies formed from stocks with more confident risk premia earn higher expected returns. The significant 0.55% monthly return difference between the two portfolios precisely captures the economic value of incorporating standard error information into trading strategies.

In the final exploration, I document interesting time-series and cross-sectional variations in the ex-ante standard errors that have important economic relevance. In the time-series, aggregate monthly standard errors (i.e., cross-sectional averages of ex-ante standard errors) reflect time-varying financial market uncertainty. Bloom (2009) and Baker, Bloom, and Davis (2016) document that market uncertainty jumps up after major shocks (e.g., Black Monday, Lehman Brothers bankruptcy). Consistent with these studies, the aggregate standard errors spike an average of at least twice the value of other periods. Because many individual predictors (e.g., size, price trends, and stock market volatility) in the NN-3 model substantially deviate from their usual distributions when markets are uncertain, risk premium predictions based on these unusual predictors would be hugely imprecise. Thus, the aggregate standard errors capture market uncertainty.

In the cross-section, the NN-3 model (*ex-ante*) confidently predicts risk premia of stocks associated with small market capital, high book-to-market ratios, high 1-year momentum returns, and high risk premium predictions. Thus, the NN-3-based investment strategies deliver more gains in the long-leg rather than the short-leg. This result contrasts with the "arbitrage asymmetry" studies, which argue that anomaly-based investment portfolios yield relatively more profits in the short-leg (e.g., Stambaugh, Yu, and Yuan (2012) and Avramov, Chordia, Jostova, and Philipov (2013)). Thus, possible mechanisms that lead to the association between the level and precision of (NN-based) risk premium predictions still need to be explored.

To summarize, this paper quantifies the *ex-ante* precision of the NN-based risk premium predictions and exploits this information to construct desirable Confident-HL investment portfolios. To statistically assess these portfolios' OOS performance, the paper shows that the existing DM tests are inadequate because they do not take into account ex-ante estimation uncertainty. I propose a bootstrap test that permits valid OOS inferences. The tests suggest that the NN-3-based Confident-HL portfolios significantly outperform the traditional NN3-HL and Lewellen-HL portfolios in terms of their OOS returns and Sharpe ratios, whereas the reported dominance of the conventional NN3-HL over the Lewellen-HL portfolio is statistically insignificant.

A. Contribution

The paper makes three crucial methodological and investment-related contributions.

Ex-ante standard errors. This paper generalizes the "dropout" procedure developed by Gal and Ghahramani (2016) to obtain standard errors of NN-based risk premium predictions. They show that an NN that employs dropout regularization is a Bayesian NN with a similar structure, and they estimate standard errors of NN-based predictions using the comparable Bayesian models' instantly available posterior variances. However, these are standard errors of individual "raw" predictions (equivalent to excess return predictions), not of "prediction means" (comparable to risk premium predictions). Moreover, they do not discuss how to obtain "joint densities" of different predictions from Bayesian NNs, which are necessary to compute portfolio-level standard errors. Nor do they show whether these Bayesian standard errors satisfy frequentist properties.

To my knowledge, this is the first paper to compute stock-level and portfolio-level standard errors of NN-based *risk premium* estimates by explicitly deriving the marginal and joint densities of expected return predictions from Bayesian NNs. I draw an equivalence between the frequentist and Bayesian standard errors and use simulations to show that the computed standard errors satisfy frequentist properties with accurate coverage probabilities. For example, simulations indicate that 95% (or any x% with $0 < x < 100$) confidence intervals constructed from risk premium predictions and their standard errors cover the true simulated risk premia with nearly 95% (x%) probability.

Out-of-Sample Comparisons. The paper relates to studies that compare competing return forecast models, including Goyal and Welch (2003, 2008), GKX, Bianchi et al. (2020), Bali et al. (2020), and Chen et al. (2020). These studies use either the OOS DM tests or assess the point estimates of OOS Sharpe ratios and OOS-R^2s, without accounting for estimation uncertainty. In contrast, this paper's block bootstrap method generalizes the DM tests by automatically accounting for non-stationarities induced by estimation uncertainty. This method can be employed to assess OOS performance of any model-based return predictions.

Investment Portfolios. The paper relates to studies, including GKX, Chinco, Clark-Joseph, and Ye (2019), and Avramov et al. (2020), that construct traditional HL portfolios based on various model-based return predictions. Alternatively, this paper shows how Confident-HL strategies could deliver superior expected returns. These strategies generally apply to all model-based return predictions, as long as their predictive standard errors are informative about their squared forecast errors.

B. Paper Overview

I organize the rest of the paper as follows. Section 2 provides the basics of model-based risk premium predictions and shows why the Confident-HL portfolios yield superior expected returns. Section 3 presents the statistical framework of NN-based risk premia and derives their standard errors. Section 4 shows how to conduct valid OOS inferences. Section 5 presents the empirical results. Section 6 concludes. Appendix includes proofs of propositions and simulations. Internet Appendix contains additional robustness checks and simulations.

2. Risk Premium Predictions and Predictive Standard Errors

This section presents the fundamental premise of measuring risk premia based on general econometric models, including the traditional linear and more advanced ML models (e.g., NN). It builds on the bias-variance tradeoff to explain why ML models' predictive standard errors are informative about their squared forecast errors, thus yielding large economic gains in terms of appropriate investment portfolios.

A. Basics of model-based risk premium predictions

In the spirit of GKX, consider a general additive prediction error model for realized stock returns in excess of the risk-free rate, given by

$$r_{i,t+1} = E_t(r_{i.t+1}) + \epsilon_{i,t+1},\ E_t(\epsilon_{i,t+1}) = 0,\ V_t(\epsilon_{i,t+1}) = \sigma^2 \tag{2}$$

where $r_{i,t+1}$ is the excess return of stock i at period $t+1$; $E_t(r_{i,t+1})$ is the stock i's unobserved conditional risk premium at period t; and $\epsilon_{i,t+1}$ is the unexpected component of returns due to new information at $t+1$, which is unpredictable at t. $E_t(.)$ and $V_t(.)$ denote the conditional expectation and variance operations, respectively. $\epsilon_{i,t+1}$ are iid over time and across stocks.

Let a flexible model $f(z_{it};\beta)$, involving stock-level predictors $\{z_{it}\}_{(it)}$ and parameters β, estimates unobserved risk premia. The set of predictors could be potentially large, containing many characteristics (e.g., size and book-to-market) and macroeconomic variables (e.g., earnings-to-price, stock market volatility). Like GKX, the parametric form of the model, $f(.)$, remains the same across different stocks and over time, thereby exploiting information from the entire panel of data to yield stable risk premium measurements. Because the true parameters, β, are unknown, the risk premia are estimated by

$$E_t(r_{i,t+1}) \approx f(z_{it};\hat{\beta}),\ \forall \text{ stocks } i, \tag{3}$$

where $\hat{\beta}$ are estimated parameters from the past data. The expected squared forecast errors of the model-based risk premium predictions are given by

$$E_t\left[\left(E_t(r_{i,t+1}) - f(z_{i,t};\hat{\beta})\right)^2\right] = E_t\left[\left(r_{i,t+1} - f(z_{i,t};\hat{\beta})\right)^2\right] - V_t(\epsilon_{i,t+1}),\ \forall i. \tag{4}$$

Because $\epsilon_{i,t+1}$ and $\{z_{it}\}_{(i,t)}$ are independent, minimizing the risk-premium squared forecast errors is equivalent to minimizing the realized return squared forecast errors. Thus, the best risk premium measurements are those that accurately predict subsequent returns. Consequently, the literature uses the following specification to estimate the true risk premia:

$$r_{i,t+1} = f(z_{it};\beta) + \eta_{i,t+1},\ E_t(\eta_{i,t+1}) = 0, \tag{5}$$

where risk premium and next period return ($\hat{r}_{i,t+1}$) predictions are given by

$$E_t(r_{i,t+1}) \approx \hat{r}_{i,t+1} = f(z_{it};\hat{\beta}) \tag{6}$$

Importantly, the expected squared forecast errors of return predictions based on (5) could be decomposed as the sum of three terms, given by

$$E_t\left[(r_{i,t+1} - f(z_{i,t};\hat{\beta}))^2\right] = \underbrace{\Big(E_t(r_{i,t+1}) - E_t(f(z_{i,t};\hat{\beta}))\Big)^2}_{Bias^2} + \underbrace{E_t\ \Big(f(z_{i,t};\hat{\beta}) - E_t(f(z_{i,t};\hat{\beta}))\Big)^2}_{Variance} + V_t(\epsilon_{i,t+1}). \tag{7}$$

The first term in the right hand side of (7), popularly known as “squared-bias”, measures the model misspecification of $f(.)$ in estimating the true risk premia. The second, known as “variance”, quantifies parameter uncertainty. The ex-ante predictive standard errors, which are the main focus of this paper, exactly equal the square root of the variance component. The final term, known as “irreducible-variance”, captures the realized return variation due to unpredictable new information. Under the assumption that $V_t(\epsilon_{i,t+1})$ is constant across the stocks, the squared-bias and variance components wholly determine the cross-sectional variation in squared forecast errors. These components also explain the squared forecast errors’ time-series variation.

Remark-1: Ex-post squared forecast errors of risk premium predictions based on simple linear models are challenging to predict ex-ante. Such models comprise few parameters and thus yield small predictive standard errors. However, they are grossly misspecified when the true risk premia are non-linear functions of many predictors. Hence, squared-bias rather than variance largely governs their forecast-squared errors. Because true risk premia are unobserved, ex-ante measurement of squared-bias is not possible, rendering simple models’ forecast-squared errors unpredictable ex-ante.

Remark-2: In contrast, ex-post forecast errors of ML-based predictions are ex-ante predictable. These predictions use many predictors and parameters and thus are less likely to be misspecified. However, their massive predictive standard errors, which reflect parameter uncertainty, predominantly determine their forecast-squared errors. These standard errors, unlike biases, are readily obtainable, rendering ML models' forecast-squared errors predictable ex-ante. For instance, in the cross-section, stocks whose ML-based risk premium predictions have large ex-ante standard errors also have large ex-post squared forecast errors.

Consistent with these remarks, the empirical section documents that the ex-ante predictive standard errors of the NN-based risk premium predictions strikingly predict their ex-post squared forecast errors, whereas those of the Lewellen-based predictions do not. The following subsection illustrates how these ex-ante standard errors could be used in real-time to form desirable investment portfolios that yield large economic gains.

B. Risk Premium Predictions, Standard Errors and Investment Portfolios

This subsection introduces the Confident-HL portfolios that deliberately exclude or downweight stocks with large predictive standard errors from the extreme predicted-return decile stocks. I restate the example provided in the introduction to illustrate why these portfolios yield superior expected returns relative to the conventional HL strategies.

Example-1. Consider two stocks A and B with true risk premia μ_A and μ_B ($< \mu_A$), respectively. Let $\hat{\mu}_A$ and $\hat{\mu}_B$ be the predicted risk premia based on an econometric model, satisfying

$$\hat{\mu}_A = \mu_A + \epsilon_A,\ \hat{\mu}_B = \mu_B + \epsilon_B,\ \epsilon_A, \epsilon_B \sim N(0, \sigma^2),\ \epsilon_A \perp \epsilon_B. \tag{8}$$

Recall that the assumption of unbiased predictions ($E(\epsilon_A), E(\epsilon_B) = 0$) is more likely to hold for ML-based rather than traditional linear models. For simplicity, (8) assumes uncorrelated predictions with the same predictive standard error, σ. Proposition-1 relaxes this assumption and generalizes for heteroskedastic standard errors.

The expected return of the traditional HL portfolio that goes long (short) on the stock with the highest predicted risk premium is then given by

$$E(HL) = (\mu_A - \mu_B)P(\hat{\mu}_A > \hat{\mu}_B) + (\mu_B - \mu_A)P(\hat{\mu}_B > \hat{\mu}_A) = (\mu_A - \mu_B)\left[2\Phi\left(\frac{\mu_A - \mu_B}{\sqrt{2}\sigma}\right) - 1\right], \quad (9)$$

where $P(.)$, $\Phi(.)$ denote the probability and standard normal distribution measures, respectively.

Thus, (9) indicates that the expected HL return monotonically increases (decreases) with the precision of risk premium predictions (σ). Mathematically, the prediction uncertainty induces bias to the maximum possible expected HL return that can be obtained when true risk premia are known. For example, the HL strategy formed from the zero standard error predictions delivers the maximum possible expected return of $(\mu_A - \mu_B)$, as the strategy always takes the long (short) position on A (B) by perfectly ranking the stocks. In contrast, the HL portfolio formed from grossly imprecise predictions ($\sigma = \infty$) earns zero expected returns, with a bias of $(\mu_A - \mu_B)$. This result follows from Jensen's inequality: " The expectations of the maximum (minimum) of a given set of risk premium predictions are lower (higher) than the maximum (minimum) of the expectations of predicted risk predicted risk premia". The lower the variance of risk premium predictions, lower will be the difference between both.

The following proposition builds on this intuition and formally establishes the Confident-HL strategies' superiority over the conventional HL portfolios.

Consider four stocks A_1, A_2, B_1, and B_2 with true risk premia μ_A, μ_A, μ_B $(< \mu_A)$, and μ_B, respectively. Predictions are unbiased, independent, and normal, but could have different predictive standard errors. To form trading strategies, stocks are sorted into two quantiles, denoted by Q_S and Q_L. Q_L (Q_S) comprises the two stocks with the highest (lowest) risk premium predictions. Now, consider the following three long-short investment strategies:

1. **HL**: The traditional HL strategy takes the EW long (short) positions on the 2 Q_L (Q_S) stocks.

2. **PW-HL**: The "precision-weighted" (PW) HL portfolio also takes the long (short) positions on the two Q_L (Q_S) stocks, but overweights ($> 50\%$) the precisely predicted stock in each quantile.

3. **Confident-HL**: This strategy takes the long (short) position only on the stock with the lowest predictive standard error in each quantile, deliberately excluding the stock with imprecise risk premium.

Then, the expected returns of these portfolios are in the order of

Proposition 1:

$$E(\text{HL}) \leq E(\text{PW-HL}) \leq E(\text{Confident-HL}). \tag{10}$$

Proof. See Appendix (A.1). □

The proof is similar to the previous example. Thus, proposition-1 indicates that the Confident-HL portfolios dominate the traditional HL portfolios in terms of earning higher expected returns. Proposition-1 makes the stylized assumption of uncorrelated predictions for mathematical tractability, as it is not possible to generalize this result with correlated predictions. However, Internet Appendix C.C1 (table A) presents an extensive simulation study to validate proposition-1 for general cases with many stocks, correlated return predictions and Confident-HL portfolios formed from various other quantile portfolios (e.g., decile).

Consistent with these results, the empirical section documents large economic gains emanating from the Confident-HL portfolios based on the NN-3 risk premium predictions and their standard errors. Such large gains would not be realized from the Lewellen-based Confident-HL portfolios, as their predictive standard errors do not predict their squared forecast errors.

Before deriving NN-based risk premia's predictive standard errors to form the Confident-HL portfolios, it is worth emphasizing a couple of important points. First, dropping stocks with imprecise risk premia improves the expected returns of HL strategies, not necessarily their variance, as it may reduce the diversification benefit. Determining the trade-off between expected HL returns and their variances is ultimately an empirical question. The empirical section shows that the Confident-HL portfolios formed using the standard decile-sorted rules deliver superior Sharpe ratios, suggesting that the expected return improvements are relatively larger. Second, the Confident-HL strategies exploit information only from the variance of risk premium predictions and not predicted return variances nor covariances. Forming optimal portfolios using all stock returns' joint predictive density requires a Bayesian framework, thus left for a future study.

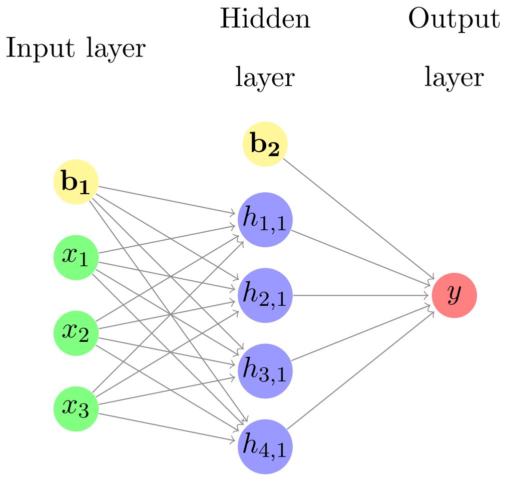
3. NN-based Risk Premia and Standard Errors

This section presents the statistical framework to predict individual stock- and portfolio-level risk premia using NN. It then theoretically derives their standard errors, shown to be easily obtainable with no additional computation cost. In particular, an NN that employs a specific regularization known as "dropout" is identical to a Bayesian NN with a similar structure (Gal and Ghahramani (2016)). A simple analogy to this identity is the equivalence between linear regressions with L_2 regularization (i.e., Ridge regressions) and Bayesian linear regressions. Thus, NN-based predictive standard errors are estimated using the comparable Bayesian models' instantly available posterior variances.

Although Bayesian posterior variances and frequentist standard errors philosophically represent different entities, the section justifies why and how the obtained standard errors satisfy critical frequentist properties with accurate coverage probabilities. This is important, because no frequentist alternative currently exists (to my knowledge) to provide standard errors.

A. Neural Networks

Figure 1. Example of a 1-layer Neural Network



Note: An example of a 1-layer, feed-forward neural network.

Like GKX, this paper considers conventional "feed-forward" NNs, which consist of an "input layer" of raw predictors, one or more "hidden layers" and an "output layer" of a final prediction, in that order. Each layer is composed of neurons that aggregate information from the neurons of (immediately) preceding layer. Thus, information hierarchically flows from the raw predictors of the input layer to the neurons in the hidden layers and finally to the final prediction in the output layer. To understand how NNs systematically conduct this prediction exercise, figure (1) shows a simple example of a 1-layer NN (NN-1) with 3 and 4 neurons in the input and hidden layers, respectively.

In figure (1), $\{x_1, x_2, x_3\}$, $\{h_{k,1}\}_{k=1}^{4}$, and y are the sets of neurons in the input, hidden, and output layers, respectively. Furthermore, $\{x_i\}_{i=1}^{3}$ are raw individual predictors, and y is the final output prediction. Each neuron in the hidden layer applies a nonlinear function (ϕ) to an aggregate signal received from the preceding (input) layer. The aggregate signal is a weighted sum of the preceding layer's neurons plus an intercept, known as "bias". Thus,

$$h_{k,1} = \phi\left(b_{1k} + \sum_{j=1}^{3} w_{1jk}x_j\right), \text{ for } k = 1, 2, 3, 4, \tag{11}$$

where b_{1k} is the intercept associated with the input (first) layer and k^{th} neuron in the (next) hidden layer, and w_{1jk} is the weight associated with the j^{th} predictor (neuron) in the input layer and the k^{th} neuron in the hidden layer. The linear sum, $(b_{1k} + \sum_{j=1}^{3} w_{1jk}x_j)$, is the aggregated signal received by the hidden layer's $h_{j,1}$ neuron from the input layer. In the spirit of GKX, the nonlinear function ϕ takes the rectified linear unit functional form (ReLU). However, the theory developed in this section holds for any general function. The ReLU is given by

$$\phi(x) = ReLU(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ otherwise.} \end{cases} \tag{12}$$

Likewise, the final output is given by

$$y_{output} = b_2 + \sum_{j=1}^{4} w_{2j}h_{j,1}, \tag{13}$$

where w_{2j} is the weight associated with the j^{th} neuron in the hidden layer and the output. Thus, given an input of Q individual predictors, x, the final prediction, y_{output}, based on a general NN-1 model with K hidden neurons can be expressed in the parametric form

$$y_{output} = b_2 + \phi(b_1 + xW_1)W_2, \tag{14}$$

where $\{W_1, W_2, b_1, b_2\}$ are the unknown parameters. W_1 and W_2 are the weight matrices connecting the imput layer to the hidden layer and hidden layer to the output layer, respectively. Intercepts b_1 and b_2 are added to the hidden and output layers, respectively. W_1 is a $Q \times K$ matrix, W_2 is a $K \times 1$ vector, b_1 is a $K \times 1$ vector, and b_2 is a scalar.

B. Parameter Estimation, Regularization, and Dropout

For simplicity, the rest of the section focuses on NN-1 models. However, the theory that follows holds in general for any feed-forward NN with an arbitrary number of hidden layers and neurons. Consider the return prediction specification in (5),

$$r_{it+1} = f(z_{it}; \beta) + \eta_{i,t+1}, \tag{15}$$

where $r_{i,t+1}$ is stock i's excess return at period $t+1$, and z_{it} is the set of stock i's raw predictors at time t. When f is an NN-1, it takes the parametric form in (14), with $\beta = \{W_1, W_2, b_1, b_2\}$.

Because the parameters are unknown, risk premia are measured as $E_t(r_{i,t+1}) \approx f(z_{it}; \hat{\beta})$, where $\hat{\beta}$ are estimated parameters of β. Given a panel of "training data", the literature typically minimizes the mean of squared forecast errors to estimate the parameters, i.e

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it} W_1) W_2)\right)^2, \tag{16}$$

where Tr is the training sample over N_{Tr} periods, and S is the total set of N_S stocks. The estimated parameters from (16) often overfit the data by taking extreme values. To alleviate this concern, the literature adds various penalties such as L_2 regularization to the usual squared forecast error loss function. Under L_2 regularization, the estimated parameters are given by

$$\begin{aligned}\hat{\beta}_\lambda = \arg\min_{\beta} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} &\left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it} W_1) W_2)\right)^2 \\ &+ \lambda \left[||W_1||^2 + ||W_2||^2 + ||b_1||^2 + ||b_2||^2\right], \end{aligned} \tag{17}$$

where $||.||$ represents the L_2 norm operator, and λ is known as the "hyperparameter". Note that the estimated parameters depend on the hyperparameter λ. From a given set of hyperparameters, the standard practice chooses the λ that minimizes the forecast-squared error mean in a panel of "validation data" that do not overlap with the training data. In particular,

$$\lambda = \arg\min_{\lambda \in \Lambda} \frac{1}{N_V N_S} \sum_{t \in V} \sum_{i \in S} \left(r_{i,t+1} - f(z_{it}, \hat{\beta}_\lambda) \right)^2, \tag{18}$$

where V is the validation sample over N_V periods, and Λ is a given set of hyperparameters.

Thus, (17) and (18) together determine the estimated parameters and hyperparameters. Because the optimal parameters that minimize (17) are not available in closed-forms, numerical algorithms start with an initial estimate (guess), and then iteratively update the parameters by feeding each observation into the training data one-by-one. This procedure could be computationally intensive. Thus, a popular algorithm known as stochastic gradient descent (SGD) considers random samples (rather than the full sample) from the training data to iteratively update the parameters until they converge.[9]

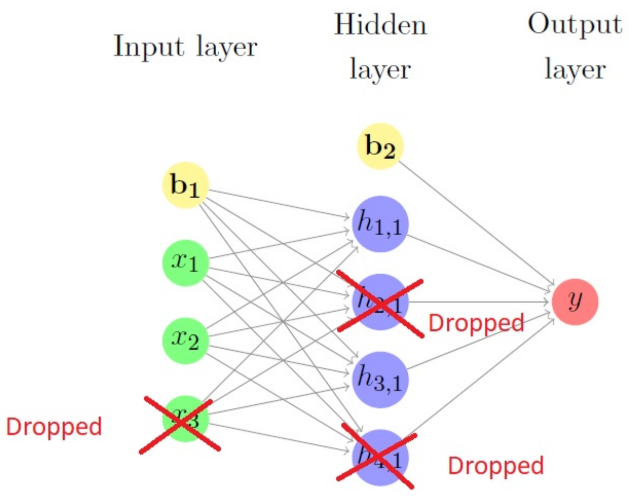
Besides L_2, GKX use several other regularizations, such as L_1, to minimize overfitting. This subsection introduces another popular regularization known as dropout that can be employed either exclusively or simultaneously with other penalties. Dropout stands out among others because it boosts the performance of NN models and helps determine predictive standard errors. GKX do not discuss the dropout procedure. In a recent working paper, Chen et al. (2020) use dropout to fit various NNs for predicting stock returns. However, they do not address how such a regularization could be exploited to obtain predictive standard errors.

Dropout. Dropout is a simple but powerful regularizations proposed by Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014).[10]

[9] See GKX for a detailed review of parameter estimation using SGD.

[10] See Géron (2019) for an excellent non-technical summary on dropout regularization.

Figure 2. NN-1 with Dropout Regularization



Note: The figure shows an NN-1 with dropout regularization. At each training iteration, a random subset of all neurons in one or more layers, including the input layer, but always excluding the output layer, is dropped. Each iteration's dropped out neurons temporarily output 0 (during that iteration), but might become active in the next iteration.

At each training iteration during parameter estimation, every neuron, including the input neurons, but always excluding the output neurons, has a probability $(1-p)$ of being temporarily dropped. These dropped out neurons are deliberately set to output 0 (equivalently, discarded) during that iteration but are allowed to become active in the next iteration. Like λ for L_2, $(1-p)$ (p) is a hyperparameter known as "dropout rate" ("retention rate"), and thus chosen (typically between 10% and 50%) to minimize the validation forecast-squared error. After training and obtaining estimated parameters, neurons are no longer dropped (i.e., to make a new prediction). Figure (2) shows an example of an NN-1 with dropout regularization.

To summarize, during parameter estimation, dropout randomly disconnects a few neurons at each iteration to avoid overfitting and improves performance. Consider a random sample of 1000 observations from training data for parameter estimation. The SGD algorithm takes 1000 iterations to estimate the parameters. Employing dropout would imply 1000 different NNs are trained, yielding 1000 distinct estimated weights. These weights are not independent but are nevertheless all different. The final estimated weights could be interpreted as an average of these distinct weights, thereby alleviating parameter uncertainty.

Estimated parameters of an NN-1 that employ dropout and L_2 regularizations satisfy

$$\hat{\beta}_{\lambda,p} = \arg\min_{\beta} \frac{1}{N_{Tr}N_S} \sum_{t\in Tr} \sum_{i\in S} (r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it}W_1))(p_{2it}W_2)))^2 \\ + \lambda \left[||W_1||^2 + ||W_2||^2 + ||b_1||^2 + ||b_2||^2\right], \tag{19}$$

where each element in p_{1it} and p_{2it} is an independent draw from a *Bernoulli* distribution with parameter (p) ((1-dropout rate)). p_{1it} and p_{2it} are $(Q \times Q)$ and $(K \times K)$ diagonal matrices, respectively. Thus, unknown parameters could be estimated by solving (19).[11] Hereafter, an NN that employs L_2 and dropout regularizations will be called a "dropout NN".

Stock-level risk premia. Given newly observed "test data" (Te) of raw predictors that do not overlap with the training and validation data sets, a dropout NN-1-based risk premium prediction is given by

$$E_t(r^*_{i,t+1}) \approx E^*_{it,Dropout} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}W_{1,\{\lambda,p\}})W_{2,\{\lambda,p\}}),\ \ r^*_{i,t+1}, z^*_{it} \in Te, \tag{20}$$

where the parameters, $\{b_{2,\{\lambda,p\}}, b_{1,\{\lambda,p\}}, W_{1,\{\lambda,p\}}, W_{2,\{\lambda,p\}}\}$, are given in (19). $E^*_{it,Dropout}$ represents the dropout NN-1-based risk premium prediction of stock i at period t. Note that no neurons are dropped out while making predictions on the test data. However, these predictions rely on estimated parameters that employ dropout regularization. In fact, Srivastava et al. (2014) establish that the predictions given in (20) are approximately equal to the sample averages of corresponding predictions that employ dropout at the test time as well. In particular,

$$E^*_{it,Dropout} \approx \frac{1}{D}\sum_{d=1}^{D}(b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1id}W_{1,\{\lambda,p\}}))(p_{2id}W_{2,\{\lambda,p\}})),\ \ r^*_{i,t+1}, z^*_{it} \in Te, \tag{21}$$

[11] The most commonly used software programs, including Python and Matlab, readily solve (19).

where each element in $\{p_{1i,d}, p_{2i,d}\}_{i=1}^{D}$ is an independent draw from $\sim Bernoulli(p)$, and D is the total number of distinct predictions drawn at the test time with dropout applied.

Portfolio-level risk premia. The risk premium prediction, $E^{*}_{Pt,Dropout}$, of portfolio P formed using a set of stock-level weights $\{\omega_{P,i,t}\}_{i=1}^{S}$ at the beginning of period $t+1$ is given by

$$E_t(r^{*}_{P,t+1}) = \sum_{i=1}^{S} \omega_{P,i,t} r^{*}_{i,t+1} \approx E^{*}_{Pt,Dropout} \approx \sum_{i=1}^{S} \omega_{P,i,t} E^{*}_{it,Dropout},\ \ r^{*}_{i,t+1} \in Te, \tag{22}$$

where $E^{*}_{it,Dropout}$ is given in (21).

Importantly, it turns out that the risk premium estimates in (20) (or (21)) and (22) are approximately equal to the respective risk premia's posterior density means under an equivalent Bayesian NN with a similar structure. Using this approximation but before formally discussing Bayesian NNs, the following subsection illustrates how to instantly obtain standard errors of general dropout NN-based risk premium predictions.

C. Standard Errors of Risk Premium Predictions based on Neural Networks

Stock-level standard errors. Given a new observation of a stock's raw predictors z^{*}_{it} in the test data, consider its risk premium prediction based on a dropout NN-1

$$E_t(r^{*}_{i,t+1}) \approx E^{*}_{it,Dropout} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^{*}_{it} W_{1,\{\lambda,p\}}) W_{2,\{\lambda,p\}}),\ \ r_{i,t+1}, z^{*}_{it} \in Te. \tag{23}$$

Then the predictive standard error of $E^{*}_{it,Dropout}$ is estimated by the sample standard deviation of distinct predictions that are obtained by randomly dropping out neurons (with probability $(1-p)$) at the test (prediction) time. In particular,

$$SE_t(E^{*}_{it,Dropout}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t+1}\right)^{2}}, \tag{24}$$

where D is the total number of distinct predictions ($\hat{E}_{i,d,t}$) drawn, with each $\hat{E}_{i,d,t}$ given by

$$\hat{E}_{i,d,t} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d}W_{1,\{\lambda,p\}}))(p_{2d}W_{2,\{\lambda,p\}})), \;\; z^*_{it} \in Te. \tag{25}$$

Every element in $p_{1,d}$, $p_{2,d}$ is an *iid* draw from the $Bernoulli(p)$ distribution. The empirical section considers $D = 100$ to estimate the standard errors, as simulations confirm that it yields well-calibrated estimates.[12]

To summarize, after estimating an NN-1 model's weights using the training and validation data sets, standard errors of risk premium predictions on the test data are quickly available by collecting predictions that deliberately assign 0 to randomly selected weights. Intuitively, as the following subsection shows, this procedure is equivalent to drawing samples from the risk premium's predictive distribution based on a comparable Bayesian NN having the same number of neurons and hidden layers as the considered NN-1.

Portfolio-level standard errors. Likewise, the predictive standard error of a portfolio-level prediction is given by

$$SE_t(E^*_{Pt,Dropout}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2}, \tag{26}$$

where

$$\hat{E}_{P,d,t} = \sum_{i=1}^{S}\omega_{P,i,t}\left(b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d}W_{1,\{\lambda,p\}}))(p_{2d}W_{2,\{\lambda,p\}})\right), \;\; z^*_{it} \in Te, \tag{27}$$

and $p_{1,d}$, $p_{2,d}$ are *iid* draws from $Bernoulli(p)$.

The procedure for computing portfolio-level standard errors deserves emphasis. Note that the dropped weights (i.e., p_{1d}, p_{2d} draws) are the *same* across the stocks that composite P, thereby preserving correlations among stock-level risk premium predictions to yield unbiased standard error estimates, as shown in the following subsection.

[12] The higher D is, the more accurate uncertainty estimates will be. However, inference time also increases with D. Thus, an ideal D trades-off between latency and accuracy.

The outlined procedure for obtaining standard errors in (24) and (26) generally applies to all predictions based on NNs with an arbitrary number of layers and neurons as long as their weights are estimated using dropout and L_2 regularizations (Gal and Ghahramani (2016)). The procedure is also robust to adding more regularizations, such as implementing the SGD algorithm with an arbitrary learning rate.

It is worth emphasizing that (24) and (26) yield standard errors of risk premium predictions and not excess return predictions. Fama and French (1997) and Pástor and Stambaugh (1999) also compute risk premium estimates' standard errors. Recall that realized excess returns equal the sum of risk premia and unexpected returns due to unpredictable new information. Thus, their predictive variances equal the sum of predictive variances of risk premium predictions and "irreducible-variance" due to unexpected returns (see (7)). The validation data's mean squared error is an asymptotically unbiased estimate of irreducible-variance (Zhu and Laptev (2017)). Thus, predictive variances of return predictions are easily obtainable as well.

D. Dropout Neural Networks and Bayesian Interpretation

This subsection illustrates a profound connection between dropout NNs and Bayesian NNs to formally validate the previously presented standard errors under a Bayesian framework.

In an influential work, Gal and Ghahramani (2016) first prove that dropout NNs have a Bayesian interpretation. In doing so, they draw upon the probability theory of Gaussian processes, thereby limiting the potential audience for their work. Moreover, they show how to estimate standard errors of *individual* NN-based "raw" predictions (analogous to return predictions) but not those of "prediction means" (equivalent to risk premium predictions). They also do not discuss how to obtain "joint densities" of different NN-based predictions, which are necessary to compute portfolio-level standard errors.

I use a simple Bayesian model to provide a straightforward but rigorous discussion of their central conclusions. In a significant contribution, I (Bayesian) theoretically derive the standard errors (24, 26) of stock and portfolio-level *risk premia.*

Bayesian Neural Network. Consider the Bayesian NN analogous to the previously considered NN-1, with the parametric form given by

$$r_{i,t+1} = b_2 + \phi(b_1 + z_{it}W_1)W_2 + \eta_{i,t+1},\ E_t(\eta_{i,t+1}^2) = \sigma_\eta^2 \tag{28}$$

where the parameters $\{W_1, W_2\}$ are random. σ_η^2 and $b = (\{b_1, b_2\})$ are assumed to be known for simplicity.[13] Denote the risk premia by μ_{it}, where

$$\mu_{i,t} = E_t(r_{it+1}) = b_2 + \phi(b_1 + z_{it}W_1)W_2. \tag{29}$$

Specify the unknown weight matrices with the standard Gaussian priors,

$$[W_1, W_2] = \mathcal{N}(0, l^{-2}I),$$

where I is the $(NK + K) \times (NK + K)$ identity matrix, and l is a hyperparameter. Then the predictive density of stock i's risk premium given a set of its raw predictors, z_{it}^*, from the test data, and the past training and validation data sets, denoted by $\{R, Z\}$, is given by

$$P(\mu_{i,t}^*|z_{it}^*, R, Z) = \int P(\mu_{i,t}^*|z_{it}^*, R, Z, W_1, W_2, b, \sigma_\eta^2)P(W_1, W_2|R, Z, b, \sigma_\eta^2)dW_1dW_2, \tag{30}$$

where $P(W_1, W_2|R, Z, b, \sigma_\eta^2)$ is the posterior density of the weight matrices given past data. Because this density is not available in a closed-form, the literature often uses one of the powerful methods known as variational inference (VI) to directly approximate the intractable posterior.

The following discussion introduces VI and shows that approximating the posterior of the weight matrices using VI and frequentist estimation the weights with dropout and L_2 regularizations, as in (17), are equivalent. Thus, dropout NNs are approximations to Bayesian NNs.

Variational Inference (VI). To approximate a given posterior density $P(W|data)$, VI first considers a family of some known densities, $\{q_\theta(W)\}$, parameterized by θ, and then finds the optimal

[13] The theory generalizes when $\{b_1, b_2\}$ are allowed to be random as well, in which case these parameters could be specified with Gaussian priors.

parameter, θ^*, such that the Kullback-Leibler divergence between $q_{\theta^*}(W)$ and the true posterior density is minimized. Thus, VI approximates the true posterior with $q_{\theta^*}(W)$, where the optimal parameter θ^* would be a function of data. The key is to consider a "good" family of densities that guarantee the (almost surely) convergence of $q_{\theta^*}(W)$ to the true posterior.[14] As a reference, in the finance literature, Allena and Chordia (2020) develop the first VI method to approximate the intractable posterior density of true stock liquidity and equilibrium prices.

Variational Inference for Bayesian Neural Networks. Gal and Ghahramani (2016) consider the following family of independent Gaussian mixture densities to approximate the posterior of the NN weight matrices

$$q_{\{M_1,M_2\}}(W_1, W_2) = q_{M_1}(W_1)q_{M_2}(W_2), \text{ with } q_{M_i}(W_i) = \prod_{k=1}^{K_i} q_{m_{iq}}(w_{iq}), \text{ for } i = 1, 2, \text{ where}$$

$$q_{m_{iq}}(w_{iq}) = p\mathcal{N}(m_{iq}, \sigma^2 I) + (1-p)\mathcal{N}(0, \sigma^2 I) \text{ for } i = 1, 2, \tag{31}$$

with $M_1 = [(m_{1q})]$ and $M_2 = [(m_{2q})]$. These are the "variational" parameters to be optimized. Also, $W_1 = [(w_{1q})]$ and $W_2 = [(w_{2q})]$. σ^2 and p are known scalars. K_i is the number of neurons in the i^{th} layer. Thus, $K_1 = Q$ and $K_2 = K$. M_1 and M_2 are matrices with the same dimensions as W_1 and W_2, respectively.

The optimal set of parameters $\{M_1^*, M_2^*\}$ that best approximate the true posterior is given by

$$\{M_1^*, M_2^*\} = \arg \min_{\{M_1,M_2\}} KL\left(q_{M_1}(W_1)q_{M_2}(W_2)||P(W_1, W_2|R, Z_b, \sigma_\eta^2)\right), \tag{32}$$

where $KL(x||y)$ represents the Kullback-Leibler divergence between the two random variables, x and y.

Bayesian and Dropout Neural Network Equivalence. Interestingly, given the sample of training data, it turns out that the optimal parameters in (32) minimize the loss function that

[14] See Blei, Kucukelbir, and McAuliffe (2017) for an excellent review of VI. They address two fundamental questions: i) what family of densities to consider? ii) how to obtain the optimal density in the family that best approximates the true posterior?

resembles a dropout NN's loss function, as in (19). In particular,

$$\{M_1^*, M_2^*\} = \arg \min_{\{M_1, M_2\}} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} (r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it} M_1))(p_{2it} M_2)))^2 \\ + \mu_1 ||M_1||^2 + \mu_2 ||M_2||^2 + \mu_3 ||b_1||^2 + \mu_4 ||b_2||^2, \quad (33)$$

where each element in p_{1it} and p_{2it} is an independent draw from a *Bernoulli* distribution with parameter (p). $\{\mu_1, \dots \mu_4\}$ are different scalars that are distinct functions of $\{l_1, \sigma_\eta^2, \sigma^2\}$.

Thus, for an appropriate choice of l_1, the variational parameters, $\{M_1^*, M_2^*\}$, that best approximate the (Bayesian) NN weight matrices' posterior density are identical to the comparable (frequentist) dropout NN's estimated weights. This implies

$$M_1^* = W_{1,\{\lambda,p\}}, \text{ and } M_2^* = W_{2,\{\lambda,p\}}. \quad (34)$$

The predictive density of a risk premium given in (29) can be approximated by

$$P(\mu_{i,t}^* | z_{it}^*, R, Z) \approx Q(\mu_{i,t}^* | z_{it}^*, R, Z) = \int P(\mu_{i,t}^* | z_{it}^*, R, Z, W_1, W_2, b, \sigma_\eta^2) q_{M_1^*, M_2^*}(W_1, W_2) dW_1 dW_2, \quad (35)$$

where $\{M_1^*, M_2^*\}$ are given in (34), and $q(.)$ in (31).

As an immediate corollary, (35) implies that the mean of a risk premium's (approximated) Bayesian predictive density is

$$E\left[Q(\mu_{i,t}^* | z_{it}^*, R, Z)\right] \approx E_{it,Dropout}^* \\ \approx \frac{1}{D} \sum_{d=1}^{D} (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z_{it}^*(p_{1id} W_{1,\{\lambda,p\}}))(p_{2id} W_{2,\{\lambda,p\}})), \; z_{it}^*, \in Te, \quad (36)$$

where each element in $p_{1id}, p_{2id} \sim Bernoulli(p)$.

Thus, the mean of a risk premium's Bayesian predictive density (36) precisely matches with the comparable dropout NN-based risk premium prediction, as in (21). In simpler words, predicting risk premia using dropout NNs and Bayesian NNs are equivalent.

Bayesian Justification for Stock-level Standard Errors. Due to (36), under usual regularity conditions (e.g., prior mass is not concentrated at a single point), and for large data, the standard deviation of a risk premium's Bayesian predictive density should proxy for the standard error of its frequentist counterpart ($E^*_{it,Dropout}$).[15] This implies

$$SE_t(E^*_{it,Dropout}) = SD\left[Q(\mu^*_{i,t}|z^*_{it}, R, Z)\right], \tag{37}$$

where $SD\left[Q(\mu^*_{i,t}|z^*_{it}, R, Z)\right]$ represents the standard deviation of $\mu^*_{i,t}$'s Bayesian predictive density. This is given by the following proposition.

Proposition 2:

$$SD\left[Q(\mu^*_{i,t}|z^*_{it}, R, Z)\right] \approx \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2}, \tag{38}$$

where $\hat{E}_{i,d,t}$ *is given in* (25).

Proof. See Appendix (A.2). □

Thus, the standard errors of dropout NN-based stock-level risk premium predictions, as in (24), are justified from a Bayesian standpoint.

Bayesian Justification for Portfolio-level Standard Errors. Likewise, the standard error of a portfolio P's risk premium prediction should satisfy

$$SE_t(E^*_{Pt,Dropout}) = SD\left[Q(\mu^*_{P,t}|\{z^*_{it}\}_{i=1}^S, R, Z)\right], \tag{39}$$

where $\mu^*_{P,t} = \sum_{i=1}^S \omega_{P,i,t}\mu^*_{i,t+1}$, and $Q(\mu^*_{P,t}|\{z^*_{it}\}_{i=1}^S, R, Z)$ is the Bayesian predictive density of P's risk premium, given a set of stock-level characteristics.

Obtaining this density is not straightforward, as it involves computing the *joint* predictive

[15] This property, known as "frequentist consistency" of posteriors, is due to the Bernstein-von Mises theorem. Whereas literature often demonstrates this result for true posteriors, Wang and Blei (2019) establish that, under standard regularity conditions, approximated posteriors using VI are consistent as well. In any case, the following subsection empirically validates this result.

density of risk premia of all stocks that compose P, $Q\left(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t} | \{z^*_{it}\}^S_{i=1}, R, Z\right)$. The following proposition formally derives the joint density to compute the standard deviation of P's posterior risk premium density.

Proposition 3:

$$SD\left[Q(\mu^*_{P,t}|\{z^*_{it}\}^S_{i=1}, R, Z)\right] \approx \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2}, \tag{40}$$

where $\hat{E}_{P,d,t}$ are given in (27).

Proof. See Appendix (A.3). □

Thus, the standard errors of dropout NN-based portfolio-level risk premium predictions, as in (26), are theoretically justified as well.

E. Frequentist Justification for Standard Errors

Recall that the paper trades the Bayesian standard errors for the frequentist standard errors, as the former are instantly available but no valid method exists to compute the latter directly (to my knowledge). This subsection justifies the obtained standard errors from a frequentist standpoint, by drawing the equivalence between both standard errors and conducting extensive Monte-Carlo simulations.

Under a large sample, observed data dominates prior, rendering Bayesian and frequentist standard errors nearly identical (see result 8 in section 4.7 of Berger (1985)). However, NN-based predictions generally employ substantial regularization, which is equivalent to starting with proper priors. In such cases, data may not always dominate prior, resulting in differences between the Bayesian and frequentist approaches under specific parameters. However, such issues typically occur at the atypical values of the parameters, such as when they approach infinity (Kyung, Gill, Ghosh, and Casella (2010)).[16] Thus, for a wide range of parameters, Bayesian and frequentist

[16] In fact, Kyung et al. (2010) motivate the same to compute the otherwise intractable standard errors of LASSO based predictions using their Bayesian counterparts.

standard errors should be equivalent.

Consistent with this result, an extensive simulation study in appendix B.1 (table (1)) confirms that the proposed standard errors are well-calibrated in the frequentist sense. Using a high dimensional predictor set, risk premia are simulated from four different data generating processes. Whereas the first two model returns as a linear function of predictors with homoscedastic and correlated residuals, respectively, the last two entertain non-linear functions. Across all models, 95% (or any x% with $0 < x < 100$) confidence intervals constructed from risk premium predictions and their standard errors cover the true simulated risk premia with nearly 95% (x%) probability.

4. Ex-ante Estimation Uncertainty and Ex-post OOS Inferences

Recall that sections 2 and 3 showed how to estimate valid standard errors of NN-based risk premium predictions and exploit them to construct desirable Confident-HL portfolios. This section derives a formal method to assess the Confident-HL or any model-based investment portfolios' ex-post OOS performance.

In doing so, the section first documents that existing tests violate the central assumption required for the DM tests' asymptotic validity. The section then presents a bootstrap methodology to deliver valid OOS comparisons in the presence of estimation uncertainty. The section concludes by showing that the method yields well-sized tests, whereas the DM tests lead to significant size distortions using simulated data.

A. Out-of-Sample Comparisons with the Diebold and Mariano (2002) Tests

OOS returns of HL strategies and DM tests. Consider any two competing model-based HL strategies, HL_1 and HL_2. These portfolio returns could be expressed as different weighted sums of excess returns, depending on which stocks comprise their long and short legs. Thus,

$$HL_{1t} = \sum_{i \in S} \hat{w}_{1,i,t-1} r_{i,t}, \;\; HL_{2t} = \sum_{i \in S} \hat{w}_{2,i,t-1} r_{i,t}, \tag{41}$$

where $r_{i,t}$ denotes the excess return of stock i at period t, and $\{\hat{w}_{1,i,t-1}\}_{i\in S}$ and $\{\hat{w}_{2,i,t-1}\}_{i\in S}$ represent the weights with which individual stocks compose the HL_1 and HL_2 portfolios, respectively. The weights are estimated using all data until $t-1$. This specification is consistent with the "recursive estimation scheme" typically employed by researchers (e.g., GKX, Bianchi et al. (2020)).

Consider the return differentials over the OOS period,

$$d_{12,t} = HL_{1t} - HL_{2t},\ t \in Te, \tag{42}$$

where Te denotes the OOS test period. Then the DM statistic to test the null of equal return means, $H_0 : E(d_{12,t} = 0)\ \forall t$, is a simple t-ratio given by

$$DM_{HL} = \frac{\bar{d}_{12}}{\hat{\sigma}_{d_{12}}} \sim \mathcal{N}(0,1), \tag{43}$$

where $\bar{d}_{12} = \frac{1}{N_{Te}} \sum_{t\in Te} d_{12,t}$ is the sample average of return differentials over N_{Te} OOS periods and $\hat{\sigma}_{d_{12}}$ is a heteroskedastic and autocorrelation robust standard error estimate for $\bar{d}_{12}$. Whereas Avramov et al. (2020) use Newey-West standard errors of return differentials as a proxy for $\hat{\sigma}_{d_{12}}$, most studies use the standard OLS standard errors.

OOS MSEs and DM tests. Likewise, existing studies employ the DM tests to compare OOS mean squared errors (MSEs) of any two competing models as well. Given two models M_1 and M_2, let $f_1(Z_{i,t-1}; \hat{\beta}_{1,t-1})$, $f_2(Z_{i,t-1}; \hat{\beta}_{2,t-1})$ be the return predictions for period t based on M_1 and M_2, respectively. Then the forecast-squared error differentials over the OOS period are given by

$$D_{12,t} = e^2_{1,t} - e^2_{2,t},\ \text{where } e^2_{k,t} = \frac{1}{N_s} \sum_{i\in S} \Big(r_{i,t} - f_k(Z_{i,t-1}; \hat{\beta}_k) \Big)^2,\ k = 1, 2,\ t \in Te, \tag{44}$$

with each $e^2_{k,t}$ representing the cross-sectional average of forecast-squared errors at period t under M_k, $k = 1, 2$. Like in the previous case, the model parameters $\hat{\beta}_{k,t-1}$ are estimated using all data until $t-1$. Then the DM statistic to test the null of equal predictive ability is given by

$$DM = \frac{\bar{D}_{12}}{\hat{\sigma}_{D_{12}}} \sim \mathcal{N}(0,1), \tag{45}$$

where $\bar{D}_{12} = \frac{1}{N_{Te}} \sum_{t \in Te} D_{12,t}$ is the sample mean of squared forecast error differentials and $\hat{\sigma}_{D_{12}}$ is a heteroskedastic and autocorrelation robust standard error estimate of $\bar{D}_{12}$. GKX use Newey-West standard errors of squared forecast error differentials as a proxy for $\bar{D}_{12}$.[17]

Asymptotic validity of DM tests. DM emphasize that their tests (43, 45) yield asymptotically valid inferences only under the assumption that the loss differentials, $\{d_{12,t}\}\{D_{12,t}\}$, are covariance stationary. Equivalently,

$$E(d_{12,t}) = \mu_1,\ cov(d_{12,t}, d_{12,t-\tau}) = \gamma_1(\tau),\ \forall t, \tau \geq 0,\ \text{and} \tag{46}$$

$$E(D_{12,t}) = \mu_2,\ cov(D_{12,t}, D_{12,t-\tau}) = \gamma_2(\tau),\ \forall t, \tau \geq 0. \tag{47}$$

However, this assumption is violated when the parameters, such as $\{\hat{w}_{k,i,t-1}\}_{i \in S}$ and $\hat{\beta}_{k,t-1}$, are estimated from econometric models. Their estimation uncertainties introduce time-varying temporal dependencies between the loss differentials, thereby breaking down the covariance stationarity assumption. A simple intuition demonstrates the central idea.

Recall that $\{\hat{w}_{1,k,t-1}\}_{i \in S}$ are estimated using all data until $t-1$. Thus, their precision (variance) increases (decreases) as time proceeds and more data are available. Consequently, the HL return differentials exhibit time-varying moments and temporal dependencies, rendering the covariance stationarity assumption infeasible.

B. Violation of Covariance Stationarity: Empirical Evidence

Consistent with the previous intuition, appendix B.2 (table 2) empirically documents that the loss differentials computed using NN-3 and Lewellen-based return predictions significantly violate the covariance stationarity assumption. This result reaffirms that the existing DM-based conclusions are misleading.

In particular, B.2 conducts covariance stationarity tests proposed by Pagan and Schwert (1990)

[17] To be precise, the DM tests were original designed for time-series data. GKX adapted these tests on panel data by cross-sectionally averaging the forecast-squared errors at each period, as in (44). In a recent working paper, Timmermann and Zhu (2019) show that this adapted statistic yields asymptotically valid inferences, of course, only under the assumption that there is no parameter uncertainty.

on three different loss differentials over the 360 OOS months. The first comprises the forecast-squared error differences between the NN-3 and Lewellen-based return predictions. The second (third) contains the return differences between the EW (VW) HL portfolios based on the NN-3 and Lewellen models.

If these loss differentials were covariance stationary, then each of their sample standard deviations over the first 180 periods should be close to those over the last 180 periods. However, the initial period standard deviations are significantly (5, 1.85, and 1.75 times) higher than those of the final period. Thus, the null of covariance stationarity is rejected across the loss differentials. Also, relatively large beginning period standard deviations may reflect a "recursive estimation scheme", in which case parameter uncertainty decreases as time progresses, when true model parameters are time-invariant.

C. Bootstrap Tests for Out-of-Sample Comparisons

This subsection presents a bootstrap test that accommodates non-stationary loss differentials. The method builds on the moving block bootstrap procedure of Kunsch (1989). Although it was initially designed for stationary processes, Gonçalves and White (2002, 2004) establish their asymptotic validity for non-stationary processes under certain assumptions that govern the degree of non-stationarity.

First, they assume that the mean heterogeneity of the given series is not too strong. The return differentials in (42) satisfy this condition, as their unconditional means are the same.[18] Second, they assume that the series is a near epoch dependent on an underlying mixing process (Billingsley (1999)). This condition is less stringent than "mixing conditions" that researchers, including DM, typically assume to derive limiting distributions. Importantly, near epoch dependent processes allow for considerable heterogeneity (of (co)variances) and also for dependence. Thus, their assumptions suit this paper's framework.

Why bootstrap works. Recall that the DM tests make two parametric approximations.

[18] It is less clear whether forecast-squared-error differentials theoretically have the same unconditional means. However, empirical tests suggest that the null of equal means over different periods do not get rejected. This result supports the assumption laid out by Gonçalves and White (2002).

The tests use heteroskedastic and autocorrelation standard errors and draw critical values from the standard normal. Such approximations likely fail under complex scenarios (e.g., when the series is not stationary). However, bootstrap-based tests do not make such parametric simplifications and thus likely yield valid asymptotic inferences even in challenging situations. Of course, even bootstrap could fail under certain circumstances (see section 4.5 from Horowitz (2001)). Thus, the literature recommends complementary simulation checks, as described in the next subsection.

I now explicitly discuss how to conduct bootstrap-based OOS inferences.

C.1. Tests of equal return means or forecast-squared errors.

Consider a series of loss differentials $\{\Delta_t\}_{t=1}^{T}$. These could be either HL return $(d_{12,t})$ or squared forecast error differentials $(D_{12,t})$. Then the procedure for obtaining critical values, or p-values, under the null hypothesis $H_0 : E(\frac{1}{T}\sum_{t=1}^{T}\Delta_t) = 0$ is as follows.

1. Choose a block-size l. For each iteration i,
   (a) draw $n = (T/l)$ random numbers, $\{b_i\}_{i=1}^{n}$, from the set $\{1, 2, \ldots, T-l\}$ with replacement,
   (b) draw a block bootstrap sample D_i=$\{\Delta_{b_1}, \Delta_{b_1+1}, \ldots \Delta_{b_1+l-1};\ \Delta_{b_2}, \Delta_{b_2+1}, \ldots \Delta_{b_2+l-1};\ \ldots; \Delta_{b_n}, \Delta_{b_n+1}, \ldots \Delta_{b_n+l-1}\}$, where D_i contains a total number of T differentials, and
   (c) impose the null and compute the bootstrap-based t-ratio, $t_i = \left(\bar{D}_i - \bar{\Delta}\right)/std(D_i)$, where $\bar{D}_i$ and $std(D_i)$ are the sample mean and standard deviation of D_i, respectively. $\bar{\Delta}$ is the sample mean of the original loss differentials.
2. Repeat step (1) many times. The generalized p-value equals the proportion of times the absolute value of t_i is greater than the original sample's realized absolute t-ratio, which equals $t = \left(\bar{\Delta}\right)/std(\Delta)$, where $std(\Delta)$ is the sample standard deviation of the loss differentials $\{\Delta_j\}_{j=1}^{T}$.

The optimal block-size l, shown in the literature to be $O(T^{1/2})$, is close to 2 years of data on a sample over 30 years. Thus, the empirical section uses a block size of 24. However, the results are qualitatively similar across other block lengths of 6, 12, 18, and 36.

C.2. Tests of equal Sharpe ratios.

I further generalize the procedure to compare OOS Sharpe ratios of any two model-based investment strategies. Let $\{HL_{1t}\}$ and $\{HL_{2t}\}$ be two such series, with squared Sharpe ratios

$$Sh_i^2 = \frac{(\frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}{\frac{1}{T}\sum_{t=1}^{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}, \text{ for } i = 1, 2. \tag{48}$$

The p-value for testing the null of equal squared Sharpe ratios, $H_0 : E(Sh_1^2) = E(Sh_2^2)$, can be computed as follows.

1. Choose a block-size l. For each iteration i.

   (a) draw $n = (T/l)$ random numbers, $\{b_i\}_{i=1}^{n}$, from the set $\{1, 2, \ldots, T-l\}$ with replacement,

   (b) normalize the returns to impose the null,

   $$HL_{it}^* = \sqrt{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it}) / \sqrt{\sum_{t=1}^{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}, \tag{49}$$

   (c) draw a block bootstrap sample $\{H_{ki}\}$ from the normalized returns;

   $$\begin{aligned} H_{ki} =& \{HL_{k,b_1}^*, HL_{k,b_1+1}^*, \ldots HL_{k,b_1+l-1}^*; HL_{k,b_2}^*, HL_{k,b_2+1}^*, \ldots HL_{k,b_2+l-1}^*; \\ & \ldots; HL_{k,b_n}^*, HL_{k,b_n+1}^*, \ldots HL_{k,b_n+l-1}^*\} \text{ for } k = 1, 2, \text{ and} \end{aligned}$$

   (d) compute the bootstrap-based squared Sharpe ratio difference, $Sh_{1i}^2 - Sh_{2i}^2$.

   $$Sh_{ki}^2 = \frac{(\frac{1}{T}\sum_{t=1}^{T} H_{kit})^2}{\frac{1}{T}\sum_{t=1}^{T}(H_{kit} - \frac{1}{T}\sum_{t=1}^{T} H_{kit})^2}, \text{ for } k = 1, 2, \text{ where } H_{kit} = t^{th}\text{element of } H_{ki}.$$

2. Repeat step (1) many times. The p-value equals the proportion of times the absolute value of $(Sh_{1i}^2 - Sh_{2i}^2)$ is greater than the absolute value of $Sh_1^2 - Sh_2^2$.

D. Performance of the Methodology: Monte Carlo Evidence

Extensive simulations in Appendix B.3 reveal that this paper's bootstrap-based tests are well-sized. In contrast, DM-based tests lead to massive size distortions.

In particular, (B.3) (figure (3)) simulates return time series with zero means under three distinct models, each allowing for a different degree of time-varying temporal dependency. It then conducts the zero return mean tests on the simulated data using three methods that include the DM-test with OLS standard errors, the DM-test with Newey-West standard errors, and this paper's bootstrap method with a block size of 24. Across all simulations, bootstrap-based 5% level tests yield accurate sizes close to 5%. However, DM-based 5% level tests deliver hugely distorted sizes between 13% and 42%, depending on how strong the temporal dependencies are.

Figure (4) shows the power curves for the three methods and confirms that bootstrap-based test "size" refinements come at the expense of only small power losses.

5. Empirical Results

This section presents the main empirical results of the paper. Recall that the theoretical sections imply two central predictions. (1) Ex-ante precision of NN-based risk premium predictions proxy for their ex-post forecast-squared errors, and thus (2) the Confident-HL investment portfolios that deliberately exclude stocks with imprecise risk premium estimates should yield huge OOS economic gains. I empirically demonstrate both of these predictions.

A. Data, Definitions, and Replication Study

A.1. Data

The sample contains monthly excess stock returns of all individual firms listed in the NYSE, AMEX, and NASDAQ exchanges between March of 1957 and December of 2016. The data include 26667 total stocks, with an average of more than 6000 stocks per month. The data also comprise a high-dimensional set of 176 raw predictors examined by GKX and Avramov et al. (2020), including

94 individual stock characteristics analyzed by Green, Hand, and Zhang (2017) (e.g., size, book-to-market, 1-year momentum returns). Another 74 are industry-sector dummy variables based on the first two digits of the Standard Industrial Classification codes. The final eight are aggregate macroeconomic variables used by Goyal and Welch (2008).[19] The Treasury-bill rate proxies for the risk-free rate.

A.2. Models

Neural Network. The paper primarily focuses on a feed-forward NN with three hidden layers (NN-3) and 32, 16, and 18 neurons per layer. This model was examined by GKX and Avramov et al. (2020). I precisely mimic their "recursive scheme" to estimate the model parameters. The scheme first divides the data into 18 years of training (1957-1974), 12 years of validation (1975-1986), and 30 years (1987-2016) of OOS test samples. It then estimates the parameters and hyperparameters using objective functions to minimize the training sample's regularized MSE (17) and the validation sample's MSE (18), respectively. At the end of each year, it re-estimates the model parameters, increasing the training sample by one year. The validation sample rolls forward every year to include the most recent year's data, maintaining the same size (12 years).

I implement this estimation framework to obtain risk premium predictions, as well as their standard errors, over the OOS test sample. Whereas GKX and Avramov et al. (2020) mainly apply L_1 regularization to estimate the parameters, I use dropout and L_2. As discussed in section 3, this approach enhances the model's predictive performance and delivers standard errors of predictions. I retain the other hyperparameters (e.g., SGD learning rate, Adam optimization, early-stopping) used by GKX. The Internet appendix tabulates all regularizations with their values.[20]

Lewellen. To compare the economic gains from NN-3-based risk premium predictions and their standard errors with those of simple benchmark models, I also examine one of Lewellen (2015)'s linear models. This Lewellen model predicts stock returns using a pooled regression on

[19] Besides these 176 predictors, GKX and Avramov et al. (2020) also consider (94 × 8) interactions between the stock characteristics and macroeconomic variables. They do so as they examine several linear models (e.g., Lasso, Instrumented Principal Components) that do not explicitly account for variable interactions. Because NNs automatically capture such interactions, this paper excludes those additional variables.

[20] See GKX for a detailed review of these regularizations.

15 firm-level characteristics (e.g., size, book-to-market, accruals, asset growth ratio). The Internet appendix describes the exact model. This model, unlike NN-3, does not entail regularization. Thus, to make a fair assessment, I estimate the regression parameters using both training and validation data-sets. The OOS test data remain the same.

A.3. Definitions of Performance Metrics

I lay out the definitions of ex-ante and ex-post precision measures that I use repeatedly throughout the rest of the paper.

Ex-ante Confidence. I compute ex-ante confidence of stock-level risk premium predictions using their absolute t-ratios

$$EC_{it} = \frac{|\hat{r}_{i,t+1}|}{se_t(\hat{r}_{i,t+1})}, \tag{50}$$

where EC is ex-ante confidence, $\hat{r}_{i,t+1}$ is the risk premium prediction of stock i at period t (for $t+1$) and $se_t(\hat{r}_{i,t+1})$ is its ex-ante predictive standard error. $|.|$ denotes the absolute value. Ex-ante confidence proxying for a prediction's precision is consistent with the notion that an estimate's standard error must always be understood in the context of the estimate's mean. See section (C.C3) in the internet appendix for a formal discussion using a simple linear model in the spirit of the capital asset pricing model.[21] However, my central conclusions are the same when I use inverse standard errors as proxies for precision. Table B in Appendix C.C1 presents the results.

Whereas I calculate the ex-ante confidence of NN-3-based risk premium predictions using the theory derived in section 3, those of Lewellen-based predictions are available in the closed-form expressions. For example, consider a linear regression model $R = Z\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, where R and Z are panels of stock-level returns and characteristics, respectively. Given a stock i's risk premium prediction $z_i\hat{\beta}$, its standard error equals $z_i^{'}(Z^{'}Z)^{-1}z_i\hat{\sigma}^2$, where $\{\hat{\beta}, \hat{\sigma}^2\}$ are the ordinary least squares (OLS) estimates of β and σ^2, respectively. The OLS standard errors are consistent

[21] Recall that proposition-1 makes a highly stylized assumption of invariant risk premia across the stocks in the top (bottom) decile. In more realistic scenarios, this assumption does not hold, in which case I argue that considering the absolute t-ratios as proxies for the precision leads to superior performance relative to the inverse standard errors. See section (C.C3).

with the model specification of GKX, given in (2).[22]

Ex-post Out-of-Sample-R^2. Given a set of risk premium predictions $\mathcal{S}$, I compute their ex-post OOS R^2 using the following measure motivated by GKX

$$\text{OOS-}R^2 = 1 - \frac{\sum_{(i,t)\in\mathcal{S}}(r_{i,+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{S}} r_{i,t+1}^2}, \tag{51}$$

where $r_{i,t+1}$ is the realized excess return of stock i at period $t+1$.

A.4. Replication of Gu, Kelly, and Xiu (2020)

To ensure that this paper's NN-3-based risk premium measurements are comparable with GKX and Avramov et al. (2020), I replicate their studies. For every period in the OOS test sample, I sort stocks into deciles, decile-1 to decile-10, according to their NN-3-based return predictions for the next month. Decile-1 (decile-10) comprises the bottom (top) 10% of stocks with the lowest (highest) return predictions. Figure 5 (6) presents the EW (VW) average OOS returns and Sharpe ratios of the decile portfolios. All of these monotonically increase from decile-1 through decile-10, thereby confirming that the realized OOS returns align with their predictions. Furthermore, the EW (VW) HL portfolio that takes long-short positions on the extreme decile portfolios (i.e., decile-10 minus decile-1) earns an enormous OOS return of 2.51% (1.47%) and an annualized Sharpe ratio of 1.56 (0.96). These results reflect the success of NN-3 in terms of impressive economic gains. They also qualitatively and quantitatively match with GKX and Avramov et al. (2020), respectively.

Having outlined the data and showing that this paper's NN-3-based return predictions match those of the previous studies, I move on to test the theoretical predictions.

B. Ex-ante Confidence and Ex-post Out-of-Sample-R^2

I first validate remarks 1 and 2 of section 2 asserting that the ex-ante precision of NN-based risk premium predictions significantly predict their ex-post precision, whereas those of Lewellen-based

[22] Alternatively, I also consider Fama-Macbeth standard errors for Lewellen-based risk premia to account for cross-sectional correlations of residuals. The conclusions are the same.

predictions do not.

Figure 7 confirms this result for NN-3. For every month, I sort stocks into deciles according to their NN-3-based ex-ante confidence. I then calculate the OOS-R^2 attained by these decile subsamples over the 30-year OOS period. Figure 7 reveals that the ex-post OOS-R^2 monotonically increases with the level of ex-ante confidence. For example, the bottom decile, containing stocks most imprecisely predicted by NN-3, attains an OOS-R^2 of 0.81%. In contrast, the top decile with the most confident predictions delivers a much improved OOS-R^2 of 2.21%. This result reinforces that the ex-post precision of NN-based predictions is ex-ante predictable.

Table 3 further shows that these OOS-R^2 refinements translate into large economic gains. In particular, I construct EW (VW) HL portfolios on each of these confident-decile subsamples, further sorting stocks into deciles according to their next period's (NN-3-based) return predictions. Table 3 demonstrates that the EW (VW) HL portfolios formed on precise deciles earn remarkably higher OOS returns and Sharpe ratios than those formed on imprecise deciles. For example, the extremely imprecise decile's HL portfolio yields a modest 0.88% (0.34) and 0.71 (0.23) average monthly return and annualized Sharpe ratios, respectively. However, those of the most confident decile's HL portfolio are 3.10% (1.59%) and 1.44 (0.80), respectively, nearly 250% (300%) and 100% (300%) larger than the imprecise decile's counterparts.

Interestingly, the HL portfolios constructed on deciles 9 and 1 have nearly the same average return predictions. However, the average realized OOS return of the relatively more precise decile's EW (VW) HL portfolio, 2.03% (1.16%), is at least twice (thrice) more than that of the imprecise decile, 0.88% (1.16). This result is in the spirit of example-1 in section 2, which shows that between any two sets of stocks with the same risk premium levels, the HL strategy formed on the relatively precisely predicted set has higher expected returns.

Figure 8 repeats the analysis for Lewellen-based predictions and supports the theory as well. Their ex-post OOS-R^2s, unlike NN-based OOS-R^2s, do not monotonically increase with the ex-ante precision. For example, decile 10, containing the stocks with the highest ex-ante precision, has a markedly lower ex-post OOS-R^2 (0.41%) than the OOS-R^2 (0.93%) of decile 7 with relatively lower ex-ante precision. This result is consistent with remark 1, which posits that "bias" rather than "variance" predominantly determines the ex-post precision of a "simple" model-based prediction, rendering it unpredictable ex-ante. Interestingly, though, predictions from the lowest ex-ante precision decile (1) also registers awful ex-post OOS-R^2. The result perhaps reflects the decile's drastically large ex-ante "variances", which dominate average "biases" across other predictions to yield cross-sectionally higher ex-post squared forecast errors.

Overall, ex-post OOS-R^2s of Lewellen-based predictions are not as conspicuously predictable as NN-3-based OOS-R^2s. Consequently, table 3 indicates that Lewellen-based HL portfolios formed on (Lewellen-based) precise deciles do not earn significantly higher OOS returns than those on imprecise deciles. This result contrasts with the massive economic gains realized by the NN-3 HL portfolios formed on precise rather than imprecise deciles.

To summarize, this subsection demonstrated that the ex-post squared forecast errors of NN-3-based predictions are ex-ante predictable. Before moving on to show how the Confident-HL portfolios exploit this result to yield spectacular economic gains, I first describe the procedure for forming various HL portfolios.

C. Portfolio Construction

1. EW(VW)-HL. These are the conventional HL portfolios. For every month, I sort stocks into deciles according to their next month's return predictions. Let L and H represent the lowest and highest prediction deciles, respectively. Then the EW(VW)-HL portfolios take EW (VW) long and short positions on H and L, respectively.

2. EW(VW)-Confident-HL. These portfolios deliberately drop stocks with imprecise risk premium predictions from the conventional HL portfolios. In particular, both L and H are further partitioned into deciles, $\{L_1, L_2, \ldots, L_{10}\}$, and $\{H_1, H_2, \ldots, H_{10}\}$, based on their ex-ante confidence. Let L_{10} (L_1) and H_{10} (H_1) denote the subsets with the highest (lowest) ex-ante confidence from L and H, respectively. Then the EW(VW)-Confident-HL portfolios take EW (VW) long and short positions only on the highest ex-ante confident subsets, H_{10} and L_{10}, respectively.

3. EW(VW)-Low-Confident-HL. In contrast, these portfolios take EW (VW) long and short positions on the lowest ex-ante confident subsets, L_1 and H_1, respectively.

4. PW-HL. Rather than completely ignoring low ex-ante confident subsets, the "precision-weighted" strategies disproportionately downweight them while forming portfolios. In particular, the portfolios take long (short) positions on each subset H_j (L_j) with the weights proportional to $1/(11-j)$, for $j = 1, 2, \ldots, 10$. Thus, the higher a subset's precision, the more weight it has.

5. LPW-HL. In contrast, the "low-precision-weighted" portfolios take long (short) positions on each subset H_j (L_j) with the weights proportional to $1/j$.

6. Matching portfolios. To fairly assess the Confident-HL portfolios' performance, I also construct several matching strategies. These portfolios, represented by "HL_{CM}", resemble conventional HL portfolios but are matched to have the same "predicted-return" averages as those of the Confident-HL portfolios. For example, based on NN-3, the EW-Confident-HL portfolio's monthly return predictions average 1.97%. It turns out that a traditional HL strategy that takes EW long (short) positions on the top (bottom) 5% of stocks with the highest (lowest) return forecasts also has an average predicted-return of 1.97%. Thus, this strategy serves as an apt benchmark for EW-Confident-HL. The difference between the two portfolios' ex-post OOS performance precisely captures the economic value of dropping stocks with low ex-ante precision.

In general, I construct the matching portfolios as follows. Every month, EW(VW)-HL_{CM} takes long (short) positions on the top (bottom) x% of the stocks with the highest (lowest) predicted returns for the next month. I choose x so that the time-series average of EW(VW)-HL_{CM} portfolio's predicted return precisely matches that of the EW(VW)-Confident-HL portfolio.[23] Likewise, I construct the "EW(VW)-HL_{LCM}", "LPW-HL_{M}", and "PW-HL_{M}" portfolios to match the average predicted-returns of the EW(VW)-Low-Confident-HL, LPW-HL, and PW-HL, respectively.

7. Double-Sorted portfolios. As an additional robustness check, I consider various double-sorted predicted-return strategies matched to contain the same number of stocks as the Confident-HL portfolios. In particular, I partition the extreme predicted return deciles, L and H, into deciles, $\{L_{1,d}, L_{2,d}, \ldots, L_{10,d}\}$, and $\{H_{1,d}, H_{2,d}, \ldots, H_{10,d}\}$, based on their predicted-returns, respectively. Let $H_{10,d}$ ($L_{10,d}$) and $H_{1,d}$ ($L_{1,d}$) denote the subsets with the highest and lowest predicted returns

[23] Because x is determined ex-post, the matching portfolios could be interpreted as counterfactual strategies.

from H (L), respectively. Then the EW(VW)-double-sorted-HL portfolios take EW (VW) long and short positions on the highest and lowest predicted return subsets, $H_{10,d}$ and $L_{1,d}$, respectively.[24] Despite containing the same number of stocks as the Confident-HL portfolios, these strategies (unlike matching portfolios) do not serve as apt benchmarks for assessing the Confident-HL portfolios' performance because they have higher predicted-returns by construction. Nevertheless, table C in Internet Appendix C.C1 reports that the Confident-HL portfolios significantly dominate these double-sorted portfolios in terms of Sharpe ratios and information ratios.

D. Economic Gains from Confident-HL Portfolios

I now establish the dominance of the Confident-HL over the conventional HL portfolios.

D.1. OOS Average Returns and Sharpe ratios of Confident-HL Portfolios

Table 4 presents the main results. The Confident-HL and precision-weighted (PW-HL) portfolios remarkably outperform the conventional HL portfolios in terms of extensive economic measures. These measures include the OOS average realized returns, Sharpe ratios, as well as abnormal returns (α) and information ratios relative to Fama and French (2015) augmented to the momentum factor (FF-5+UMD) and Stambaugh and Yuan (2017) (SY) models. For example, the traditional EW(VW)-HL portfolio earns an impressive OOS average monthly return of 2.52% (1.48%) and an annualized Sharpe ratio of 1.5 (0.9). However, the EW(VW)-Confident-HL portfolio outperforms this strategy with the same measures of 3.61%(2.21%) and 1.75 (1.09). These are massive 43% (49%) and 17% (21%) increases, respectively. Likewise, the PW-HL also outperforms the EW-HL with an average return and Sharpe ratio of 2.87% and 1.67, respectively.

Note that the matching EW(VW)-HL_{CM} and the EW(VW)-Confident-HL portfolios have the same average NN-3-based predicted-returns. However, the former yields a considerably lower average return and Sharpe ratio than the latter. The 0.54% (0.48%) monthly return difference between the two signifies the economic value of incorporating the ex-ante precision information into forming

[24] Simply, the double-sorted portfolios take long (short) positions on stocks that have predicted risk premia higher (lower) than the top (bottom) 1% of all stocks.

NN-3-based HL portfolios. In contrast, the Low-Confident-HL and low-precision-weighted (LPW-HL) portfolios containing stocks with imprecise risk premium predictions underperform the traditional HL and Confident-HL portfolios. For example, although the EW(VW)-Low-Confident-HL portfolio has higher average predicted-returns than that of the EW(VW)-HL, it earns a drastically lower average-return and Sharpe ratio. Particularly, the VW-Low-Confident-HL strategy's annualized Sharpe ratio and the FF-6-adjusted and SY-adjusted information ratios are almost or even less than half the corresponding measures of the VW-HL portfolio. This result demonstrates the enormous imprecision of Low-Confident-HL portfolios.

Table 4 reveals that the expected returns of the EW-HL, PW-HL, and EW-Confident-HL portfolios are in increasing order, thus validating proposition-1 of section 2. Of course, all inferences drawn so far are based on the OOS point-estimates of various economic measures. To establish their statistical significance, I conduct pairwise comparisons using the moving block bootstrap tests developed in section 4.

Table 5 presents the bootstrap results, and the central conclusions are the same. The OOS annualized squared-Sharpe and squared-information ratio differences between the Confident-HL and conventional HL portfolios and between the Confident-HL portfolios and their matching HL strategies are significant at the 1% level. Likewise, the corresponding differences between the PW-HL and conventional EW-HL portfolios and between the precision-weighted HL portfolio and its matching HL strategy are also significant at 1%. Even the OOS average return and alpha differences between the Confident-HL and conventional-HL are significant at 1%. Thus, these results statistically validate the superiority of the Confident-HL portfolios.

Similarly, squared Sharpe and squared information ratios of the low-Confident-HL and low-precision-weighted-HL (LPW-HL) portfolios are significantly lower than those of their matching portfolios and the conventional HL portfolios at the 1% level. Interestingly, though, a seemingly large 0.17% monthly average return difference between the EW(VW)-HL and EW(VW)-Low-Confident-HL is statistically insignificant. Because the Low-Confident-HL portfolio returns are excessively imprecise (volatile), zero-mean comparison tests with them perhaps have less "power" to reject the null. However, Sharpe ratio tests vividly indicate the underwhelming performance of the Low-Confident-HL portfolios.

To summarize, the statistical tests distinctly reject the conventional HL portfolios in favor of the Confident-HL portfolios. As mentioned in section 4, the bootstrap tests use a block-size of 24. However, the conclusions are the same for block-lengths of 6, 12, 18, and 36.

D.2. Robustness of Confident-HL Portfolios on Non-Microcaps

In a recent working paper, Avramov et al. (2020) document that NN-3-based HL strategies primarily extract economic gains from microcap stocks. Thus, to investigate the extent to which these stocks drive the Confident-HL portfolio results, I retrain NN-3 on non-microcaps by excluding microcaps.

Table 6 presents the portfolios' OOS performance. Table 7 shows their statistical significance. Even on the non-microcap subsample, the EW Confident-HL portfolio significantly outperforms comparable alternative HL strategies. For example, the VW-Confident-HL and its matching VW-HL_{CM} have the same average predicted-returns. However, the difference between the former and latter portfolio's average monthly return is a large 0.48% (5.76% at the annual level), which is statistically significant at 5%. Likewise, the former portfolio yields a 15% higher annualized Sharpe ratio (1.00) compared with the latter (0.87), statistically distinct at the 1% level.

D.3. Robustness to Higher-Moment Risks and Transaction-Costs

Higher-Moment Risks. Because NN-3-based HL portfolios are known to display positive skewness and excess kurtosis (Avramov et al. (2020)), I also examine several higher-moment-adjusted performance measures that reflect the portfolios' downside risk. I consider Omega, Sortio, and upside-potential ratio measures that asymmetrically penalize portfolio losses more than rewarding gains, typically examined by practitioner-researchers as alternatives for Sharpe ratios.[25]

Table 8 presents the results. The Confident-HL and PW-HL handily outperform the conventional HL and equivalent matching portfolios across the higher-order measures. Thus, dropping or

[25] See the following Wikipedia pages for the definitions of these measures: Omega, Sortino, and up-side potential.

downweighing stocks with lower ex-ante precision from an investment portfolio also mitigates its downside risk.

Transaction-Costs. To evaluate whether the economic gains from the Confident-HL portfolios come at the expense of high transaction-costs, I calculate their portfolio turnovers. I find that the Confident HL-portfolios deliver impressive transaction-adjusted returns as well. The "Turnover" column of table 8 shows the portfolio turnovers, representing their average monthly percentage change in holdings. The higher the turnover, the larger the transaction costs. In fact, Avramov et al. (2020) extrapolate that a deduction of (0.005× turnover) from a portfolio's realized return roughly approximates the portfolio's transaction-cost adjusted returns.

The Confident-HL portfolio turnovers, thereby transaction costs, are significantly higher relative to the conventional HL portfolios. This result is expected, as they predominantly take long-short positions on a much smaller subset of stocks, thereby requiring more rebalancing. However, the Confident-HL portfolios' trading-cost adjusted returns are substantially larger than the conventional HL and corresponding matching portfolios. For example, the adjusted returns of the EW(VW)-Confident-HL are 2.68% (1.89%), whereas those of the EW(VW)-HL are much lower, 1.26% (0.79%), respectively.

In summary, I demonstrate that the NN-3-based Confident-HL portfolios statistically outperform the traditional HL counterparts across various economic measures. Plus, these results are robust on non-microcaps and to transaction-costs and higher-moment risks. Now, I compare these portfolios with the benchmark Lewellen-based HL portfolios.

E. Reassessing NN-3 and Lewellen Model Comparisons Using Bootstrap Tests

Recall from section 4 that the OOS model comparisons conducted by the existing studies (GKX) using the DM tests are inadequate, as they do not account for estimation uncertainty. This section reevaluates the predictive performance of NN-3 relative to the benchmark Lewellen model using the bootstrap tests. I assess the models' performance in terms of their OOS MSEs and the HL portfolios' average returns and Sharpe ratios.

E.1. NN-3 versus Lewellen: Out-of-Sample Mean Squared Error Comparisons

First, I test the null hypothesis that the MSEs of the NN-3 and Lewellen models are equal. Figure 9 presents the p-values computed using the bootstrap tests and the DM tests on various subsamples. In particular, every month, I sort stocks into deciles according to their NN-3-based risk premium predictions' ex-ante confidence, NN-3-EC. The blue line (yellow dotted-line) displays the bootstrap (DM) p-values on the subsamples that dropout 10%, 20%, ..., and 90% of the stocks with the lowest NN-3-EC, respectively. These subsamples contain the forecasts that NN-3 confidently predicts. In contrast, the red line (purple dotted-line) represents the p-values on subsamples comprising the forecasts that NN-3 imprecisely predicts.

Figure 9 reveals that the DM-based p-value is less than 0.01 on the entire OOS data comprising all stocks. Thus, consistent with GKX, the DM test rejects the Lewellen model in favor of NN-3 at the 1% level. However, with a p-value of 3.03%, the bootstrap test does not reject the null at 1% significance. Although the null of equal predictive abilitiy is rejected at the 5% significance in favor of NN-3, the difference between both p-values suggest that the DM-based tests over reject the null.

Interestingly, figure 9 illustrates that the predictive dominance of NN-3 monotonically increases with the level of ex-ante confidence. For example, dropping out 10%, 50%, and 90% of stocks with the lowest NN-3-EC significantly decreases the p-value to 2.86%, 2.24%, and 1.01%, respectively. Thus, the likelihood in favour of NN-3 increases considerably on the subsamples containing forecasts confidently predicted by NN-3. In contrast, excluding 10%, 50%, and 90% of the stocks with the highest NN-3-EC substantially increases the p-values to 4.11%, 5.72%, and 7.91%.

Of course, p-value comparisons may not provide adequate information about the models' performance on different subsamples. For example, consider the effect of changing the sample size, holding the model MSEs constant. The smaller samples would yield larger standard errors and larger p-values, although the true MSEs remain the same. Thus, to draw more informative inferences, the following subsection compares the two models in terms of their HL portfolios' OOS returns and Sharpe ratios

E.2. NN-3 versus Lewellen: High-Low Portfolio Comparisons

Fig 10 plots the OOS return and Sharpe ratio differences between both models' VW HL portfolios on various subsamples. Like in the previous figure, the economic gains from the NN-3 monotonically increase with the NN-3-EC. For example, on the entire sample containing all stocks, the difference between NN-3 and Lewellen HL portfolios' average returns (squared Sharpe ratios) is 0.38% (0.02), and statistically insignificant (at 10%). However, the difference soars to a highly significant 0.82% (0.52) on the subsample comprising the top 10% stocks with the highest NN-3-EC. In contrast, for the bottom 10% of stocks with the lowest NN-3-EC, Lewellen statistically outperforms NN-3. The average return (square-Sharpe ratio) difference between NN-3 and Lewellen HL portfolios is significantly negative -1.2% (-0.58).

Finally, I compare the conventional and Confident-HL portfolios formed from the NN-3 and Lewellen models. The portfolio definitions and notations remain the same as in section 5.C. In addition, I denote all Lewellen-based HL portfolios by attaching the subscript "$_{\text{L}}$" to HL. For example, the conventional EW-HL portfolio based on the Lewellen model is represented by EW-HL_{L}.

Table 9 presents the results. It reveals that the difference between the conventional EW (VW) NN-3-HL and Lewellen-HL portfolios' squared-Sharpe ratios is statistically insignificant at 1% (10%). Moreover, the analogous difference between the NN-3-Low-Confident-HL and Lewellen-HL is significantly negative, suggesting the Lewellen model's dominance on the subsample of forecasts imprecisely predicted by NN-3. In contrast, the corresponding difference between the NN-3-Confident-HL and Lewellen-HL portfolios is highly positive and significant at 1%. These results confirm the superiority of NN-3-based Confident-HL portfolios.

To make a fair assessment, I also compare the NN-3-Confident-HL portfolios with Lewellen-Confident-HL portfolios. The conclusions are the same. The NN-3-Confident-HL portfolios remarkably outperform in terms of squared-Sharpe ratios. This result is expected, as Confident-HL portfolios' performance hinges on ex-ante precision predicting ex-post squared forecast errors. Because it is less likely to hold for the benchmark Lewellen model (as shown in 2 and 5.B), the

Lewellen-Confident-HL portfolios do not deliver superior performance.

In sum, this section shows that existing studies significantly overestimate the overall predictive performance of NN-3 relative to the Lewellen model. The difference between the performance of both models' conventional HL portfolios' is moderately significant or insignificant. However, NN-3 exceptionally outperforms on subsamples of forecasts that it confidently predicts. Likewise, the NN-3-based Confident-HL portfolios statistically dominate the comparable Lewellen model's portfolios.

In the following two sections, I explore the time-series and cross-sectional properties of NN-3-based ex-ante precision.

F. Time-Series Variation in Ex-ante Standard Errors

To understand the time-series variation in the estimation uncertainty of NN-3-based risk premia, I compute the cross-sectional average of their ex-ante standard errors and call these "aggregate standard errors". Figure 11 plots the time-series of the aggregate standard errors. The series clearly reflects time-varying financial market uncertainty. For example, Bloom (2009) and Baker et al. (2016) document that market uncertainty appears to jump up after major shocks, such as Black Monday, the Dotcom Bubble, the Russian default, the failure of Lehman Brothers, and the 2011 debt ceiling dispute. Consistent with these studies, the aggregate standard errors spike after such shocks.

Table 10 presents the time-series average of aggregate standard errors over the OOS period and periods of shocks. Whereas the average monthly standard error across all periods is 1.06%, it is 2.31% during crisis periods. Because many individual predictors (e.g., size, price trends, and stock market volatility) in the NN-3 model substantially deviate from their usual distributions during these crisis periods, resulting risk premium predictions would also be hugely imprecise. Thus, the aggregate standard errors proxy for market uncertainty. For example, the standard errors are 38% correlated with the widely-used uncertainty proxy, the monthly market return standard deviation computed using daily data.

G. Cross-sectional Variation in Ex-ante Confidence

Table 11 presents the cross-sectional properties of various ex-ante confidence sorted deciles. It reveals that NN-3 confidently predicts stocks with small market capital, high book-to-market ratios and high 1-year momentum returns. Because these characteristics associate with higher expected returns, NN-3-based HL portfolios deliver more gains in the long-leg rather than the short-leg. This result contrasts with the "arbitrage asymmetry" studies that argue, under trading frictions, anomaly-based investment portfolios yield relatively more profits in the short-leg (e.g., Stambaugh et al. (2012)). Avramov et al. (2020) note similar observations, albeit examining *ex-post* OOS long-leg and short-leg returns of investment portfolios based on various ML models, including NN-3. Possible reasons for understanding the association between the level and precision of NN-based risk premium predictions warrant a future study.

Moreover, NN-3 confidently predicting risk premia of small-sized stocks lends support to Avramov et al. (2020), who argue that NN-3-based HL portfolios derive more economic gains from microcaps. Table 11 shows why. Because such stock risk premia are more confidently predicted, HL portfolios containing microcaps yield relatively larger economic gains.

Interestingly, I find that a significant proportion of non-microcaps have confidently risk premium predictions. Table 12 presents the results. It shows that 34% of the stocks with the most precise risk premium predictions have market caps greater than the median size across all individual stocks. Thus, NN-3-based Confident-HL portfolios yield impressive gains even on sub-samples containing large-sized stocks.

6. Conclusions

I develop an easy-to-implement method to estimate ex-ante standard errors of risk premium predictions from neural networks. To my knowledge, this is the first paper to explicitly derive the precision of NN-based risk premia at the stock-level and portfolio-level. I show that considering ex-ante standard errors leads to enhanced investment portfolios and out-of-sample statistical inferences.

The neural-network-based confident high low trading strategies that take long-short positions on stocks that have more risk premium estimates yield at least 40% higher returns and 15% higher Sharpe ratios than the neural-network-based conventional high-low portfolios. In evaluating whether these improvements are statistically significant, this paper shows that existing out-of-sample inferences that do not account for ex-ante standard errors are inadequate. I develop a bootstrap method, robust to estimation uncertainty, to compare OOS returns and Sharpe ratios of any two model-based investment strategies. The method also can be employed to compare mean squared errors of any two competing return predictions.

The bootstrap tests suggest that the neural-network-based confident high-low portfolios significantly outperform the neural-network-based conventional high-low portfolios, as well as the traditional high-low and confident high-low portfolios formed using the benchmark Lewellen model. However, the difference between the conventional neural-network-based and Lewellen-based high-low portfolios' out-of-sample returns and Sharpe ratios are either statistically insignificant or moderately significant. Thus, considering ex-ante standard errors is necessary for both real-time trading strategies and ex-post out-of-sample inferences.

A. Appendix: Proofs

1. Proof of Proposition-1:

Let the risk premium predictions of A_1, A_2, A_3, and A_4 be $\hat{a}_1$, $\hat{a}_1$, $\hat{b}_1$, and $\hat{b}_2$, respectively. Let pse_{a1}, pse_{a2}, pse_{b1}, and pse_{b2} be the predictive standard errors of A_1, A_2, A_3, and A_4, respectively.

The expected HL return equals the sum of the following measures

$$\begin{aligned} E(HL) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+ (\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) \\ &+ 0 \times p_3, \end{aligned} \tag{52}$$

where $p_3 = 1 - P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) - P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right)$.

Case1: When $pse_{a1} \geq \{pse_{b1}, pse_{b2}\}$ and $pse_{a2} \geq \{pse_{b1}, pse_{b2}\}$, the expected Confident-HL return equals

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+ (\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) \\ &+ 0 \times p_3 \\ &= E(HL). \end{aligned} \tag{53}$$

Case2: Similarly, when $pse_{b1} \geq \{pse_{a1}, pse_{a2}\}$ and $pse_{b2} \geq \{pse_{a1}, pse_{a2}\}$

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+ (\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) \\ &+ 0 \times p_3 \\ &= E(HL). \end{aligned} \tag{54}$$

Case3: When predictive standard errors do not align with either case1 or case2, without loss of generality, let $pse_{a1} \leq pse_{b1} \leq pse_{a2} \leq pse_{b2}$. Then,

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+(\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) + (\mu_a - \mu_b) \times p_4 + (\mu_b - \mu_a) \times p_5, \end{aligned} \tag{55}$$

where $p_4 = P\left(\{\hat{a}_1, \hat{b}_2\} \in Q_L\right)$, $p_5 = P\left(\{\hat{a}_2, \hat{b}_1\} \in Q_L\right)$, and $P(.)$ is the probability measure. Because $\hat{a}_1$ and $\hat{b}_1$ are (relatively) precisely measured, $\hat{a}_1$ and $\hat{b}_2$ are more likely to be in Q_L and Q_S, respectively. Consistent with this intuition, it turns out that $p_4 > p_5$. Thus,

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+(\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) + (\mu_a - \mu_b) \times (p_4 - p_5) \\ &> E(HL) \end{aligned} \tag{56}$$

Similarly, the expected return of PW-HL is given by

$$\begin{aligned} E(\text{PW-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+(\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) + (2w - 1) \times (\mu_a - \mu_b) \times (p_4 - p_5), \end{aligned} \tag{57}$$

where w (> 0.5) is the weight assigned to the precise stock in each quantile. When $w = 1$, PW-HL reduces to Confident-HL, as it takes long (short) position only on the stock with the precise risk premium prediction. Thus,

$$E(\text{HL}) \leq E(\text{PW-HL}) \leq E(\text{Confident-HL}) \tag{58}$$

2. Proof of Proposition-2

Proof. Using Gal and Ghahramani (2016), the following expressions are directly obtained for the (approximated) Bayesian marginal predictive distribution of returns and their variances, respectively.

$$\begin{aligned}
&Q(r^*_{i,t+1}|z^*_{it}, R, Z) = P(r_{i,t+1}|z^*_{it}, R, Z, \Omega)q(\Omega) \\
&q(\Omega) = \prod_{k=1}^{K} p_{i,k}, \text{ where each } p_{i,k} \sim Bern(p), \\
&P(r_{i,t+1}|z^*_{it}, R, Z, \Omega) = \mathcal{N}(\hat{E}_{i,\Omega,t}, \sigma^2_\eta I),
\end{aligned} \tag{59}$$

where $Bern()$ represents Bernoulli distribution. $\hat{E}_{i,\Omega,t}$ is given by (25), with d replaced by Ω. And

$$Var\left[Q(r^*_{i,t+1}|z^*_{it}, R, Z)\right] \approx \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2 + \sigma^2_\eta \tag{60}$$

Denote $Var\left[Q(r^*_{i,t+1}|z^*_{it}, R, Z)\right]$ by $V_Q(r^*_{i,t+1})$, where V_Q represents the variance operation under the probability distribution $Q(r^*_{i,t+1}|z^*_{it}, R, Z)$. Note that by the law of total variance

$$V_Q(r^*_{i,t+1}) = V_Q(E(r^*_{i,t+1}|W_1, W_2)) + E_Q(V(r^*_{i,t+1}|W_1, W_2)), \tag{61}$$

where W_1, W_2 are the unknown weight matrices of the NN-1, and E_Q represents the expectation operation under the probability distribution $Q(r^*_{i,t+1}|z^*_{it}, R, Z)$.

(61) further implies that

$$V_Q(r^*_{i,t+1}) = V_Q(\mu^*_{i,t}) + \sigma^2_\eta, \tag{62}$$

because $E(r^*_{i,t+1}|W_1, W_2) = \mu^*_{i,t}$, and $V(r^*_{i,t+1}|W_1, W_2) = \sigma^2_\eta$, which is assumed to be known.

Thus, (60) and (61) implies

$$V_Q(\mu^*_{i,t}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2. \tag{63}$$

□

3. Proof of Proposition-3

Proof. To compute portfolio-level standard errors, joint (approximated) density of return predictions are required. Straightforward algebra implies that it is given by

$$Q(r^*_{1,t+1}, r^*_{2,t+1}, \ldots r^*_{S,t+1} | \{z^*_{it}\}_{i=1}^S, R, Z) = P(r^*_{1,t+1}, r^*_{2,t+1}, \ldots r^*_{S,t+1} | \{z^*_{it}\}_{i=1}^S, R, Z, \Omega) q(\Omega)$$

$$q(\Omega) = \prod_{k=1}^{K} p_{i,k}, \text{ where each } p_{i,k} \sim Bern(p),$$

$$P(r^*_{1,t+1}, r^*_{2,t+1}, \ldots r^*_{S,t+1} | \{z^*_{it}\}_{i=1}^S, R, Z, \Omega) = \mathcal{N}(\hat{E}_{S,\Omega,t}, \sigma^2_\eta I), \text{ where } \hat{E}_{S,\Omega,t} = \begin{bmatrix} \hat{E}_{1,\Omega,t} \\ \hat{E}_{2,\Omega,t} \\ \vdots \\ \hat{E}_{S,\Omega,t} \end{bmatrix}, \tag{64}$$

with each $\hat{E}_{i,\Omega,t}$ given by (25). The key is to use the same Ω across the stocks, as discussed in the main section of the paper. Then, the predictive variance of the portfolio P is given by

$$V_Q(r^*_{P,t+1}) = E_Q\left(V(r^*_{P,t+1}|\Omega)\right) + V_Q\left(E(r^*_{P,t+1}|\Omega)\right), \tag{65}$$

where $r^*_{P,t+1} = \sum_{i \in S} \omega_{P,i,t} r^*_{i,t+1}$. Moreover, $V(r^*_{P,t+1}|\Omega) = \sum_{i \in S} \omega^2_{P,i,t} \sigma^2_\eta$. And due to (64), $V_Q\left(E(r^*_{P,t+1}|\Omega)\right)$ can be approximated by

$$V_Q\left(E(r^*_{P,t+1}|\Omega)\right) \approx \frac{1}{D} \sum_{d=1}^{D} \left(\hat{E}_{P,d,t} - \frac{1}{D} \sum_{d=1}^{D} \hat{E}_{P,d,t} \right)^2, \tag{66}$$

with $\hat{E}_{P,d,t}$, and $p_{1,d}$, $p_{2,d}$ given in (27).

Thus, (65) further implies that

$$V_Q(r^*_{P,t+1}) = \sum_{i \in S} \omega^2_{P,i,t} \sigma^2_\eta + \frac{1}{D} \sum_{d=1}^{D} \left(\hat{E}_{P,d,t} - \frac{1}{D} \sum_{d=1}^{D} \hat{E}_{P,d,t} \right)^2. \tag{67}$$

Now, to compute the predictive variance of P's risk premium, note that

$$V_Q(r^*_{P,t+1}) = E_Q\left(V(r^*_{P,t+1}|W_1, W_2)\right) + V_Q\left(E(r^*_{P,t+1}|W_1, W_2)\right) = \sum_{i \in S} \omega^2_{P,i,t}\sigma^2_\eta + V_Q(\mu^*_{P,t}). \quad (68)$$

Thus, from (67) and (68),

$$V_Q(\mu^*_{P,t}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2 \quad (69)$$

□

B. Appendix: Simulations and Testing the Diebold and Mariano (2002) Assumption

1. Validity of Standard Errors: Monte Carlo Evidence

Table 1
Calibration of the Confidence Intervals: Monte Carlo Evidence
This table validates the proposed standard errors using Monte Carlo simulations. The data comprise monthly stock risk premia and their raw predictors simulated under four different models 1-4. On the simulated data, confidence intervals (CIs) of various levels are constructed using NN-based risk premium predictions and their standard errors. Each row presents the confidence level and probabilities with which the corresponding level's confidence intervals cover the true simulated risk premia under the four models.

	Probability that CI contains true risk premium			
Confidence level	Model 1	Model 2	Model 3	Model 4
1%	1.26%	1.49%	1.08%	0.91%
5%	6.23%	6.65%	4.64%	3.63%
10%	11.81%	13.16%	8.98%	7.57%
20%	23.83%	26.26%	17.78%	16.17%
50%	48.72%	61.62%	46.85%	43.64%
60%	57.73%	73.10%	59.38%	55.52%
80%	78.94%	90.73%	83.60%	79.66%
90%	90.24%	96.48%	93.72%	90.36%
95%	96.03%	98.56%	97.39%	95.20%
99%	99.33%	99.74%	99.36%	98.75%

2. Tests of Covariance Stationarity

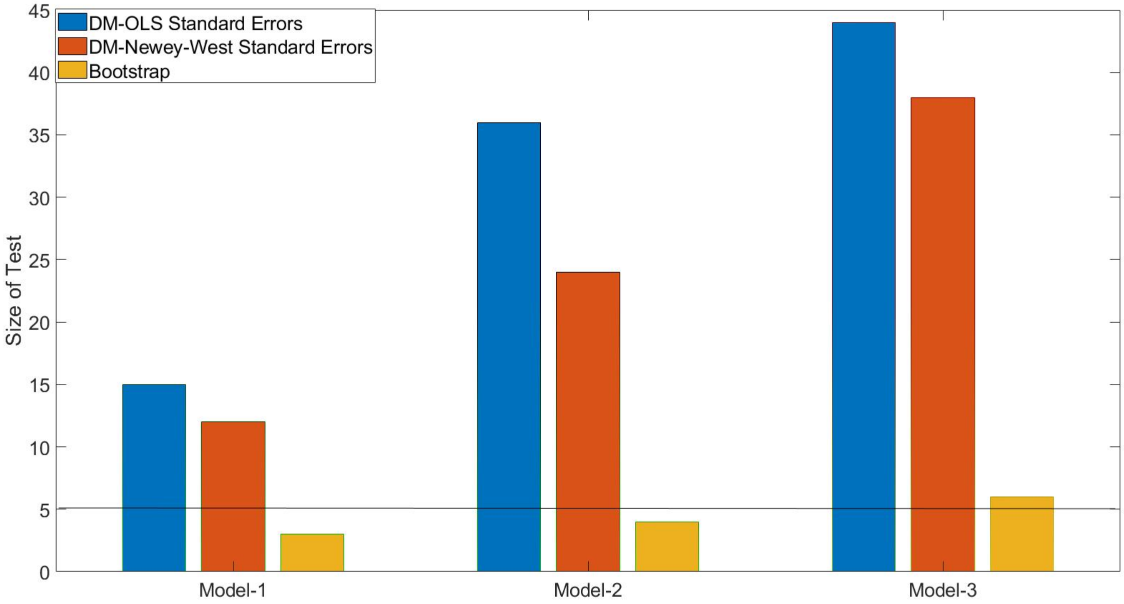
Table 2
Violation of Diebold and Mariano (2002) conditions : Non-Stationarities due to Estimation Uncertainty

This table shows that the model-based loss differentials violate the covariance stationarity assumption required for the Diebold and Mariano (2002) tests' asymptotic validity. The table presents three loss differential series over the 360 out-of-sample periods. The first comprises the forecast-squared error differences between the NN-3 and Lewellen-based return predictions. The second contains the return differences between the equal-weighted high-low portfolios based on the NN-3 and Lewellen-based models. The third includes the return differences between the value-weighted high-low portfolios based on the NN-3 and Lewellen-based models. The First 180 Months column presents the loss differentials' sample standard deviations over the first 180 OOS periods, whereas the Last 180 Months column shows those over the last 180 periods. The Ratio column presents the ratio of the first and last 180 month standard deviations. The *p*-value column presents the *p*-value under the hypothesis that the ratio equals one, with critical values based on Pagan and Schwert (1990).

(NN-3 − Lewellen) Differentials	Standard Deviation of Loss Function			
Loss Function	First 180 Months	Last 180 Months	Ratio	p-value
Mean Squared Forecast Errors	0.12%	0.02%	5.03	< 0.001
Equal-weighted High-low Returns	0.35%	0.19%	1.85	< 0.001
Value-weighted High-low Returns	0.47%	0.27%	1.75	< 0.001

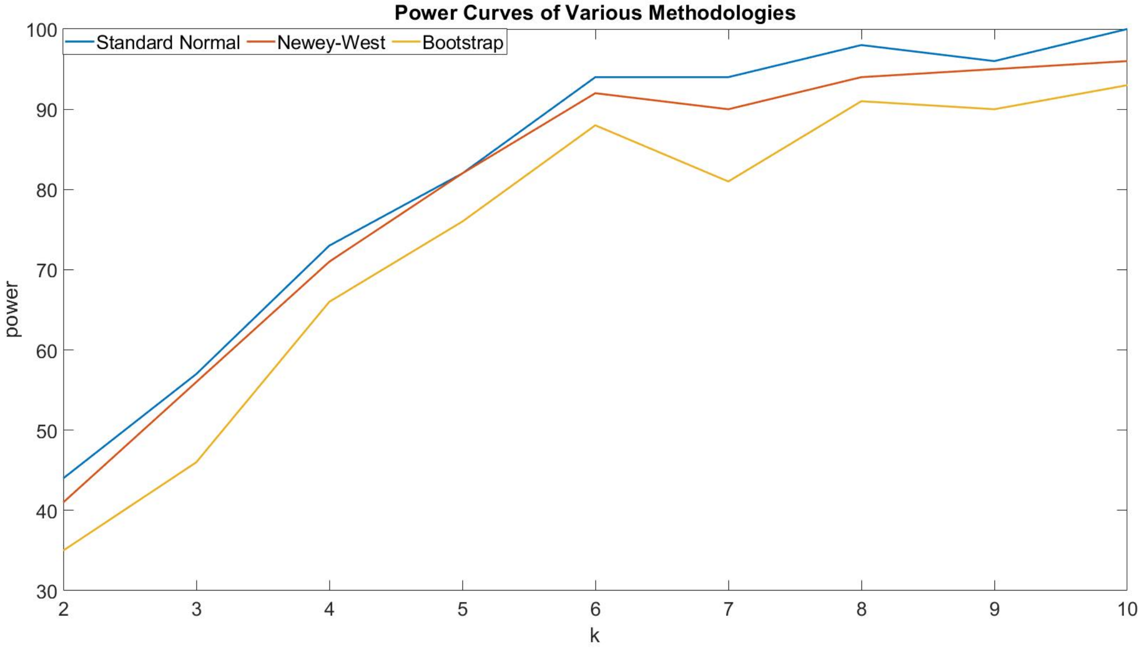
3. Performance of this paper's OOS Comparison Method: Monte Carlo Evidence

Figure 3. Test Sizes of OOS Comparison Methodologies



Note: This figure presents the "test sizes" of various methodologies at the 5% level. Test size represents the probability of *incorrectly rejecting* the null when it is true. Return time series with zero means are simulated under three distinct models, each imposing a different degree of time-varying temporal dependency. On the simulated data, tests of zero return means are conducted using three methods. The first (in blue) performs DM tests with the OLS standard errors. The second (in red) executes DM tests with Newey-West standard errors. The third (in orange) implements this paper's bootstrap method.

Figure 4. Power Curves of OOS Comparison Methodologies



Note: This figure presents the "power curves" of various methodologies at the 5% level. Power represents the probability of *correctly rejecting* the null when it is not true. Return time series are simulated under nine models, denoted by k, allowing for time-varying temporal dependencies. The mean return under model k equals k$\times\sigma$, where σ is a known scalar calibrated to match the standard deviation of the market risk premium. On the simulated data, tests of zero return means are conducted using three methods. The first (in blue) performs DM tests with the OLS standard errors. The second (in red) executes DM tests with Newey-West standard errors. The third (in orange) implements this paper's bootstrap method.

References

Allena, Rohit, 2020a, Comparing Asset Pricing Models with Non-Traded Factors and Principal Components, *SSRN Electronic Journal* .

Allena, Rohit, 2020b, Industry Costs of Equity with Machine Learning, *Working Draft, Goizueta Business School* .

Allena, Rohit, and Tarun Chordia, 2020, True Liquidity and Equilibrium Prices: US Tick Pilot, *Working Paper, Goizueta Business School* .

Avramov, Doron, Si Cheng, and Lior Metzker, 2020, Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability, SSRN Scholarly Paper ID 3450322, Social Science Research Network, Rochester, NY.

Avramov, Doron, Tarun Chordia, Gergana Jostova, and Alexander Philipov, 2013, Anomalies and financial distress, *Journal of Financial Economics* 108, 139–159.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, 2016, Measuring Economic Policy Uncertainty*, *The Quarterly Journal of Economics* 131, 1593–1636.

Bali, Turan G., Amit Goyal, Dashan Huang, Fuwei Jiang, and Quan Wen, 2020, The Cross-Sectional Pricing of Corporate Bonds Using Big Data and Machine Learning, SSRN Scholarly Paper ID 3686164, Social Science Research Network, Rochester, NY.

Berger, James O., 1985, *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics, second edition (Springer-Verlag, New York).

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni, 2020, Bond Risk Premiums with Machine Learning, *The Review of Financial Studies* .

Billingsley, Patrick, 1999, *Convergence of Probability Measures*, second edition (Wiley).

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe, 2017, Variational Inference: A Review for Statisticians, *Journal of the American Statistical Association* 112, 859–877.

Bloom, Nicholas, 2009, The Impact of Uncertainty Shocks, *Econometrica* 77, 623–685.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2020, Deep Learning in Asset Pricing, SSRN Scholarly Paper ID 3350138, Social Science Research Network, Rochester, NY.

Chinco, Alex, Adam D. Clark-Joseph, and Mao Ye, 2019, Sparse Signals in the Cross-Section of Returns, *The Journal of Finance* 74, 449–492.

Diebold, Francis X., 2015, Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests, *Journal of Business & Economic Statistics* 33, 1–1.

Diebold, Francis X, and Roberto S Mariano, 2002, Comparing Predictive Accuracy, *Journal of Business & Economic Statistics* Vol.20(1), p.134-144.

Fama, Eugene F., and Kenneth R. French, 1997, Industry costs of equity, *Journal of Financial Economics* 43, 153–193.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Gal, Yarin, and Zoubin Ghahramani, 2016, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, *Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016* 10.

Gonçalves, Sílvia, and Halbert White, 2002, The bootstrap of the mean for dependent heterogeneous arrays, *Econometric Theory* 18, 1367–1384.

Gonçalves, Sílvia, and Halbert White, 2004, Maximum likelihood and the bootstrap for nonlinear dynamic models, *Journal of Econometrics* 119, 199–219.

Gonçalves, Sílvia, and Halbert White, 2005, Bootstrap Standard Error Estimates for Linear Regression, *Journal of the American Statistical Association* 100, 970–979.

Goyal, Amit, and Ivo Welch, 2003, Predicting the Equity Premium with Dividend Ratios, *Management Science* 49, 639–654, Publisher: INFORMS.

Goyal, Amit, and Ivo Welch, 2008, A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, *The Review of Financial Studies* 21, 1455–1508.

Green, Jeremiah, John R. M. Hand, and X. Frank Zhang, 2017, The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns, *The Review of Financial Studies* 30, 4389–4436.

Géron, Aurélien, 2019, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, second edition (O'Reilly Media, Inc.).

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies* 33, 2223–2273.

Horowitz, Joel L., 2001, Chapter 52 - The Bootstrap, in James J. Heckman, and Edward Leamer, eds., *Handbook of Econometrics*, volume 5, 3159–3228 (Elsevier).

Kunsch, Hans R., 1989, The Jackknife and the Bootstrap for General Stationary Observations, *The Annals of Statistics* 17, 1217–1241.

Kyung, Minjung, Jeff Gill, Malay Ghosh, and George Casella, 2010, Penalized regression, standard errors, and Bayesian lassos, *Bayesian Analysis* 5, 369–411.

Lewellen, Jonathan, 2015, The Cross-section of Expected Stock Returns, *Critical Finance Review* 4, 1–44.

Pagan, Adrian R., and G. William Schwert, 1990, Testing for covariance stationarity in stock market data, *Economics Letters* 33, 165–170.

Pástor, Ľuboš, and Robert F. Stambaugh, 1999, Costs of Equity Capital and Model Mispricing, *The Journal of Finance* 54, 67–121.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 2014, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research* Vol.15, 1929–1958.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan, 2012, The short of it: Investor sentiment and anomalies, *Journal of Financial Economics* 104, 288–302.

Stambaugh, Robert F., and Yu Yuan, 2017, Mispricing Factors, *The Review of Financial Studies* 30, 1270–1315.

Timmermann, Allan, and Yinchu Zhu, 2019, Comparing Forecasting Performance with Panel Data, *SSRN Electronic Journal* .

Wang, Yixin, and David M. Blei, 2019, Frequentist Consistency of Variational Bayes, *Journal of the American Statistical Association* 114, 1147–1161.

Zhu, Lingxue, and Nikolay Laptev, 2017, Deep and Confident Prediction for Time Series at Uber, in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 103–110, ISSN: 2375-9259.

Figure 5. Out-of-Sample (OOS) Performance of Equal-weighted Deciles Based on NN-3 Predictions.

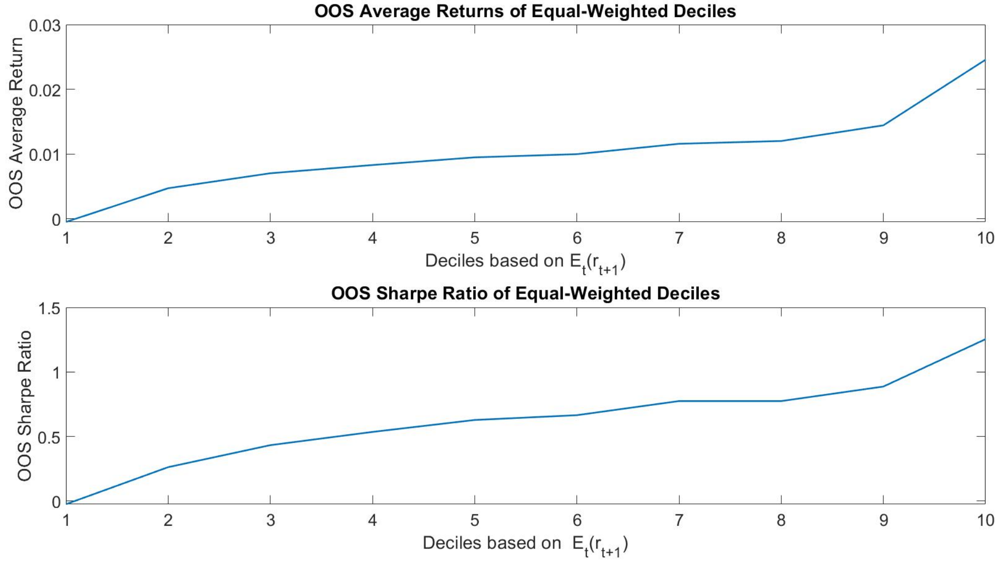
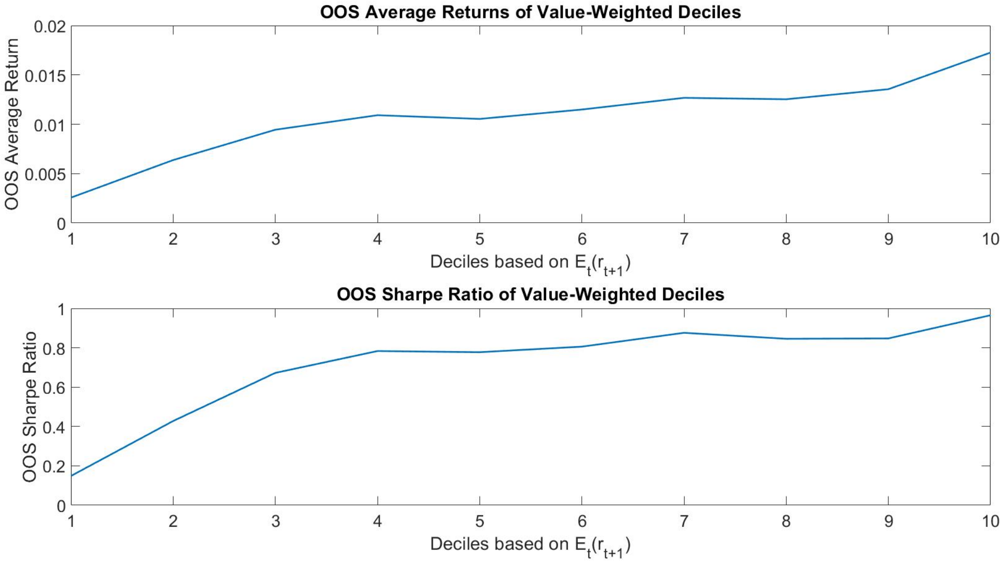
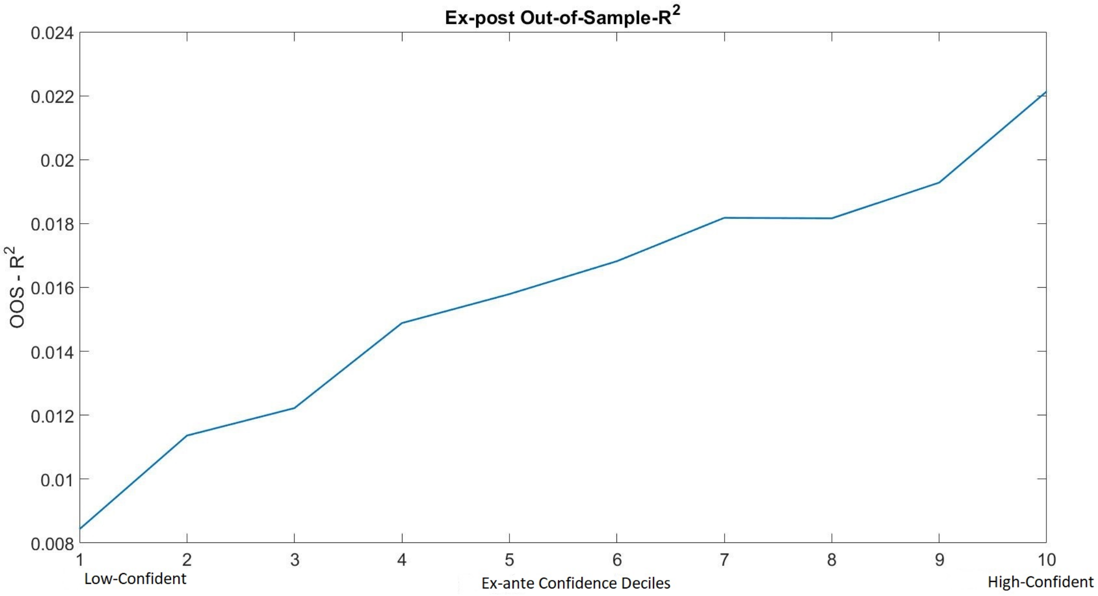


Figure 6. Out-of-Sample (OOS) Performance of Value-weighted Deciles Based on NN-3 Predictions.



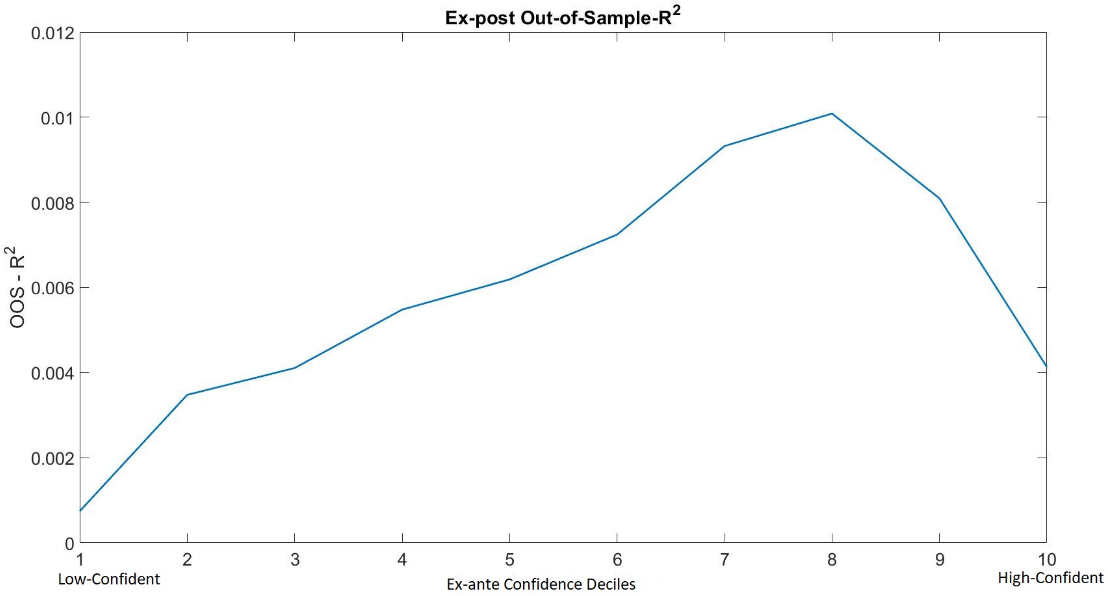
Note: Figure 5 (6) presents the performance of equal-weighted (value-weighted) prediction-sorted portfolios over the 30-year out-of-sample. At each period, stocks are sorted into deciles according to their NN-3-based risk premium predictions. Decile-10 (decile-1) comprises the top (bottom) 10% stocks with the lowest (highest) return predictions. The top figure shows the average monthly returns of each decile, whereas the bottom represents their annualized Sharpe ratios.

Figure 7. Ex-ante Confidence and Ex-post OOS-R^2: NN-3-based Predictions and Standard Errors



Note: This figure presents the out-of-sample (OOS) R^2s of various ex-ante confidence-sorted subsamples over the 30-year test sample. At each period, stocks are sorted into deciles according to their NN-3-based risk premium predictions' ex-ante confidence (EC). Decile-10 (decile-1) comprises the top (bottom) 10% stocks with the lowest (highest) precision. The y-axis represents the ex-post OOS R^2s attained by the decile subsamples.

Figure 8. Ex-ante Confidence and Ex-post OOS-R^2: Lewellen-based Predictions and Standard Errors



Note: This figure presents the out-of-sample (OOS) R^2s of various ex-ante confidence-sorted subsamples over the 30-year test sample. At each period, stocks are sorted into deciles according to their Lewellen-based risk premium predictions' ex-ante confidence. Decile-10 (decile-1) comprises the top (bottom) 10% of stocks with the lowest (highest) precision. The y-axis represents the ex-post OOS R^2s attained by the decile subsamples.

Table 3

Long-short Portfolios' Performance on Subsamples with Different Levels of Ex-ante Confidence

This table reports the performance of model-based high-low (HL) portfolios over the 30-year out-of-sample (OOS) period on various subsamples. Each period, stocks are first sorted into deciles according to their ex-ante confidence levels of model-based risk premium predictions. On each decile, equal-weighted (value-weighted) HL portfolios are formed by further sorting stocks into deciles according to their next month's model-based return predictions and taking long-short positions on the extreme deciles. The NN-3-HL and Lewellen-HL columns present each precision-decile's HL portfolio's performance under the NN-3 and Lewellen models, respectively. The Pred Ret column reports the HL portfolio's average return predictions. The Avg Ret, Std, Sharpe columns respectively represent the average, standard deviation, and Sharpe ratio of the HL portfolio's realized returns. Panels A and B present the equal-weighted and value-weighted strategies, respectively.

Panel A: Performance of equal-weighted-HL on various precision-sorted subsamples

	NN-3-HL				Lewellen-HL			
Precision decile	Pred Ret	Avg Ret	Std	Sharpe	Pred Ret	Avg Ret	Std	Sharpe
1 (Low-Confident)	0.72%	0.88%	4.29%	0.71	1.83%	0.81%	6.85%	0.41
2	0.52%	1.14%	4.80%	0.83	2.97%	1.89%	6.53%	1.00
3	0.54%	0.75%	4.62%	0.56	2.30%	1.53%	6.88%	0.77
4	0.58%	1.31%	4.74%	0.96	1.83%	1.80%	7.60%	0.82
5	0.62%	1.31%	5.15%	0.88	1.62%	1.70%	7.02%	0.84
6	0.64%	1.77%	5.42%	1.13	1.49%	1.44%	5.99%	0.83
7	0.66%	1.40%	5.44%	0.89	1.49%	1.93%	6.12%	1.09
8	0.68%	1.78%	5.59%	1.10	1.50%	1.53%	5.27%	1.01
9	0.71%	2.03%	7.43%	0.95	1.43%	2.01%	4.99%	1.40
10 (High-Confident)	0.88%	3.10%	7.48%	1.44	1.07%	1.42%	4.90%	1.00

Panel B: Performance of value-weighted-HL on various precision-sorted subsamples

	NN-3-HL				Lewellen-HL			
Precision decile	Pred Ret	Avg Ret	Std	Sharpe	Pred Ret	Avg Ret	Std	Sharpe
1 (Low-Confident)	0.70%	0.34%	5.12%	0.23	1.79%	1.00%	5.46%	0.64
2	0.49%	0.65%	5.82%	0.39	2.89%	1.27%	8.64%	0.51
3	0.52%	0.86%	5.60%	0.53	2.16%	1.57%	7.39%	0.74
4	0.56%	0.65%	5.21%	0.43	1.76%	1.07%	6.39%	0.58
5	0.60%	0.80%	5.55%	0.50	1.45%	1.06%	6.30%	0.58
6	0.62%	0.68%	5.59%	0.42	1.32%	1.01%	5.43%	0.64
7	0.62%	0.43%	6.02%	0.25	1.29%	1.27%	5.40%	0.82
8	0.67%	0.67%	6.52%	0.36	1.35%	1.13%	5.11%	0.77
9	0.70%	1.16%	7.68%	0.52	1.33%	1.33%	5.98%	0.77
10 (High-Confident)	0.89%	1.59%	6.86%	0.80	0.99%	0.66%	5.93%	0.39

Table 4
Performance of Confident and Low-Confident Long-Short Portfolios: All Stocks

This table reports the performance of various NN-3-based long-short portfolios over the 30-year out-of-sample (OOS) period. EW(VW)-HL represents the traditional equal(value)-weighted long-short portfolio. EW(VW)-Confident-HL and EW(VW)-Low-Confident-HL denote the equal(value)-weighted Confident and Low-Confident long-short portfolios that only include stocks with the most *confident* and *imprecise* risk premium predictions, respectively. LPW-HL and PW-HL are the "imprecision" and "precision" weighted portfolios that overweight stocks with imprecise and precise return predictions, respectively. EW(VW, LPW)-HL$_{\mathrm{LCM}}$ is the conventional EW(VW, LPW) HL portfolio matched to have the same average predicted returns as that of the EW-Low-Confident-HL (EW-Low-Confident-HL, LPW-HL) portfolio. EW(VW)-HL$_{\mathrm{CM}}$ is a traditional EW(VW)-HL portfolio matched to have the same average predicted returns as that of the EW-Confident-HL (VW-Confident-HL) portfolio. Likewise, LPW(PW)-HL$_{\mathrm{M}}$ is a traditional EW-HL portfolio matched with LPW(PW)-HL. See section 5.C for a detailed description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD) and Stambaugh-Yuan 4-factor (SY) models. The "pred ret" column represents the average predicted returns. The "avg ret" column shows the average realized returns. The "α" columns indicate abnormal returns. The "t" columns denote the t-stats of "average returns" and "α". The "SR" and "IR" columns represent the annualized Sharpe and Information ratios, respectively.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL$_{\mathrm{LCM}}$, HL$_{\mathrm{CM}}$ and HL$_{\mathrm{M}}$ are matching high-low portfolios.

Panel A: Equal-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	1.69%	2.52%	8.21	1.50	2.20%	7.63	1.39	2.18%	7.15	1.31
EW-HL$_{\mathrm{LCM}}$	1.77%	2.64%	8.20	1.50	2.34%	7.7	1.41	2.33%	7.25	1.32
EW-Low-Confident-HL	1.79%	2.35%	6.46	1.18	1.97%	5.65	1.03	1.96%	5.28	0.96
EW-HL$_{\mathrm{CM}}$	1.97%	3.07%	8.65	1.58	2.77%	8.26	1.51	2.75%	7.8	1.42
EW-Confident-HL	1.97%	3.61%	9.58	1.75	3.29%	9.02	1.65	3.27%	8.6	1.57

Panel B: Value-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
VW-HL	1.62%	1.48%	4.95	0.90	0.90%	3.26	0.59	0.77%	2.68	0.49
VW-HL$_{\mathrm{LCM}}$	1.77%	1.50%	4.61	0.84	0.87%	2.87	0.52	0.76%	2.38	0.44
VW-Low-Confident-HL	1.78%	1.31%	3.02	0.55	0.48%	1.15	0.21	0.39%	0.88	0.16
VW-HL$_{\mathrm{CM}}$	1.90%	1.73%	4.92	0.90	1.12%	3.39	0.62	1.02%	2.95	0.54
VW-Confident-HL	1.90%	2.21%	5.95	1.09	1.79%	4.77	0.87	1.43%	3.82	0.70

Panel C: Precision-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	1.69%	2.52%	8.21	1.50	2.20%	7.63	1.39	2.18%	7.15	1.31
LPW-HL$_{\mathrm{M}}$	1.69%	2.52%	8.21	1.50	2.20%	7.63	1.39	2.18%	7.15	1.31

Table 5
Statistical Comparison of Long-Short Portfolios: All Stocks
This table conducts pairwise statistical comparisons of the out-of-sample (OOS) performance of various NN-3-based long-short portfolios. The tests are based on the moving block bootstrap procedure developed in section 4, with a block-length of 24. The Investment Strategy column shows the comparing pair of portfolios. The avg ret column presents the average return differences between the pair of investment strategies. The α column shows the average abnormal return differences. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios, respectively. The numbers in parenthesis are p-values. *, ** and *** denote significance at the 1%, 5% and 10% levels, respectively. See table 4 and section 5.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL_{LCM}, HL_{CM} and HL_{M} are matching high-low portfolios.

Panel A : OOS Performance Differences of Equal-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-Low-Confident-HL	0.17% (0.373)	0.859*** (0)	0.23% (0.207)	1.008*** (0)	0.22% (0.267)	0.941*** (0)
$\text{EW-HL}_{\text{LCM}}$ − EW-Low-Confident-HL	0.30% (0.142)	0.853*** (0)	0.36%* (0.06)	1.049*** (0)	0.36%* (0.083)	0.998*** (0)
EW-Confident-HL − EW-HL	1.10%*** (0)	0.808*** (0)	1.09%*** (0)	0.884*** (0)	1.09%*** (0)	0.92*** (0)
EW-Confident-HL − EW-Low-Confident-HL	1.27%*** (0.001)	1.666*** (0)	1.32%*** (0)	1.892*** (0)	1.31%*** (0.001)	1.861*** (0)
EW-Confident-HL − EW-HL_{CM}	0.55%** (0.03)	0.563*** (0)	0.52%** (0.039)	0.502*** (0)	0.52%** (0.043)	0.527*** (0)

Panel B : OOS Performance Differences of Value-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
VW-HL − VW-Low-Confident-HL	0.17% (0.542)	0.511*** (0.001)	0.42%* (0.094)	0.356*** (0.001)	0.38% (0.136)	0.258*** (0.002)
$\text{VW-HL}_{\text{LCM}}$ − VW-Low-Confident-HL	0.19% (0.503)	0.404*** (0.002)	0.39% (0.144)	0.266*** (0.003)	0.37% (0.173)	0.198*** (0.008)
VW-Confident-HL − VW-HL	0.73%*** (0.003)	0.364*** (0.003)	0.89%*** (0)	0.467*** (0)	0.66%*** (0.007)	0.3*** (0.001)
VW-Confident-HL − VW-Low-Confident-HL	0.90%** (0.032)	0.875*** (0)	1.31%*** (0)	0.823*** (0)	1.04%*** (0.009)	0.558*** (0)
VW-Confident-HL − VW-HL_{CM}	0.48%* (0.086)	0.374*** (0.004)	0.67%*** (0.003)	0.433*** (0.001)	0.41% (0.128)	0.238*** (0.008)

Panel C : OOS Performance Differences of Precision-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − LPW-HL	0.15%** (0.031)	0.307*** (0)	0.18%*** (0.007)	0.38*** (0)	0.38% (0.136)	0.258*** (0.002)
LPW-HL_{M} − LPW-HL	0.15%** (0.031)	0.307*** (0)	0.18%*** (0.007)	0.38*** (0)	0.37% (0.173)	0.198*** (0.008)
PW-HL − EW-HL	0.36%*** (0)	0.535*** (0)	0.37%*** (0)	0.658*** (0)	0.66%*** (0.007)	0.3*** (0.001)
PW-HL − LPW-HL	0.51%*** (0.001)	0.842*** (0)	0.55%*** (0)	1.038*** (0)	1.04%*** (0.009)	0.558*** (0)
PW-HL − PW-HL_{M}	0.23%** (0.014)	0.541*** (0)	0.23%*** (0.007)	0.617*** (0)	0.41% (0.128)	0.238*** (0.008)

Table 6

Performance of Confident and Low-Confident Long-Short Portfolios: Non-Microcap Stocks

This table reports the performance of various NN-3-based long-short portfolios over the 30-year out-of-sample (OOS) period. Every period, the sample excludes microcap stocks with market capital smaller than the 20th NYSE size percentile. See table 4 and section 5.C for a description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD) and Stambaugh-Yuan 4-factor (SY) models. The pred ret column represents the average predicted returns. The avg ret column shows the average realized returns. The α columns indicate abnormal returns. The t columns denote the t-stats of average returns and α. The SR and IR columns represent the annualized Sharpe and Information ratios, respectively.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL$_{\text{LCM}}$, HL$_{\text{CM}}$ and HL$_{\text{M}}$ are matching high-low portfolios.

Panel A: Equal-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
EW-HL$_{\text{LCM}}$	0.74%	1.83%	5.57	1.02	1.51%	4.76	0.87	1.37%	4.13	0.75
EW-Low-Confident-HL	0.74%	1.50%	3.98	0.73	1.10%	2.96	0.54	0.89%	2.32	0.42
EW-HL$_{\text{CM}}$	0.74%	1.83%	5.57	1.02	1.51%	4.76	0.87	1.37%	4.13	0.75
EW-Confident-HL	0.74%	2.25%	6.68	1.22	2.04%	6.03	1.10	1.93%	5.49	1.00

Panel B: Value-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
VW-HL	0.66%	1.42%	4.64	0.85	1.09%	3.58	0.65	0.98%	3.1	0.57
VW-HL$_{\text{LCM}}$	0.73%	1.58%	4.76	0.87	1.25%	3.76	0.69	1.10%	3.2	0.59
VW-Low-Confident-HL	0.74%	1.25%	3.13	0.57	0.88%	2.26	0.41	0.74%	1.83	0.33
VW-HL$_{\text{CM}}$	0.73%	1.58%	4.76	0.87	1.25%	3.76	0.69	1.10%	3.2	0.59
VW-Confident-HL	0.72%	2.07%	5.48	1.00	1.84%	4.78	0.87	1.64%	4.14	0.76

Panel C: Precision-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
LPW-HL$_M$	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
LPW-HL	0.69%	1.60%	4.99	0.91	1.26%	4.06	0.74	1.13%	3.47	0.63
PW-HL$_{\text{M}}$	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
PW-HL	0.69%	1.80%	5.93	1.08	1.52%	5.17	0.94	1.41%	4.57	0.83

Table 7

Statistical Comparison of Long-Short Portfolios: Non-Microcap Stocks

This table conducts pairwise statistical comparisons of the OOS performance of various NN-3-based long-short portfolios. Every period, the sample excludes microcap stocks with market capital smaller than the 20th NYSE size percentile. The tests are based on the moving block bootstrap procedure developed in section 4, with a block-length of 24. The Investment Strategy column shows the comparing pair of portfolios. The avg ret column presents the average return differences between the pair of investment strategies. The α column shows the average abnormal return differences. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios. The numbers in parenthesis are p-values. *, ** and *** denote significance at the 1%, 5% and 10% levels, respectively. See table 4 and section 5.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL_{LCM}, HL_{CM} and HL_{M} are matching high-low portfolios.

Panel A : Performance Differences of Equal-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-Low-Confident-HL	0.16% (0.393)	0.454*** (0.000)	0.25% (0.183)	0.469 (0.000)	0.35%* (0.064)	0.427*** (0.000)
EW-HL_{LCM} − EW-Low-Confident-HL	0.33%* (0.076)	0.505*** (0.000)	0.41%** (0.023)	0.535*** (0.000)	0.48%*** (0.008)	0.471** (0.000)
EW-Confident-HL − EW-HL	0.59%*** (0.000)	0.505*** (0.000)	0.69%*** (0.000)	0.588*** (0.000)	0.69%*** (0.000)	0.572*** (0.000)
EW-Confident-HL − EW-Low-Confident-HL	0.75%** (0.016)	0.959*** (0.000)	0.94%*** (0.002)	1.058*** (0.001)	1.03%*** (0.001)	0.999*** (0.000)
EW-Confident-HL − EW-HL_{CM}	0.42%** (0.015)	0.454*** (0.000)	0.53%*** (0.001)	0.523*** (0.000)	0.56%*** (0.001)	0.528*** (0.000)

Panel B : Performance Differences of Value-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
VW-HL − VW-Low-Confident-HL	0.17% (0.509)	0.391*** (0.000)	0.20% (0.438)	0.296*** (0.000)	0.24% (0.341)	0.253*** (0.001)
VW-HL_{LCM} − VW-Low-Confident-HL	0.33% (0.214)	0.428*** (0.000)	0.37%** (0.166)	0.348*** (0.000)	0.36%* (0.168)	0.280** (0.001)
VW-Confident-HL − VW-HL	0.65%*** (0.005)	0.285*** (0.000)	0.75%*** (0.001)	0.382*** (0.000)	0.66%*** (0.005)	0.304*** (0.000)
VW-Confident-HL − VW-Low-Confident-HL	0.82%** (0.029)	0.676*** (0.000)	0.95%*** (0.009)	0.679*** (0.000)	0.90%** (0.012)	0.557* (0.000)
VW-Confident-HL − VW-HL_{CM}	0.48%** (0.041)	0.248*** (0.001)	0.59%** (0.011)	0.331*** (0.000)	0.54%** (0.024)	0.277*** (0.000)

Panel C : Performance Differences of Precision-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − LPW-HL	0.06% (0.348)	0.152*** (0.000)	0.09% (0.146)	0.172*** (0.000)	0.11%* (0.082)	0.157*** (0.000)
LPW-HL_{M} − LPW-HL	0.06% (0.348)	0.152*** (0.000)	0.09% (0.146)	0.172*** (0.000)	0.11%* (0.082)	0.157*** (0.000)
PW-HL − EW-HL	0.14%** (0.014)	0.192*** (0.000)	0.17%*** (0.002)	0.222*** (0.000)	0.17%*** (0.001)	0.198*** (0.000)
PW-HL − LPW-HL	0.20%* (0.088)	0.343*** (0.000)	0.27%** (0.015)	0.394*** (0.000)	0.28%** (0.011)	0.355*** (0.000)
PW-HL − PW-HL_{M}	0.14%** (0.014)	0.192*** (0.000)	0.17%*** (0.002)	0.222*** (0.000)	0.17%*** (0.001)	0.198*** (0.000)

Table 8

Transaction Costs and Higher-Moment Adjusted Performance of Confident-HL Portfolios

This table reports the transaction costs and higher-moment-risk-adjusted performance of various NN-3-based long-short portfolios over the 30-year out-of-sample period. The Turnover column presents a portfolio's average monthly percentage change in holdings (i.e., turnover). A deduction of (0.005×Turnover) from a portfolio's realized return roughly approximates its transaction-cost-adjusted returns. The Omega, Sortino and Upside columns respectively represent the Omega, Sortino and Upside potential ratios. These ratios measure the higher-moment-risk-adjusted performance of portfolios, explicitly penalizing losses more than realizing gains. See table 4 and section 5.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL$_{\text{LCM}}$, HL$_{\text{CM}}$ and HL$_{\text{M}}$ are matching high-low portfolios.

Equal-Weighted Portfolios: Higher-Moment Adjusted Performance

	All Stocks				Non-Microcaps			
Investment Strategy	Turnover	Omega	Sortino	Upside	Turnover	Omega	Sortino	Upside
EW-HL	1.27	4.22	0.98	1.28	1.12	2.46	0.51	0.86
EW-HL$_{\text{LCM}}$	1.37	4.18	0.96	1.27	1.23	2.49	0.54	0.89
EW-Low-Confident-HL	1.88	2.83	0.71	1.10	1.89	1.89	0.37	0.80
EW-HL$_{\text{CM}}$	1.53	4.44	1.05	1.36	1.45	2.49	0.54	0.89
EW-Confident-HL	1.85	4.70	1.28	1.62	1.84	2.84	0.66	1.01

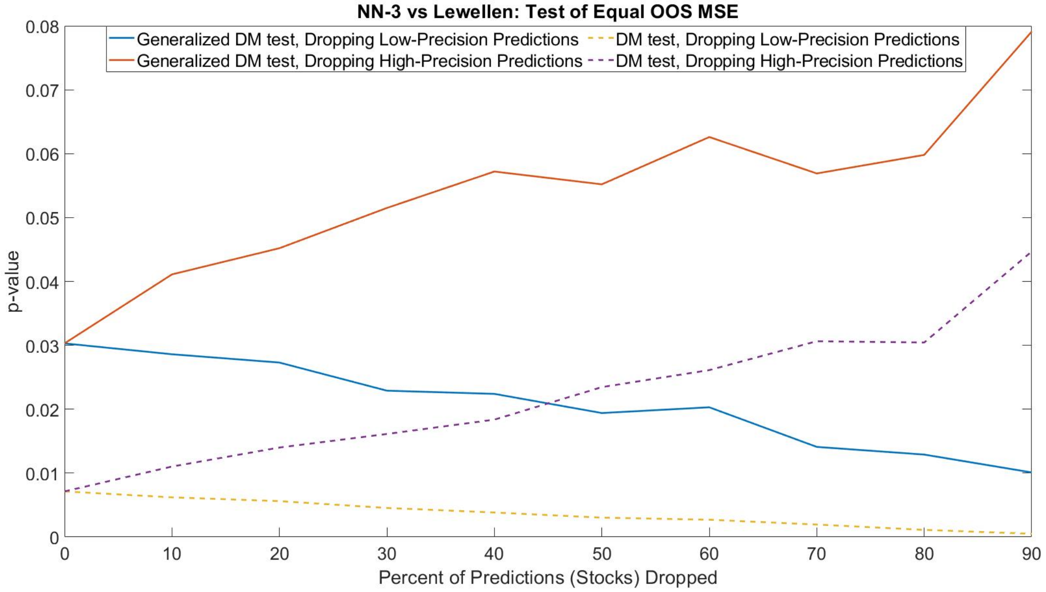
Value-Weighted Portfolios: Higher-Moment Adjusted Performance

	All Stocks				Non-Microcaps			
Investment Strategy	Turnover	Omega	Sortino	Upside	Turnover	Omega	Sortino	Upside
VW-HL	1.37	2.24	0.53	0.96	1.2	2.12	0.43	0.82
VW-HL$_{LCM}$	1.51	2.12	0.49	0.93	1.37	2.14	0.46	0.86
VW-Low-Confident-HL	1.90	1.58	0.26	0.71	1.86	1.59	0.26	0.71
VW-HL$_{\text{CM}}$	1.62	2.23	0.54	0.98	1.5	2.14	0.46	0.86
VW-Confident-HL	1.89	2.43	0.63	1.07	1.88	2.43	0.56	0.96

Precision-Weighted Portfolios: Higher-Moment Adjusted Performance

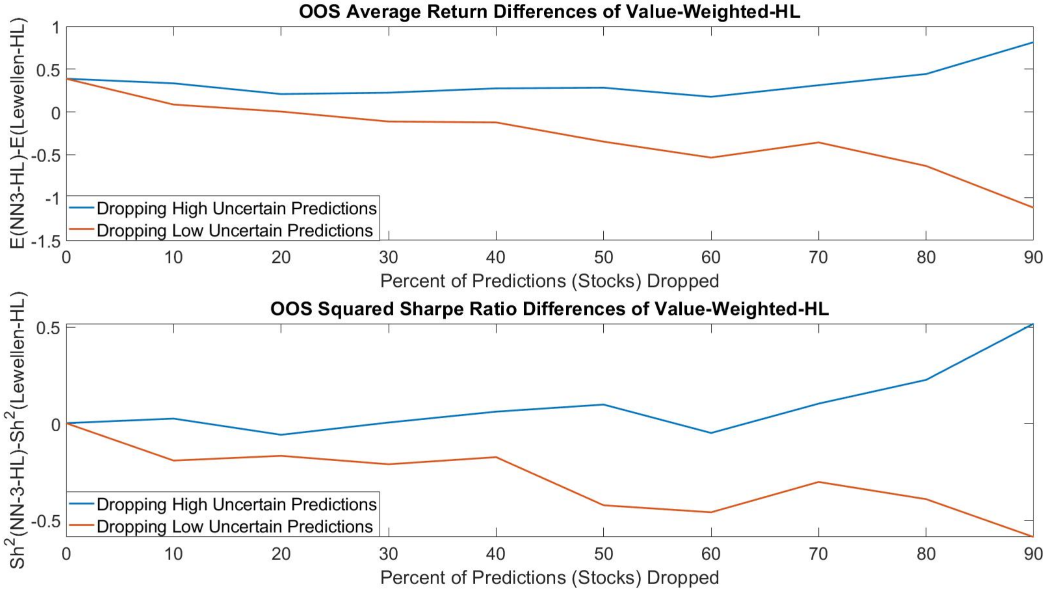
	All Stocks				Non-Microcaps			
Investment Strategy	Turnover	Omega	Sortino	Upside	Turnover	Omega	Sortino	Upside
PW-HL	1.27	4.22	0.98	1.28	1.12	2.46	0.51	0.86
PW-HL$_{\text{M}}$	1.27	4.22	0.98	1.28	1.12	2.46	0.51	0.86
PW-Low-Confident-HL	1.54	3.74	0.91	1.24	1.43	2.26	0.47	0.85
PW-HL$_{\text{M}}$	1.37	4.18	0.96	1.12	1.38	2.46	0.51	0.86
PW-Confident-HL	1.51	4.80	1.13	1.42	1.43	2.66	0.56	0.90

Figure 9. Comparing predictive performance of NN-3 with the benchmark Lewellen (2015) model



Note: This figure presents the p-values under the null hypothesis that the mean squared error of the NN-3 and Lewellen models are equal on various subsamples over the 30-year out-of-sample period. Every month, stocks are sorted into deciles according to their NN-3-based risk premium predictions' ex-ante confidence, NN-3-EC. The blue line (yellow dotted-line) displays the bootstrap (DM) p-values on the subsamples that dropout 10%, 20%, . . . and 90% of the stocks with the lowest NN-3-EC, respectively. Thus, these subsamples contain the forecasts that NN-3 confidently predicts. In contrast, the red line (purple dotted-line) represents the p-values on the subsamples comprising the forecasts that NN-3 imprecisely predicts, excluding the 10%, 20%, . . . and 90% stocks with the highest NN-3-EC, respectively.

Figure 10. Comparing predictive performance of NN-3 with the benchmark Lewellen (2015) model



Note: This figure presents the out-of-sample average return and squared-Sharpe-ratio differences between the value-weighted high-low (HL) portfolios formed using the NN-3 and Lewellen models on various subsamples. Every month, stocks are sorted into deciles according to their NN-3-based risk premium predictions' ex-ante confidence, NN-3-EC. The blue line in the top (bottom) of the figure displays the HL portfolios' average return (squared-Sharpe-ratio) differences on the subsamples that dropout 10%, 20%, . . . , and 90% of the stocks with the lowest NN-3-EC, respectively. Thus, these subsamples contain the forecasts that NN-3 confidently predicts. In contrast, the red line at the top (bottom) of the figure corresponds to the subsamples comprising the forecasts that NN-3 imprecisely predicts, excluding the 10%, 20%, . . . and 90% highest NN-3-EC stocks, respectively.

Table 9

Statistical Comparison of Long-Short Portfolios: NN-3 versus Lewellen (2015)

This table conducts pairwise statistical comparisons of the OOS performance of various long-short portfolios based on the NN-3 and Lewellen models. The tests are based on the moving block bootstrap procedure developed in section 4, with a block-length of 24. The Investment Strategy column shows the comparing pair of portfolios. The avg ret column presents the average return differences between the pair of investment strategies, the α column shows the average abnormal return differences. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios. The "HL" and "HL$_{\text{L}}$" portfolios are based on the NN-3 and Lewellen models, respectively. The numbers in parenthesis are p-values. *, **, and *** denote significance at the 1%, 5% and 10% levels, respectively. See table 4 and section 5.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low portfolio based on NN-3; HL$_{\text{L}}$=high-low portfolio based on Lewellen

Panel A : Performance Differences of Equal-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-HL$_{\text{L}}$	0.72%** (0.016)	0.247** (0.036)	0.66%** (0.036)	0.255** (0.025)	0.70%** (0.033)	0.446*** (0.002)
EW-Low-Confident-HL − EW-HL$_{\text{L}}$	0.55%** (0.089)	−0.611*** (0.002)	0.44% (0.23)	−0.753 (0)	0.49% (0.21)	−0.495*** (0)
EW-Confident-HL − EW-HL$_{\text{L}}$	1.82%*** (0)	1.055*** (0)	1.75%*** (0)	3.071*** (0)	1.80%*** (0)	1.366*** (0)
EW-Low-Confident-HL − EW-Low-Confident-HL$_{\text{L}}$	1.94%*** (0)	1.33*** (0)	1.64%*** (0)	1.225*** (0)	1.61%*** (0)	1.08*** (0)
EW-Confident-HL − EW-Confident-HL$_{\text{L}}$	0.99%* (0.059)	1.034*** (0.001)	1.25%** (0.02)	2.511*** (0)	1.42%*** (0.009)	1.253*** (0)

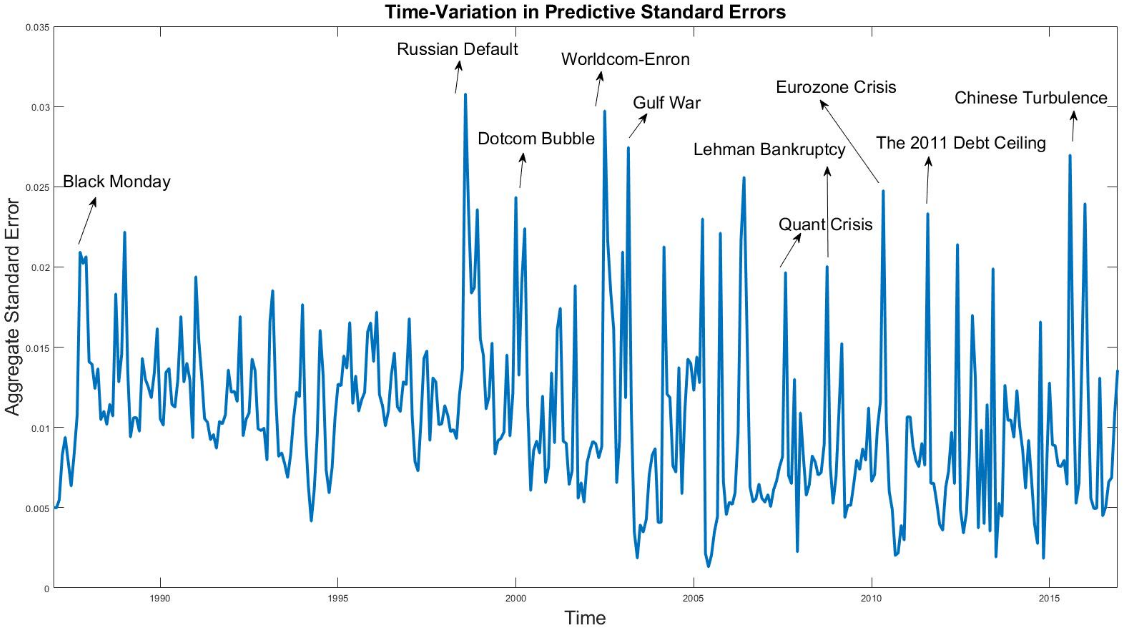
Panel B : Performance Differences of Value-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
VW-HL − VW-HL$_{\text{L}}$	0.39% (0.249)	0.002 (0.925)	0.33% (0.256)	−0.004 (0.869)	0.27% (0.36)	0.036 (0.423)
VW-Low-Confident-HL − VW-HL$_{\text{L}}$	0.22% (0.659)	−0.509*** (0.004)	−0.09% (0.847)	−0.36*** (0.005)	−0.12% (0.793)	−0.221** (0.023)
VW-Confident-HL − VW-HL$_{\text{L}}$	1.12%*** (0.003)	0.366** (0.013)	1.22%*** (0.001)	0.873*** (0)	0.93%** (0.015)	0.337*** (0.004)
VW-Low-Confident-HL − VW-Low-Confident-HL$_{\text{L}}$	0.98% (0.109)	0.281** (0.036)	0.20% (0.715)	0.051 (0.293)	0.12% (0.818)	0.013 (0.672)
VW-Confident-HL − VW-Confident-HL$_{\text{L}}$	0.44% (0.344)	0.377** (0.014)	0.86%** (0.03)	0.855*** (0)	0.81%* (0.072)	0.419*** (0.001)

Panel C : Performance Differences of Precision-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-HL$_{\text{L}}$	0.72%** (0.016)	0.247** (0.04)	0.66%** (0.033)	0.255** (0.023)	0.70%** (0.035)	0.446*** (0.002)
LPW-HL − EW-HL$_{\text{L}}$	0.57%** (0.049)	−0.06 (0.319)	0.49% (0.127)	−0.125 (0.11)	0.52% (0.127)	0.076 (0.225)
PW-HL − EW-HL$_{\text{L}}$	1.08%*** (0.002)	0.782*** (0.002)	1.03%*** (0.002)	2.796*** (0)	1.07%*** (0.002)	1.071*** (0)
LPW-HL − LPW-HL$_{\text{L}}$	1.06%*** (0.001)	0.798*** (0.001)	1.05%*** (0.001)	1.787*** (0)	1.05%*** (0.003)	0.978*** (0)
PW-HL − PW-HL$_{\text{L}}$	0.60%* (0.099)	0.273** (0.046)	0.82%*** (0.03)	1.977*** (0)	0.90%*** (0.023)	0.529*** (0.002)

Figure 11. Time-Series Variation in Standard Errors of NN-based Risk Premia



Note: This figure plots the time-series of aggregate standard errors, which are the cross-sectional averages of NN-3-based risk premium predictions' ex-ante standard errors . The labels, such as "Black Monday", "Russian Default", represent periods of major shocks.

Table 10
Aggregate Standard Errors of NN-3-based Risk Premia
This table reports time-series averages of aggregate standard errors over different periods. The aggregate standard errors equal the cross-sectional averages of NN-based risk premium predictions' standard errors.

Panel A: Overall Period

Event	Standard Error	Time Period
Overall Data	1.06%	Jan 1987 to Dec 2016

Panel B: Periods of major Shocks

Event	Standard Error	Time Period
Black Monday	2.05%	Oct 1987 to Nov 1987
Russian LTCM Defualt	3.08%	Sep 1998 to Sep 1998
Dotcom Bubble	2.24%	Apr 2000 to Apr 2000
Worldcom and Enron	2.33%	Jul 2002 to Sep 2002
Gulf War	2.75%	Mar 2003 to Mar 2003
Quant Crisis	1.97%	Aug 2007 to Aug 2007
Lehman Bankruptcy	2.00%	Oct 2008 to Oct 2008
The 2011 Debt-Ceiling	2.32%	Aug 2011 to Aug 2011
Crisis Period Average	2.31%	
Non-Crisis Period Average	1.02%	

Table 11
Cross-sectional Characteristics of Confidence-sorted Deciles
This table reports average characteristics of various confidence-sorted deciles. Every month, stocks are sorted into deciles according to their ex-ante confidence of NN-3-based risk premium predictions. Each row under All Stocks Columns represents the equal-weighted average of various characteristics across all stocks in the corresponding precision-sorted decile. The table also presents the characteristics of confidence-sorted portfolios from the long and short legs, separately. Every period stocks are first sorted into deciles according to their NN-based risk premia, with H and L representing the deciles containing the highest and lowest predicted returns. Both H and L are further partitioned into deciles according to their ex-ante confidence. The Long-Leg columns represent the average characteristics of confidence-sorted deciles of H, whereas Short-Leg columns show those of L.

Ex-ante Precision	All Stocks			Long-Leg			Short-Leg		
Decile	Size	BM	mom12m	Size	BM	mom12m	Size	BM	mom12m
1	1811	1.62	0.01	816	3.45	0.23	1939	0.76	-0.11
2	1836	1.76	0.05	810	3.37	0.23	2003	0.88	-0.08
3	1838	1.97	0.07	793	3.33	0.24	2084	0.92	-0.06
4	1788	2.12	0.08	877	3.20	0.25	2043	0.99	-0.06
5	1750	2.29	0.10	846	3.58	0.26	2102	1.04	-0.06
6	1627	2.39	0.11	805	3.58	0.26	2049	1.03	-0.05
7	1521	2.54	0.12	829	3.50	0.29	2188	0.97	-0.05
8	1394	2.62	0.13	798	3.56	0.31	2206	0.99	-0.05
9	1233	2.72	0.16	706	3.74	0.34	2283	0.89	-0.05
10	988	3.16	0.22	628	4.53	0.42	2347	1.02	-0.07

Table 12
Characteristics Distributions of Stocks in the Decile Containing the Most Confident Risk Premium Predictions

This table reports various characteristic distributions of stocks in the top decile with the most confident risk premium predictions. Every month, stocks are sorted into deciles according to their ex-ante confidence. The first row of the Size column presents the proportion of stocks in the top-most confident decile that have market capital lower than the 10^{th} percentile of sizes across all stocks. Similarly, the second (third, . . . , tenth) row of the Size column shows the proportion of stocks in the top-most confident decile that have market capital between the 10^{th} and 20^{th} (20^{th} and 30^{th}, . . . , 90^{th} and 100^{th}) percentile of sizes across all stocks. The BM, mom12m, and illiq columns represent equivalent proportions for book-to-market, 1-year momentum and illiquidity characteristics.

Decile	Size	BM	mom12m	illiq
1 (Low-Characteristic)	18.50%	10.02%	9.58%	7.23%
2	15.05%	8.21%	8.33%	6.94%
3	12.61%	8.34%	7.98%	7.03%
4	10.38%	11.39%	8.25%	7.53%
5	8.96%	14.09%	7.89%	8.14%
6	7.92%	11.61%	7.96%	9.21%
7	7.17%	7.64%	9.47%	10.61%
8	6.62%	10.55%	10.88%	12.36%
9	6.56%	13.43%	13.07%	14.54%
10 (High-Characteristic)	6.51%	15.10%	17.04%	16.50%

C. Internet Appendix

C1. Internet Appendix: Simulation Results and Robustness Checks

Table A
Performance of High-Low and Confident High-Low Portfolios: Simulation Evidence
This table compares the performance of the confident high-low portfolios with the conventional high-low portfolios on simulated data. The data contains 200 stock-level simulated true risk premia, NN-3-based estimated risk premia and their standard errors over 60 out-of-sample periods. Every period, the "True High-Low" portfolios take long (short) positions on the stocks with the simulated true risk premia greater (lower) than the $x\%$ $(100 - x\%)$ percentile of the true risk premia across 200 stocks. x equals 80, 70 and 90 under rule 1, 2 and 3, respectively. The "High-Low" portfolios take long (short) positions on the stocks with NN-3-based risk premium estimates greater (lower) than the $x\%$ $(100 - x\%)$ percentile of the predicted risk premia in the cross-section. Extreme predicted-return deciles are further partitioned into quantiles according to their precision measures. Panel A (Panel B) presents the results using the absolute t-ratios (inverse standard errors) as proxies for the precision. The "Confident High-Low" portfolios take long-short positions on the top $y\%$ subset of stocks in the extreme predicted return deciles that have the highest precision. y equals 80, 80 and 50 under rule 1, 2 and 3, respectively. The "Matching High-Low" portfolios take (short) positions on the stocks with NN-3-based risk premium predictions greater (lower) than the $z\%$ $(100 - z\%)$ percentile of the predicted risk premia in the cross-section. See section (C.C2) and equation (75) for a detailed description of the simulated data.

Panel A: Confident-HL Portfolios Constructed Using Absolute t-ratios

	Rule 1		Rule 2		Rule 3	
Portfolio	pred ret	avg ret	pred ret	avg ret	pred ret	avg ret
True High-Low	2.45%	2.45%	2.16%	2.16%	2.74%	2.74%
High-Low	3.04%	1.69%	2.60%	1.45%	3.57%	1.88%
Matching High-Low	3.64%	1.90%	3.45%	1.84%	3.72%	1.92%
Confident High-Low	3.65%	2.31%	3.47%	2.23%	3.74%	2.23%

Panel B: Confident-HL Portfolios Constructed Using Standard Errors

	Rule 1		Rule 2		Rule 3	
Portfolio	pred ret	avg ret	pred ret	avg ret	pred ret	avg ret
True High-Low	2.45%	2.45%	2.16%	2.16%	2.74%	2.74%
High-Low	3.04%	1.69%	2.60%	1.45%	3.57%	1.88%
Confident High-Low	2.72%	2.18%	2.34%	1.99%	3.41%	2.18%

Table B

Performance of Various Long-Short Portfolios: Inverse Standard Errors as Precision Measures

This table reports the performance of various NN-3-based long-short portfolios over the 30-year out-of-sample (OOS) period. This table uses inverse standard errors (rather than the absolute t-ratios) of risk premium predictions as proxies for ex-ante precision (i.e., ex-ante confidence). See table 4 and section 5.C for a description of the portfolios. The pred ret column represents the average predicted returns. The avg ret column shows the average realized returns. The t, SR and SR^2 columns denote the t-stats of the average returns, annualized Sharpe ratios and squared Sharpe ratios, respectively. Notes: EW = equal-weighted; VW = value-weighted

All Stocks: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
EW-HL	1.69%	2.52%	8.21	1.50	2.25
EW-Low-Confident-HL	1.92%	3.02%	7.62	1.39	1.93
EW-Confident-HL	1.69%	3.07%	8.46	1.54	2.39
EW-Confident-HL − EW-HL		0.55%** (0.013)			0.14*** (0.046)
EW-Confident-HL − EW-Low-Confident-HL		0.05% (0.916)			0.45*** (0.001)

All Stocks: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
VW-HL	1.62%	1.48%	4.95	0.90	0.82
VW-Low-Confident-HL	1.88%	1.13%	2.47	0.45	0.20
VW-Confident-HL	1.64%	1.83%	5.68	1.04	1.08
VW-Confident-HL − VW-HL		0.35%* (0.067)			0.26*** (0.022)
VW-Confident-HL − VW-Low-Confident-HL		0.70%* (0.071)			0.87*** (0.000)

Non-Microcaps: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
EW-HL	0.68%	1.66%	5.43	0.99	0.980
EW-Low-Confident-HL	0.72%	1.30%	3.53	0.64	0.35
EW-Confident-HL	0.66%	1.87%	5.95	1.08	1.17
EW-Confident-HL − EW-HL		0.23%** (0.041)			0.19** (0.02)
EW-Confident-HL − EW-Low-Confident-HL		0.57%*** (0.000)			0.82*** (0.000)

Non-Microcaps: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
VW-HL	0.66%	1.42%	4.64	0.85	0.72
VW-Low-Confident-HL	0.71%	1.25%	2.90	0.53	0.27
VW-Confident-HL	0.65%	1.91%	5.68	1.04	1.08
VW-Confident-HL − VW-HL		0.49%** (0.041)			0.36** (0.001)
VW-Confident-HL − VW-Low-Confident-HL		0.66%* (0.0723)			0.81*** (0.000)

Table C
Comparing Confident-HL Portfolios with Double-sorted HL Portfolios

This table compares the out-of-sample performance of the Confident-HL portfolios with the HL portfolios that are double sorted on predicted-returns. EW(VW)-Confident-HL represents the equal(value)-weighted Confident long-short portfolio that only include stocks with the most confident risk premium predictions. See section 5.C for a detailed description of the portfolios. Each period, stocks are sorted into quantiles according to their NN-based risk premia. EW-double-sorted-HL and VW-double-sorted-HL denote the HL portfolios that take equal-weighted and value-weighted long (short) positions on stocks that have greater (lower) predicted-returns than the predicted-return of the 99^{th} (1^{st}) quantile, respectively. The avg ret column presents the average return differences between the pair of investment strategies. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios. The numbers in parenthesis are p-values. *, ** and *** denote significance at the 1%, 5% and 10% levels, respectively.

All Stocks: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
EW-Confident-HL	1.97%	3.61%	9.58	1.75	3.06	3.12	2.99
EW-double-sorted-HL	2.54%	3.99%	8.58	1.57	2.46	2.49	1.87
Difference		-0.37% (0.168)			0.60^{***} (0.000)	0.96^{***} (0.000)	1.12^{**} (0.000)

All Stocks: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
VW-Confident-HL	1.90%	2.21%	5.95	1.09	1.18	0.87	0.59
VW-double-sorted-HL	2.51%	2.39%	5.28	0.96	0.93	0.5	0.42
Difference		-0.18% (0.61)			0.25^{**} (0.02)	0.37^{**} (0.016)	0.17^{**} (0.03)

Non-Microcaps: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
EW-Confident-HL	0.66%	2.25%	6.68	1.22	1.49	1.39	1.22
EW-double-sorted-HL	1.02%	2.39%	5.56	1.01	1.02	0.87	0.66
Difference		-0.13% (0.62)			0.47^{**} (0.000)	0.52^{**} (0.000)	0.56^{**} (0.000)

Non-Microcaps: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
VW-Confident-HL	0.72%	2.07%	5.48	1.00	1.00	0.97	0.69
VW-double-sorted-HL	1.01%	2.20%	4.71	0.86	0.74	0.69	0.44
Difference		-0.13% (0.73)			0.26^{**} (0.000)	0.28^{**} (0.000)	0.25^{**} (0.000)

C2. Internet Appendix: Simulation Details

To assess the finite sample performance of this paper's standard errors and Confident-HL portfolios, I replicate the simulation exercise of GKX.[26] I simulate a 3-factor model for excess returns, for $t = 1, 2, \ldots, T$:

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1},\ \ e_{i,t+1} = \beta_{i,t} v_{t+1} + \epsilon_{i,t+1},\ \ z_{i,t} = (1, x_t)' \otimes c_{i,t},\ \ \beta_{i,t} = (c_{i1,t}, c_{i2,t}, c_{i3,t}), \quad (70)$$

where c_t is a 200×180 matrix of characteristics, v_{t+1} is a 3×1 vector of factors, x_t is a univariate time series, and ϵ_{t+1} is a 200×1 vector of idiosyncratic errors. I choose $v_{t+1} = 0,\ \forall t$ under models 1 and 3 and $v_{t+1} \sim \mathcal{N}(0, 0.05^2 \times I)$ under models 2 and 4, respectively. I specify $\epsilon_{i,t+1} \sim \epsilon_{i,t+1} \sim \mathcal{N}(0, 0.05^2)$. These parameters are calibrated so that the average time series R^2 is 50% (40%) and annualized volatility is 24% (30%) under models 1 and 3 (2 and 4). The OOS-R^2 of NN-3-based risk premium predictions on the simulated data is 3.8% (3.2%) under models 1 and 3 (2 and 4).

I simulate the panel of characteristics by

$$c_{ij,t} = \frac{2}{N+1} CSrank(\bar{c}_{ij,t}) - 1,\ \ \bar{c}_{ij,t} = \rho_j \bar{c}_{ij,t-1} + \epsilon_{ij,t},\ \text{ for } 1 \leq i \leq 200,\ \ 1 \leq j \leq 180, \quad (71)$$

where $CSrank$ denotes the cross-sectional rank.

And the time-series x_t is given by

$$x_t = \rho x_{t-1} + u_t, \quad (72)$$

where $u_t \sim \mathcal{N}(0, 1 - \rho^2)$, and $\rho = 0.95$ so that x_t is highly persistent.

Under models 1 and 2, the parametric form of $g(.)$ is linear and given by

$$g(z_{i,t}) = (c_{i1,t}, c_{i2,t}, c_{i3,t})\theta_0,\ \text{ where } \theta_0 = (0.02, 0.02, 0.02)'. \quad (73)$$

[26] I thank GKX for making their code publicly available.

In contrast, under models 3 and 4, $g(.)$ takes the following non-linear functional form

$$g(z_{i,t}) = (c_{i1,t}^2, c_{i1,t} \times c_{i2,t}, sgn(c_{i3,t} \times x_t))\theta_0, \text{ where } \theta_0 = (0.04, 0.03, 0.012)^{'}. \tag{74}$$

To summarize, the simulated true risk premia are linear in characteristics under models 1 and 2, whereas they are non-linear under models 3 and 4. Models 1 and 3 do not entertain cross-sectional temporal residual correlations, whereas models 2 and 4 do.

Lastly, I divide the whole time-series into three consecutive subsamples of equal length (60) for training, validation, and testing, respectively. Although this paper's standard errors are derived under the assumption that the residual errors are uncorrelated in the time-series and cross-section, table (1) of the main section indicates that the standard errors are well-calibrated even under models 2 and 4.

Simulations for table (A) of the Internet Appendix use the non-linear specification of model 3, given by

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1}, \; e_{i,t+1} = \epsilon_{i,t+1}, \; z_{i,t} = (1, x_t)^{'} \otimes c_{i,t}, \tag{75}$$

where $\epsilon_{i,t+1} \sim \epsilon_{i,t+1} \sim \mathcal{N}(0, 0.05^2)$, $g(z_{i,t})$ is given by (74) and $c_{i,t}$ is given by (71).

C3. Why Confidence-levels are Better Measures of Precision Relative to Inverse Standard Errors

In this section, I present a simple example showing why the absolute t-stat is a better measure relative to the inverse standard error for constructing Confident-HL portfolios. Consider regressing a given cross-section of excess stock returns on one of stock characteristics (e.g., betas)

$$r_i = \lambda\beta_i + \epsilon_i,\ \epsilon_i \sim MVN(0, \sigma^2 I),\ i = 1, 2, \ldots N \tag{76}$$

where r_i, β_i are assumed to be given. λ, which can be interpreted as the market premium, is an unknown parameter. Assume $\lambda > 0$ without loss of generality. Let $\hat{\lambda}$ be the OLS estimate of λ obtained from the cross-sectional regression in (76).

Now, consider four stocks in the out-of-sample that have betas β_1^*, β_2^*, β_3^* and β_4^*, respectively. Let $0 < \beta_1^* < \beta_2^* < \beta_3^* < \beta_4^*$. Their predicted excess returns are then given by $\beta_1^*\hat{\lambda}$, $\beta_2^*\hat{\lambda}$, $\beta_3^*\hat{\lambda}$ and $\beta_4^*\hat{\lambda}$, respectively. Straightforward algebra implies that these predictions' standard errors equal $\frac{\beta_1^*\hat{\sigma}}{\sum\beta_i^2}$, $\frac{\beta_2^*\hat{\sigma}}{\sum\beta_i^2}$, $\frac{\beta_3^*\hat{\sigma}}{\sum\beta_i^2}$ and $\frac{\beta_4^*\hat{\sigma}}{\sum\beta_i^2}$, respectively. $\hat{\sigma}$ is the OLS estimate of σ in (76).

Thus, the standard errors are proportional to the stock betas. In contrast, the absolute t-ratios are invariant across stocks. In other words, the "confidence-level" of predicting returns is the same across all stocks. In the following paragraph, I show that Confident-HL-se portfolios formed using the standard errors yield sub-optimal returns relative to the traditional HL portfolios. In contrast, Confident-HL-t portfolios formed using the absolute "t-ratios" do not.

Consider the following trading strategies using these four stocks' predicted returns and their precision measures.

1. Conventional-HL: Takes equal-weighted long (short) positions on the top (bottom) stocks with the highest (lowest) predicted returns.

2. Confident-HL-t: Sort stocks into two quantiles based on their predicted returns. Take the long (short) position on the stock in the top (bottom) quantile that has the highest absolute t-ratio. If two stocks have the same absolute t-ratios, take the equal-weighted average.

3. Confident-HL-se: Sort stocks into two quantiles based on their predicted returns. Take the long (short) position on the stock in the top (bottom) quantile that has the lowest standard error. If two stocks have the same absolute standard errors, take the equal-weighted average.

Then the expected return of these three strategies are given by

$$E(\text{Conventional-HL}) = E(\text{Conventional-HL-t}) = \left[\frac{(\beta_3^* + \beta_4^*)}{2} - \frac{(\beta_1^* + \beta_2^*)}{2}\right]\Big(P(\hat{\lambda} > 0) - P(\hat{\lambda} < 0)\Big) \tag{77}$$

$$E(\text{Confident-HL-se}) = (\beta_3^* - \beta_1^*)\Big(P(\hat{\lambda} > 0) - P(\hat{\lambda} < 0)\Big) \tag{78}$$

For sufficiently large β_4^*, $E(\text{Conventional-HL}) > E(\text{Confident-HL-se})$. Thus, standard errors must always be evaluated relative to the "level" of predictions to obtain better measures of precision.

Comparing Asset Pricing Models with Non-Traded Factors and Principal Components

Rohit Allena [*]
Goizueta Business School
Emory University

March 31, 2021

Abstract

This paper develops a Bayesian methodology to compare asset pricing models containing non-traded factors and principal components. Existing comparison procedures are inadequate when models include such factors due to estimation uncertainties in mimicking portfolios and return covariances. Furthermore, regressions of test assets on such factors are interdependent, rendering comparisons with recently proposed priors sensitive to subsets of the test assets. Thus, I derive novel, non-informative priors that deliver invariant inferences. Simulations suggest that my methodology substantially outperforms existing methods in identifying true non-traded models. I find that macroeconomic models dominate several recent benchmark models with traded factors and principal components.

Keywords: Macroeconomic Factors, Principal Components, Novel Non-Informative Priors, Conditional Test Assets Irrelevance, Invariant Comparisons, Out-of-Sample Model Comparisons

[*] I am grateful to Jay Shanken for his invaluable suggestions and support. I benefited from discussions with Francisco Barillas, Tarun Chordia, Christian Julliard, Chandra Mangipudi, Bryan Kelly, Alejandro Lopez-Lira, William Mann, Jegadeesh Narasimhan, Cesare Robotti, Tim Simin, Dacheng Xiu and other participants at the annual conferences of Western Finance Association (WFA 2020), European Finance Association (2019), University of Chicago - Machine Learning and the New Empirical Asset Pricing (2018), Northern Finance Association (2018), Society for Financial Econometrics (2018), Econometric Society European Meetings (2018), and Emory University, and I'm thankful for their comments and feedback. All errors are mine.
Address: 1300 Clifton Road, Atlanta, 30329. E-mail Address: rohit.allena@emory.edu

7. Introduction

Research has examined a plethora of factor-based asset pricing models to explain the cross-section of expected returns. These factors are either tradable portfolios (e.g., market returns); non-tradable macroeconomic factors (e.g., consumption growth); or statistically estimated latent factors (e.g., principal components). Comparing models with these factors and identifying a relatively superior model is important, not only for adequately summarizing the cross-section of stock returns (Fama and French (2016)) but also for evaluating the performance of managed-fund portfolios (Fama and French (2010)). In recent studies, Barillas and Shanken (2018), Barillas and Shanken (2020) (BS henceforth), and Chib, Zeng, and Zhao (2020b) (CZZ henceforth) propose Bayesian procedures to compare models that exclusively comprise traded factors. This paper complements studies in this literature by developing a Bayesian methodology that permits simultaneous comparison of models containing traded, non-traded factors, and principal components (PCs).

Non-traded factors are popular in the literature, as they directly relate to the aggregate macroeconomic activities or business cycles and thus appear to explain the expected returns. More recently, He, Kelly, and Manela (2017) document that a single non-traded factor, the intermediary capital ratio, significantly explains cross-sectional variation in the expected returns of a wide range of assets, including equities, bonds and commodities. Koijen, Lustig, and Van Nieuwerburgh (2017) argue that two macroeconomic bond factors jointly determine the expected returns of equities and bonds, whereas Campbell, Giglio, Polk, and Turley (2018) argue that an intertemporal capital asset pricing model (ICAPM), which includes a stochastic volatility factor, explains a wide range of anomalies. In practical applications, investment managers that rely on factor investing argue that macroeconomic factors substantially explain the expected returns across various asset classes.[1]

Despite their popularity, evaluating models with non-traded (rather than traded) factors, is extremely challenging. When all factors of an asset pricing model are tradable portfolios, Jensen, Black, and Scholes (1972) have shown that the model holds iff the time series intercepts in the regression of excess returns on the factors, the alphas, equal zero. BS build on this insight and

[1] In the official BlackRock factor investing commentary, Andrew Ang and Ked Hogan argued that six macroeconomic factors explain more than 90% of the expected returns across asset classes.

evaluate asset pricing models with traded factors based on their posterior densities of alphas. Such a simple condition on alphas does not hold when the factors are non-traded. Breeden (1979) notes that the zero-alpha condition holds when the non-traded factors are substituted with their mimicking portfolios. However, these portfolios are unknown and thereby can only be estimated. Thus, zero-alpha tests with non-traded factors require an additional uncertainty adjustment. I address this challenge by using a Bayesian framework to provide an exact methodology for evaluating an individual asset pricing model, as well as for simultaneously comparing multiple models containing non-traded factors.

Another burgeoning literature in asset pricing advocates the use of latent factors, which are estimated from PCs of the covariance matrices of returns and other advanced machine-learning methods. In the absence of near-arbitrage conditions, Kozak, Nagel, and Santosh (2018) theoretically argue that a sparse model with the first few PCs characterizes the cross-section of expected stock returns. Kozak et al. (2019) further show that a stochastic discount factor (SDF) formed from a small number of leading PCs can adequately explain the cross-section of various anomalies, whereas a sparse SDF formed from the traditional factor models cannot. Both the theoretical as well as empirical arguments rely on an assumption that the covariance matrix of returns is known. Studies of Kan and Zhou (2007) and Britten-Jones (1999) document large sampling errors in the estimation of the return covariances. Thus, recognizing the significance of this estimation uncertainty, I develop a formal zero-alpha based model comparison methodology with the PCs, which accounts for the estimation error in the covariance of returns.

Simulations suggest that sampling errors of various parameters (e.g., alphas, betas and residual covariances) and weights of mimicking portfolios and PCs, significantly impact model comparison inferences. Existing procedures, including comparing models based on their out-of-sample ordinary and the generalized least squared R^2s (OOS-R^2, GLS-R^2), that rely on these estimated parameters, often fail to measure true performance of models and yield misleading inferences. By shrinking the parameters toward economically motivated priors, the Bayesian procedures of BS and CZZ perform better in comparing models that exclusively comprise traded factors. However, applying their methods to compare models containing non-traded factors and PCs would be misleading, as the sample estimates of mimicking portfolios and PCs are significantly imprecise. For instance, Chib, Huang, Zhao, and Zhou (2020a) use the CZZ method to compare models containing PCs without adjusting for their estimation uncertainty. Similarly, BS note that their method could be employed to compare models containing non-traded factors by substituting them with their mimicking portfolios. Although they indicate that additional estimation issues might arise, they do not discuss how to address such estimation issues.

This paper provides the first Bayesian methodology to compare models with traded, non-traded factors, and PCs in a unified framework. The methodology reduces the influence of sampling errors of the estimated weights of mimicking portfolios and PCs by shrinking the posterior Sharpe ratios of unknown mimicking portfolios and PCs toward economically motivated priors. The methodology is equivalent to comparing models based on their out-of-sample prediction records, after adjusting for the associated estimation uncertainties. Simulations indicate that the methodology enhances selection rates of simulated null models that comprise non-traded factors by more than 200% and 100%, compared with the existing OOS-R^2 and Bayesian procedures, respectively. Likewise, the methodology increases the average of null models' posterior probabilities by more than 40%, and decreases the standard deviation of a given null model's probabilities across simulation-iterations by an impressive 15%, compared with the existing Bayesian procedures.

I begin by showing that Bayesian zero-alpha tests with non-traded factors and PCs are fundamentally challenging because of the following interdependence related to test assets: Consider a set of 25 test assets and an asset pricing model with 5 non-traded factors. The zero-alpha tests involve regression of the test assets on the 5 mimicking portfolios, which themselves are various linear combinations of the test-asset returns. Thus, it is sufficient to regress any 20 (25-5) linearly independent test assets on the mimicking portfolios. This is because the regression parameters, which include the alphas, and the nuisance parameters (betas and residual covariances), completely determine the regression parameters of the remaining 5 assets. The traditional Jeffreys (1998) priors for the test assets' regression parameters used in BS and in CZZ would be inadequate in this framework. These priors do not impose the interdependence apriori, thereby rendering the posterior densities of alphas and Bayesian inferences sensitive to the initial choice of the 20 test assets.

I tackle this challenge by constructing novel, non-informative priors for the nuisance parameters to preserve the parameter interdependence and yield inferences that are *invariant* to the initial choice of the subset of the test assets. Interestingly, these priors also deliver inferences that are invariant to the scale of traded and non-traded factors and PCs. As researchers often use a variety of scaling procedures, such a scale-invariant measure would be particularly necessary. For example, a recent study by Kozak et al. (2019) scale various models' mean-variance efficient portfolio returns and factor returns to have the same standard deviation as that of the market returns.

Using the novel priors for the nuisance parameters and economically motivated priors for the alphas, which reflect the extent of models' mispricing, I first derive a Bayesisan statistic to test an individual asset pricing model containing non-traded factors or PCs. This statistic is analytically shown to be a modified function of the marginal likelihoods of BS and CZZ. The modification entails two fundamental adjustments. The first adjusts for the estimation uncertainty of the mimicking portfolios or the PCs, and the second adjusts for the parameter interdependence to ensure invariance to the subsets of the basis test assets.

I next develop a procedure that permits *simultaneous* comparison of various asset pricing models with traded and non-traded factors and PCs. The posterior model probabilities are readily obtained from the modified marginal likelihoods. This follows from a fundamental result showing that model comparison with non-traded factors and PCs only requires examination of each model's ability to price the mimicking portfolios and the PCs in the other models, after adjusting for their estimation uncertainties. As a consequence, I show that conditional on the posterior densities of the mimicking portfolios and PCs, test assets would be irrelevant for model comparisons. Barillas and Shanken (2017) have drawn similar conclusions for several likelihood-based model comparison procedures, which include the differences in likelihoods and Sharpe ratios. I formally prove the *conditional test assets irrelevance* in a Bayesian framework. This key result entails the novel prior specification and would not be achievable with the existing priors.

Interestingly, a straightforward decomposition of the marginal likelihoods reveals that model comparison with my methodology is equivalent to ranking models based on their out-of-sample predictive performance, after adjusting for the estimation uncertainties. Given the prevalence of recent studies that emphasize out-of-sample model comparisons to avoid models that overfit, this result is particularly relevant. Although existing studies often compare models based on their estimated out-of-sample R^2s and Sharpe ratios, it is unclear whether such measures are statistically significant. For example, Kan, Robotti, and Shanken (2013) have provided many examples of large R^2 differences among models that are statistically insignificant. Deriving asymptotic distributions of the out-of-sample R^2s that entail estimation uncertainties would be arduous. An advantage with my methodology is that it has a natural out-of-sample interpretation that reflects estimation uncertainty.

In an empirical application, I implement my methodology on monthly data and compare prominent asset pricing models with traded, non-traded factors and PCs[2]. These models include the Capital Asset Pricing Model (CAPM); the three and five factors models of Fama and French (1992, 2015) (FF-3, FF-5); the investment-based model of Hou, Xue, and Zhang (2015) (HXZ); the mispricing model of Stambaugh and Yuan (2017) (SY); the six-factor model of BS (BS-6); the ICAPMs of Petkova (2006), Campbell and Vuolteenaho (2004) (CV), Campbell, Giglio, and Polk (2013) (CGP), and Campbell et al. (2018) (CGPT); and two models with the first five and six PCs (PC1-5, PC1-6), respectively. PC1-5 was earlier examined by Kozak et al. (2018). I use two sets of test assets. The first one consists of excess returns of 52 anomalies (52-anom) that have been examined by Kozak et al. (2019). Because inferences from such characteristics sorted portfolios could be misleading (Lewellen, Nagel, and Shanken (2010)), I also consider a second set that adds excess returns of 10 industry portfolios to the 52 anomalies (52-anom+10-ind). The monthly data are from January 1974 to December 2016.

The empirical study begins by comparing traditional models with traded and non-traded factors, excluding PCs. I find that SY and BS-6 have the highest posterior probabilities, which are followed by the ICAPMs of CGPT, CGP, and CV. Importantly, these ICAPMs dominate the benchmark models of Fama and French (2015) and Hou et al. (2015). Given that the benchmark

[2] My methodology also can be applied for comparing *set of all possible models* formed from a given set of factors. Such comparisons, as Harvey, Liu, and Zhu (2016) and more recently Fama and French (2018) emphasize, create an overwhelming multiple comparisons problem that preempts statistical inference. Thus, to alleviate this data mining concern, I restrict myself to comparing prominent models that are either theoretically or empirically well-established in the literature.

models' factors are directly constructed from sorted portfolios of various anomalies, the superior performance of the economically motivated ICAPMs makes them even more impressive. Whereas Campbell et al. (2018) document that each anomaly individually attains lower alphas with respect to their ICAPM, this paper formally recognizes the dominance of the ICAPMs by jointly examining the alphas of all the anomalies, after adjusting for the associated estimation uncertainties.

The results are qualitatively similar for different prior expectations on the mispricing and across the test assets. I also report the posterior probabilities based on the Bayesian procedure of BS and CZZ, where the estimated mimicking portfolios are substituted for the non-traded factors, without adjusting for their uncertainties. I find that the biases from failing to adjust for estimation uncertainty are significant across the models. For example, ignoring estimation uncertainty suggests that HXZ and CGP are equally likely to summarize the expected returns. In contrast, my methodology suggests that CGP is 1.23 times more likely to summarize the cross-section of expected returns. Thus, the uncertainty adjustment could substantially affect model comparison inferences.[3]

The differences in results between BS and CZZ and my methodology are consistent with the simulation results and intuition that the sample estimated mimicking portfolios often underestimate the explanatory "power" of true non-traded factor models because of significant sampling errors. By shrinking the weights of mimicking portfolios toward economically motivated priors, my methodology reduces the influence of sampling errors and thus better captures the true performance of non-traded factor models.

I then move on to a more comprehensive comparison by adding the models with the PCs to the traditional models. I find that SY and BS-6 have the highest posterior model probabilities across the test assets. Interestingly, the ICAPMs of CGPT, CGP, and CV dominate the benchmark PC1-5 model, which was earlier examined by Kozak et al. (2018), across both sets of test assets.

[3] Interestingly, the difference between BS and CZZ and my methodology could be important from an investment perspective. In particular, Pástor and Stambaugh (2000) have shown that the optimal portfolio of an investor facing model uncertainty is a function of posterior probabilities of the models. Because the investor is uncertain about the true mimicking portfolios, the optimal portfolio will be a function of this paper's probabilities that account for their estimation uncertainty. Thus, the investor deems the portfolio based on BS probabilities sub-optimal, as it could lead to a significant utility loss.

The uncertainty adjustment turns out to be essential for the models with PCs as well. Ignoring estimation uncertainty suggests that odds in favor of PC1-5 are 1.18 times more than the CGPT model. In contrast, my methodology suggests that CGPT is 1.64 times more likely than PC1-5 to summarize the cross-section of expected returns. I find that PC1-6 performs almost on par with the BS-6 and CGPT models on the 52-anom test assets. However, its performance significantly deteriorates on the 52-anom+10-ind test assets.

The empirical results relate to and partially contrast with Kozak et al. (2019), who compare models based on their out-of-sample R^2s. They argue that traditional models with sparse characteristic-based factors would not summarize the cross-section of expected returns, whereas the first few principal components would. Inferences from such point estimates of OOS-R^2s could be misleading because these measures rely on various estimated parameters that have significant sampling errors. In contrast, this paper *statistically* establishes dominance or comparable performances of various sparse-factor models, including SY, BS-6, and CGPT, over models with the leading PCs.[4] Of course, the majority of these traditional factor models, including SY and BS-6, consist of factors that are empirically motivated and directly constructed from various sorted or anomaly portfolios. Thus, the influence of data mining in the construction of these empirically motivated factors still needs to be explored. I also find that theoretically motivated ICAPMs outperform the benchmark FF-5 and HXZ models. Studies in this literature have not statistically established this result before.

My results are not directly comparable with Kozak et al. (2019) for two main reasons, among others. First, Kozak et al. (2019) construct a sparse SDF involving a few PCs using an elastic-net procedure, and compare it with the SDFs implied by the other models. Mathematically, their SDF comprises a few PCs with the highest return means, which differ from the conventional leading PCs that instead explain the highest return covariances.[5] For example, with the daily anomaly return data, their SDF comprises the PCs-$\{1, 2, 3, 5, 10\}$. Because the true return means of PCs are unknown, comparing models with these sophisticated factors requires an additional uncertainty adjustment. Second, Kozak et al. (2019) use data at a daily level. My methodology

[4] BS-6 is a sparse characteristic-based factor model.

[5] Lettau and Pelger (2020) analytically establish this result.

relies on an assumption that returns are uncorrelated in the time series. Kozak et al. (2019) made the same assumption. However, daily returns are known to be non-synchronous, and thus it is unclear whether this assumption suits the dynamics of daily data (Shanken (1987)). A formal methodology that accommodates potential temporal dependence of the daily returns and adjusts for the uncertainty of PCs' return means is left for a future exercise.

The paper contributes to the literature that evaluates and compares various asset pricing models. In the classical setting, Kan et al. (2013) and Maio (2019), and more recently Barillas, Kan, Robotti, and Shanken (2019) derive asymptotic distributions of cross-sectional R^2 and Sharpe ratio differences among competing models, respectively. These studies do not accommodate models with principal components. Moreover, these methodologies compare models based on their in-sample measures, which do not reflect the models' true predictive ability. For example, Kan, Wang, and Zheng (2019) argue that in-sample Sharpe ratios rely on ex-post tangency portfolios, which are not attainable by investors. Although they derive distributions for out-of-sample Sharpe ratios that investors achieve using individual asset pricing models, they do not provide a method for comparing multiple models. My methodology addresses this challenge by simultaneously comparing models containing non-traded factors and PCs based on their out-of-sample predictions.

In a contemporaneous working paper, Bryzgalova, Huang, and Julliard (2020) conduct Bayesian comparisons for models containing traded and non-traded factors using the cross-sectional Fama-Macbeth approach. Their framework fundamentally differs from mine in several ways. First, I examine time-series (rather than cross-sectional) regressions using mimicking portfolios. This approach naturally restricts the risk premium (γ) of a traded factor to be equal to its expected excess return ($E(F)$), as advocated by theory. Bryzgalova et al. (2020) do not impose this key restriction, which can lead to unduly favoring models containing traded factors, by relying on implausible estimates of the risk premia for those traded factors (Maio (2019), Lewellen et al. (2010)).[6] Second, Bryzgalova et al. (2020) use the traditional Jeffreys priors (rather than the novel priors) for the

[6] The restriction, $\gamma = E(F)$, should hold when the zero-beta rate equals risk-free rate and models are not misspecified. In the majority of Bayesian papers, including BS, Kozak et al. (2019), and Bryzgalova et al. (2020), models are either i) assumed to be not missspecified, in which case $\gamma = E(F)$ must be imposed apriori; ii) or priors for the misspecification are specified to be centered around zero, in which case priors for γ should be centered around $E(F)$. Bryzgalova et al. (2020) impose neither of these restrictions on γ apriori.

test assets' regression parameters, and thus their method is not suitable for comparing models containing PCs.

I organize the rest of the paper as follows. Section 2 provides the premise of linear asset pricing models and challenges associated with non-traded factors. Section 3 develops a Bayesian framework for testing the validity of an individual asset pricing model with non-traded factors. Section 4 extends the framework for simultaneous model comparisons. Section 5 extends the framework for PCs. Section 6 shows the equivalence of the methodology to comparing models based on their out-of-sample prediction records. Section 7 provides a simulation study on the performance of the methodology. Section 8 describes the data and provides empirical results. Section 9 concludes. Appendix includes proofs of propositions. Internet Appendix-I includes a critical discussion of what priors to use, when, and why, to compare asset pricing models in general. Internet Appendix-II provides additional details about the simulations and lists the definitions of the non-traded factors that I examine in the empirical section.

8. Asset Pricing Models with Non-Traded Factors

This section briefly discusses challenges associated with testing asset pricing models that comprise non-traded factors and gives a heuristic explanation of how these can be addressed in a Bayesian framework. First, under the assumption that the zero-beta rate equals the risk-free rate, an asset pricing model with K factors, $\{f_{1t}, f_{2t}, \ldots, f_{Kt}\}$, summarizes the cross-section of expected stock returns if there exist risk premia, $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$, such that

$$E(R_i) = \gamma_1\beta_{i1} + \gamma_2\beta_{i2} + \cdots + \gamma_k\beta_{ik}, \forall \text{ assets } i, \tag{1}$$

where $E(R_i)$ denotes the expected excess return of the asset i and $\{\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK}\}$ are coefficients in the time-series regression of excess returns, R_{it}, on the factors.

$$R_{it} = \alpha_i + \beta_{i1}f_{1t} + \beta_{i2}f_{2t} + \cdots + \beta_{iK}f_{Kt} + \epsilon_{it} \tag{2}$$

When models exclusively comprise traded factors, Jensen (1968) and Jensen et al. (1972) have shown that $\gamma_k = E(f_{kt})$, and thus the cross-sectional condition in (1) holds iff the time-series intercepts, α_i, equal zero, for all assets, i. Thus, evaluating an asset pricing model with traded factors requires examination of only time series alphas[7]. BS used this insight and conducted model comparisons based on the posterior densities of time-series alphas.

For non-traded factors, there is no such relation between γ and $E(f)$, and thus the zero alpha condition will not necessarily hold. Thus, evaluating models containing non-traded factors is not straightforward. However, Breeden (1979) and Huberman, Kandel, and Stambaugh (1987) have shown that there exist mimicking portfolios, $\{f_1^m, f_2^m, \ldots, f_k^m\}$, such that the zero alpha restriction holds when the excess returns are instead regressed on these mimicking portfolios. Equivalently, this implies, $\alpha_i^m = 0$, in the regression below, for all assets i:

$$R_{it} = \alpha_i^m + \beta_{i1}^m f_{1t}^m + \beta_{i2}^m f_{2t}^m + \cdots + \beta_{iK}^m f_{Kt}^m + \epsilon_{it}^m, \tag{3}$$

where the mimicking portfolio, f_{kt}^m, of the non-traded factor f_{kt} is a weighted sum of excess returns across all the assets. The weights are proportional to the coefficients in the time-series regression of the non-traded factor on all the assets' excess returns. In particular, the mimicking portfolio of the non-traded factor f_{kt} is obtained by

$$f_{kt} = c_k + w_k^T R_t + \eta_{kt}, \;\; f_{kt}^m = w_k^T R_t / 1^T w_k \;\forall\; k, \tag{4}$$

where f_{kt}^m is the mimicking portfolio of the factor, f_{kt}, R_t denotes the set of excess returns of given test assets at period t, 1 is the vector of ones with the same dimension as w_k, $w_k/1^T w_k$ is the vector of unobserved true weights of the mimicking portfolio, f_{kt}^m.

Thus, an asset pricing model with non-traded factors is equivalent to implementing two time-series regressions – i) f_t on R_t (4); and ii) R_t on $w^T R_t$ (3), where $\alpha_i^m = 0$, $\forall\; i$. Although the true weights (w) of the mimicking portfolios are unobserved, both the regressions, as well as the

[7] Gibbons, Ross, and Shanken (1989) (GRS) developed a formal procedure to evaluate an individual asset pricing model with traded factors. GRS jointly tests $\alpha_i = 0$, $\forall$ assets, i.

zero-alpha restriction can be easily implemented in a Bayesian framework because of the following product identity

$$ML = P(f, R) = \int \underbrace{P(f|R, par, w)}_{\text{Implements eq (4)}} \underbrace{P(R|par, w)}_{\text{Implements eq (3)}} P(par, w)dw\ d(par). \tag{5}$$

Note that the first term on the right-hand side of (5) implements the regression of the non-traded factors on the excess returns (equation (4)) given w and other regression parameters (par). The second term implements the regression of the excess returns on the mimicking portfolios (equation (3)). In this regression, the zero-alpha restriction can be additionally imposed.

Thus, the marginal likelihood of returns and non-traded factors, ML, can be easily obtained by first taking the product of (3) and (4), and then integrating the product over the prior densities of w and other parameters (par). In the following sections, I build on this insight and formally derive Bayesian tests for evaluating an individual asset pricing model as well as comparing multiple models simultaneously.

9. General Test of a K-Factor Model

In this section, I develop a Bayesian framework to test the validity of a K factor asset pricing model (APM) against a general hypothesis that it is not true. For simplicity, I start with a K-factor model, where all of its factors are non-traded. The case of a general factor model that includes traded and non-traded factors can be easily extended, which I will discuss in the model comparison framework (next section). I use the following notations to represent the variables and parameters of the model throughout this section.

r_t denotes the $N \times 1$ vector of excess returns of N test assets at time t, f_t^N is the $K \times 1$ vector of K non-traded factors at time t, f_t^m is the $K \times 1$ vector of mimicking portfolios for the K non-traded factors, $R_{(T\times N)}$ is the $(T \times N)$ matrix form of excess returns for the N test assets over T periods, where each t^{th} row equals $r_t^{'}$, $F^N_{(T\times K)}$ is the $(T \times K)$ matrix form of K non-traded factors over T periods, where each t^{th} row equals $f_t^{N'}$, and $F^m_{(T\times K)}$ is the $(T \times K)$ matrix of mimicking portfolios for the K non-traded factors over T periods, where each t^{th} row equals $f_1^{m'}$.

An asset pricing model with non-traded factors can be expressed as two multivariate linear regressions. The first one is the regression of the non-traded factors, $F^N_{T\times K}$, on the excess returns of the test assets, $R_{T\times N}$,

$$F^N_{T\times K} = 1_{T\times 1}C^T + R_{T\times N}W_{N\times K} + \eta_{T\times K},\ vec(\eta_t) \sim MVN(0, \Sigma_\eta \otimes I), \tag{6}$$

where $MVN(\mu, \Sigma)$ denotes the multivariate normal density with mean μ, and covariance Σ. The variance specification is consistent with the standard assumption that errors are cross-sectionally correlated but uncorrelated in the time-series.[8] Recall that the mimicking portfolios, F^m, are weighted sums of the excess returns that are maximally correlated with their corresponding non-traded factors. In particular, $F^m_{T\times K} = RW$.[9]

The second one is the regression of N excess returns, $R_{T\times N}$, on the mimicking portfolios, $F^m_{T\times K}$,

$$R_{T\times N} = 1\alpha^T + F^m_{T\times K}B_{K\times N} + \epsilon_{T\times K}, \tag{7}$$

which is equivalent to,

$$R_{T\times N} = 1\alpha^T + RWB_{K\times N} + \epsilon_{T\times K}. \tag{8}$$

Regressing the set of N excess returns (R) on its own sub-space (RW), which has a smaller dimension (K), is equivalent to regressing only, and any, $N-K$ linearly independent excess returns on RW. This is because coefficients in this regression will uniquely determine the regression coefficients of the remaining K returns, given the mimicking portfolios RW.[10] Thus, the remaining K

[8] Because non-traded factors are usually the first-order innovations of various macroeconomic variables, it is reasonable to assume that the factors are uncorrelated in the time-series.

[9] To be precise, $F^m_{T\times K} = RW_m$, where W_m is the scaled matrix of W, where the each column sums to one. Thus, if W consists of columns W_j so that $W = \begin{bmatrix} W_1 & W_2 & \dots W_K \end{bmatrix}$, then $W_m = \begin{bmatrix} W_{m1} & W_{m2} & \dots W_{mK} \end{bmatrix}$, where each column $W_{mj} = W_j/1^TW_j$ and 1 is the vector of ones having the same dimension as any column W_j. However, I show that the final test statistic is invariant to such scaling of the weights. Thus, I obtain the same results whether I use RW_m or RW.

[10] This can be easily proved as below. Let $\bar{R}$ denote the considered $N-K$ returns and R^* denote the remaining K returns. Let $\bar{R} = \bar{\alpha} + RWB + \epsilon$, $\epsilon \sim MVN(0, \Sigma)$. Because $\{\bar{R}, RW\}$ span the entire cross-section of returns, there exist matrices C_1, C_2 such that $R^* = \bar{R}C_1 + RWC_2$. This implies, $R^* = \bar{\alpha}C_1 + RWBC_1 + RWC_2 + \epsilon C_1 \implies R^* = \bar{\alpha}C_1 + RW(BC_1 + C_2) + \epsilon C_1$. Thus, the regression coefficients of R^* are uniquely determined by the coefficients of $\bar{R}$.

returns can be ignored from the regression. Now, consider the regression of any $N-K$ returns, denoted by $\bar{R}$, on the mimicking portfolios :

$$\bar{R}_{T\times(N-K)} = 1\bar{\alpha}^T + RW\bar{B}_{K\times(N-K)} + \bar{\epsilon}_{T\times(N-K)},\ \bar{\epsilon} \sim MVN(0, \bar{\Sigma}_\epsilon \otimes I), \tag{9}$$

where $\bar{R}_{T\times(N-K)}$ can be expressed as $R\bar{I}$, for some $N\times(N-K)$ matrix $\bar{I}$. For example, if $\bar{R}$ denotes the returns of the last $N-K$ test assets, then $\bar{I}$ equals the last $N-K$ columns of the $N\times N$ identity matrix.

Finally, consider the following specification for the mimicking portfolios, given their true weights, W

$$RW|W = 1\alpha_{mim}^T + \epsilon_{mim},\ \epsilon_{mim} \sim MVN(0, \Sigma_{mim} \otimes I). \tag{10}$$

Because $[\bar{R}, RW]$=$(R[\bar{I}, W])$, where $[\bar{I}, W]$ is a full-rank $N\times N$ square matrix, (9) and (10) together determine the joint density of the excess returns given the corresponding parameters. I denote $[\bar{I}, W]$ by $\bar{I}_W$. Given the regressions in (6), (9), and (10), the null hypothesis (that the model is true) holds if and only if

$$H_0:\ \bar{\alpha} = 0, \quad \text{whereas, under the alternative } H_1:\ \bar{\alpha} \neq 0 \tag{11}$$

Now, I lay out the priors for all the required parameters to test the hypothesis in (11).

9a. Prior specification

The total set of parameters from the three regressions in (6), (9) and (10) are

$$\text{Parameters } = \{C, W, \Sigma_\eta, \bar{\alpha}, \bar{B}, \bar{\Sigma}_\epsilon, \alpha_{mim}, \Sigma_{mim}\}. \tag{12}$$

$\bar{\alpha}$ is the only restricted parameter, which equals zero when the null is true. All the other parameters are unrestricted and can take any values under both the hypothesis. However, the parameters $\{\bar{B}, \bar{\Sigma}\}$ vary with the choice of $N-K$ returns, $\bar{R}$. For example, when a different set of returns R^* are regressed on the mimicking portfolios, different coefficients $\{B^*, \Sigma^*\}$ will arise. Thus, Bayesian tests might depend on the choice of returns that are regressed on the mimicking portfolios. However, given the existence of one-one mapping between both the set of parameters $\{\bar{B}, \bar{\Sigma}\}$ and $\{B^*, \Sigma^*\}$ (see footnote 10 for a formal proof), in what follows, I construct novel, non-informative priors by imposing this mapping apriori. This yields a unified Bayesian test that is invariant to any choice of $(N-K)$ linearly independent subset of returns.

The traditional Jeffreys (1998) priors for the test assets' regression parameters do not impose this mapping apriori. Thus, applying these priors directly in the non-traded framework would yield undesirable Bayesian tests that violate the invariant property. Such tests may favor the null or alternative in unanticipated ways, depending on the choice of $\bar{R}$.

The novel priors also deliver another desirable property. Bayesian tests that entail the novel priors are invariant to the scale of factors. This property is particularly important to evaluate models containing either non-traded factors or PCs. Whereas the standard theories advocate that weights of mimicking portfolios and PCs should sum to one, researchers often use a variety of scaling procedures. For example, Kozak et al. (2019) scale the weights of various models' mean-variance efficient portfolio returns to have the same standard deviation as that of the market returns. Bayesian inferences that rely on the existing priors would be sensitive to the scale of factors.

I begin by specifying the unrestricted parameters $\{C, W, \Sigma_\eta\}$ with the traditional Jeffreys (1998) priors

$$P(C, W, \Sigma_\eta) \propto |\Sigma_\eta|^{-(K+1)/2} \text{ , under both } H_0,\ H_1. \tag{13}$$

These improper priors are defined only up to constants. But these parameters, and thereby constants, commonly appear under both the null and alternative and subsequently drop out. Thus, the traditional Jeffreys priors are justifiable for this set of parameters.

Rather than directly specifying the sets of other unconstrained parameters, $\{\bar{B}, \bar{\Sigma}\}$ and $\{\Sigma_{mim}\}$, with two traditional Jeffreys, I construct the novel priors that yield invariance as follows: i) I first specify the mean and covariance (Σ_R) of the test assets, R, with the traditional Jeffreys; ii) then, I establish a one-one correspondence between the sets of nuisance parameters, (Σ_R), and $\{\bar{B}, \bar{\Sigma}, \Sigma_{mim}\}$; iii) lastly, I use this one-one map to induce the priors for $\{\bar{B}, \bar{\Sigma}, \Sigma_{mim}\}$. By the usual property of Jeffreys, the induced priors would yield marginal likelihoods that are invariant to any reparametrization. Thus, the marginal likelihoods remain the same even when a different subset of test assets R^* (yielding different parameters $\{B^*, \Sigma^*, \Sigma_{mim}\}$) is used, as long as the priors for the parameters are induced as outlined above. The exact details of the prior construction and proof follow.

Consider the regression of test assets on a constant

$$R = \mu_R + \epsilon_R, \;\; \sim MVN(0, \Sigma_R \otimes I). \tag{14}$$

The priors in the spirit of BS for the mean and nuisance parameters of this regression are given by

$$P(\mu_R, \Sigma_R) = P(\mu_R|\Sigma_R) \times P(\Sigma_R) \propto P(\mu_R|\Sigma_R) \times |\Sigma_R|^{-(N+1)/2}. \tag{15}$$

Given W, $[\bar{R}, RW]$ $(= R\bar{I}_W)$ is a (full rank) linear transformation of R. Thus, by the usual Jeffreys rule, the induced priors for the variance parameters of $R\bar{I}_W$ would also be traditional diffuse. In particular, if

$$R\bar{I}_W = 1\bar{\mu}_{RW}^T + \epsilon_{RW}, \; \epsilon_{RW} \sim MVN(0, \bar{\Sigma}_{RW}), \tag{16}$$

then the induced prior for the parameters $\{\bar{\mu}_{RW}, \bar{\Sigma}_{RW}\}$ equals:

$$p(\bar{\mu}_{RW}, \bar{\Sigma}_{RW}) = P(\bar{\mu}_{RW}|\bar{\Sigma}_{RW}) \times P(\bar{\Sigma}_{RW}) \propto P(\bar{\mu}_{RW}|\bar{\Sigma}_{RW}) \times |\bar{\Sigma}_{RW}|^{-(N+1)/2} \tag{17}$$

Now, note that there is a one-one transformation between the nuisance parameters in (17), $\{\bar{\Sigma}_{RW}\}$, and the nuisance parameters in (9) and (10), $\{\bar{B}, \bar{\Sigma}_{\epsilon}, \Sigma_{mim}\}$. This is because the joint multivariate normal density of any two random vectors is equivalent to i) the linear regression of the first vector on the second, and ii) the linear regression of the second on a constant. Thus, I use this transformation and induce the priors for the parameters $\{\bar{B}, \bar{\Sigma}, \Sigma_{mim}\}$.

Proposition 1: *The induced priors for the parameters $\{\bar{B}, \bar{\Sigma}_\epsilon, \Sigma_{mim}\}$ based on the traditional Jeffreys prior for the parameters μ_R, Σ_R in (15) are then given by*

$$P(\bar{B}, \bar{\Sigma}_\epsilon, \Sigma_{mim}) \propto |\bar{\Sigma}_\epsilon|^{-(N+1)/2} \times |\Sigma_{mim}|^{-(2K-N+1)/2}. \tag{18}$$

Proof. The proof follows because of the one-one correspondence between the parameters under various reparametrizations. An analogous correspondence was first recognized by BS, which was later used by CZZ, in an entirely different context of comparing *multiple* traded factor models. Thus, the computations would be straightforward from the propositions 3 and 4 of CZZ. □

The final priors for the set of all nuisance parameters could be specified as

$$\begin{aligned} P(C, W, \Sigma_\eta) &= |\Sigma_\eta|^{-(K+1)/2}, \\ P(\bar{B}, \bar{\Sigma}_\epsilon, \Sigma_{mim}|W) &= |\bar{\Sigma}_\epsilon|^{-(N+1)/2}|\Sigma_{mim}|^{-(2K-N+1)/2}. \end{aligned} \tag{19}$$

Note that the novel priors differ from the traditional Jeffreys (1998). The novel priors modify the exponent of $\bar{\Sigma}_\epsilon$ to the total number of test assets, $(N+1)$, whereas the exponent based on the traditional Jeffreys (1998) equals the number of regressands, $(N - K^N)$. Similarly, the novel priors modify the exponent of Σ_{mim} to $(2K - N + 1)/2$, whereas the traditional Jeffreys (1998) has an exponent of $(K+1)$. Because the nuisance parameters commonly appear under both the null and the alternative, I use the same priors as in (19) under both hypothesis.

It is worth pointing out important differences between the novel priors and the more recent priors advocated by BS and CZZ (BS-CZZ priors).[11] These priors serve fundamentally different purposes. The BS-CZZ priors facilitate *simultaneous comparisons of multiple models* with traded factors when the models have different sets of nuisance parameters. These priors are not needed in the context of testing an individual asset pricing model with traded factors, because in such a framework, the same set of nuisance parameters (e.g., betas and residual covariances) appear under both the null and alternative hypothesis.

[11] See the next section for a formal definition of the BS-CZZ priors.

In contrast, this paper's priors are designed to test *individual* models with non-traded factors, because the traditional Jeffreys (1998) do not yield invariant inferences. Unlike traded-factor models, testing a non-traded model involves regression of returns on mimicking portfolios, which themselves belong to the return subspace. As a consequence, Bayesian tests with the conventional Jeffreys (1998) priors would be sensitive to the subsets of test assets that are chosen to regress on the mimicking portfolios.

Lastly, recall that $\bar{\alpha} = 0$ under the null. Under the alternative, I specify the following informative prior for $\bar{\alpha}$, which was used by BS

$$\bar{\alpha}|\bar{\Sigma}_{\epsilon}, W = MVN(0, k\bar{\Sigma}_{\epsilon}), \tag{20}$$

where the parameter $k > 0$. Economically, this informative prior was motivated in studies by MacKinlay (1995) and Pástor and Stambaugh (2000).[12] These studies stress the desirability of a positive relation between the absolute value of α and Σ, which makes extremely large Sharpe ratios and arbitrage opportunities less likely. Furthermore, note that $E(\alpha^{'}\Sigma^{-1}\alpha) = k \times N$. Thus, k reflects the prior belief about the expected increase in the Sharpe ratio ($E(\alpha^{'}\Sigma^{-1}\alpha)$) when K factors are added to the N test assets.

Because various studies such as Cochrane and Saa-Requejo (2000) argue that extreme Sharpe ratios as high as twice or a few multiples of the market portfolio's Sharpe ratio are usually arbitraged away and unlikely to survive, similar to BS I choose k such that the prior expectation of the maximum Sharpe ratio of the test assets (Sh^2_{max}) equals various multiples (p_m) of the Sharpe ratio of the $S\&P$ 500 returns (Sh^2_{mkt}). For example, p_m of 1.5 implies a value of $k = (1.5^2)Sh^2_{mkt}/N$. I consider different prior specifications yielding maximum (prior) Sharpe ratios of $\{1.25^2 \times Sh^2_{mkt}, 1.5^2 \times Sh^2_{mkt}, 1.75^2 \times Sh^2_{mkt}\}$.

In the similar spirit, I specify the following priors for the mean of the mimicking portfolios

[12] Kozak et al. (2019) use a slightly different prior. In my framework, their prior translates to $P(\bar{\alpha}|\bar{\Sigma}_{\epsilon}, W) = MVN(0, k\bar{\Sigma}^2_{\epsilon})$.

under both the null and alternative

$$\alpha_{mim}|\Sigma_{mim}, W = MVN(0, k\Sigma_{mim}). \tag{21}$$

This economically motivated prior also offers an additional advantage. It shields against potential spurious non-traded factors, which are uncorrelated with the cross-section of stock returns. The sample Sharpe ratio estimates of spurious factors' mimicking portfolios are not well-behaved and tend to overestimate their true values more often than it should. The prior in (21) alleviates this concern by shrinking the Sharpe ratios toward the Sharpe ratio of the market portfolio.[13] Large and economically implausible values of k overweight the evidence from sample estimates of mimicking portfolios' Sharpe ratios, which could unduly favor models containing spurious factors. In the empirical section, I provide results from an extensive simulation analysis to show that my methodology is robust to spurious factors.

Finally, I use the prior specifications in (19), (20), and (21) to derive marginal likelihoods under both the null and alternative.

9b. Bayes Factor

In a Bayesian framework, the Bayes Factor (BF) quantifies the null hypothesis's relative likelihood against the alternative. It is the ratio of marginal likelihoods (ML) under the null and alternative, $ML(H_0/ML(H_1)$, where each ML is the likelihood of data obtained by integrating over the prior density of the parameters under the corresponding hypothesis. MLs are usually not well-defined under improper priors, as their marginalization constants could be arbitrarily specified. However, recall that the nuisance parameters that are specified with the improper priors commonly appear under both the hypothesis. Thus, these marginalization constants drop out in the computation of BF.

[13] In the cross-sectional framework, Kan and Zhang (1999) have shown that conventional procedures tend to overestimate the true risk-premia of spurious factors, which technically should be zero. Note that risk-premia of non-traded factors are functions of their mimicking portfolios' true means and Sharpe ratios (see Lewellen et al. (2010)). By shrinking both of them, the prior in (21) shrinks the risk-premia towards zero, thereby alleviating the problem of spurious factors.

Because the observed data are set of excess returns, (R), and non-traded factors, F^N,

$$ML(H_i) = P(F^N, R|H_i) = \int P(F^N, R|par, H_i)P(par|H_i), \tag{22}$$

where $par = \{W, c, \Sigma_\eta, \alpha, \bar{B}, \bar{\Sigma}_\epsilon\}$ and $P(par|H_i)$ is the prior density of the parameters under H_i. Furthermore,

$$P(F^N, R|par, H_i) = P(F^N|R, W, c, \Sigma_\eta, H_i) \times P(R|W, \alpha, B, \Sigma_\epsilon, H_i). \tag{23}$$

Because both the conditional likelihoods, the non-traded factors given the excess returns, and the excess returns given the mimicking portfolios are known from the equations (6), (9), & (10), MLs and BF can be easily derived. In particular, proposition-2 shows that these can be readily obtained by modifying the MLs of BS and CZZ. Denoting the ML under the null that imposes the zero alpha restriction by ML_R, and under the alternative that does not impose the restriction by ML_U, I restate the following result by BS for notational convenience.

Result: The unrestricted marginal likelihood, $ML_U(R, F; T-K)$, the restricted marginal likelihood, $ML_R(R, F; T-K)$, and the Bayes Factor, BF^T, for an asset pricing model with K traded factors, F, and N test assets, R are given by

$$ML_R(R, F; T-K) \propto |F'F|^{-N/2}|S_R|^{-(T-K)/2}, \tag{24}$$

$$ML_U(R, F; T-K) \propto |F'F|^{-N/2}|S|^{-(T-K)/2}Q, \tag{25}$$

$$BF^T = \frac{ML_R(R, F; T-K)}{ML_U(R, F; T-K)}, \tag{26}$$

where S and S_R are the $N \times N$ cross-product matrices of the estimated ordinary least squares (OLS) residuals in the multivariate regression of R on F, with and without the intercepts, respectively. $(T-K)$ is the degrees of freedom. Furthermore, the scalar Q is a function of the GRS-statistic. BF^T can be used to test the validity of an asset pricing model with traded factors.

For models containing non-traded factors, the marginal likelihoods require two additional adjustments. The first adjusts for the estimation uncertainty in the mimicking portfolios and the second to ensure invariance across $N-K$ linearly independent subsets of the test assets. I formally state the proposition below:

Proposition 2: *The unrestricted marginal likelihood, $ML_U^N(R, F^N)$, and the restricted marginal likelihood, $ML_R^N(R, F^N)$, for a model with K non-traded factors, F^N, and excess returns, R, are proportional to*

$$ML_U^N(R, F^N) = E\left[C_I \times ML_U(\bar{R}, RW, T) \times ML_U(RW, 1; T-(N-K))\right], \tag{27}$$

$$ML_R^N(R, F^N) = E\left[C_I \times ML_R(\bar{R}, RW, T) \times ML_U(RW, 1; T-(N-K))\right], \tag{28}$$

where RW are the unknown mimicking portfolios of the non-traded factors F^N, $\bar{R}$ is any subset of $N-K$ linearly independent returns from R. The *invariant-constant*, C_I, which ensures invariance, is a function of the mimicking portfolio weights. In particular, $C_I = |\bar{I}_W|^T$, where $\bar{I}_W$ satisfies $[\bar{R}, RW] = R\bar{I}_W$. $E(.)$ denotes the expectation, which is taken over the random variable W, which has the following density

$$g(vec(W)) \sim MVN\left(vec(\hat{W}), \Sigma_\eta\right), \; \Sigma_\eta \sim IW\left(S_{F^N}, T-N-1\right), \tag{29}$$

where vec denotes the vectorized form of a matrix (stacking up all columns under one column). $\hat{W}$ and S_{F^N} are the OLS estimates of the true mimicking portfolio weights (regression coefficients) and $K \times K$ cross-product of residuals in the regression of non-traded factors F^N on the excess returns R, respectively. MVN and IW are notations for the multivariate normal and Inverse-Wishart densities respectively.

Proof. See Appendix Xa. □

The expectation, $E(.)$, adjusts for the estimation uncertainty in the mimicking portfolios by averaging (integrating) the MLs of BS and CZZ over the density of the true mimicking portfolio weights (W) given in (29). Conditional on W, ML_R, and ML_U could be interpreted as "explained return variances" by the restricted and unrestricted hypothesis. The explained-variance of a hypothesis is sensitive to the scale of data. One could make it arbitrarily big by multiplying the data with an appropriate constant. Thus, inferences solely based on the expected MLs could be misleading.

However, the "proportion of explained-variance", which equals the ratio of the explained-variance to the total variance of the data, would be invariant to such a scaling. My methodology favors the hypothesis with the best proportion of explained-variance. I show in the appendix, as well as in the next section, that the invariant-constant, C_I, is proportional to the inverse of the total variance of the data, and thereby yielding the invariant result.

Finally, BF for testing the validity of a model with non-traded factors is given by

$$BF^N = \frac{ML_R^N(R, F^N)}{ML_U^N(R, F^N)} = \frac{E\left[C_I \times ML_R(\bar{R}, RW, T) \times ML_U(RW, 1, T-(N-K))\right]}{E\left[C_I \times ML_U(\bar{R}, RW, T) \times ML_U(RW, 1, T-(N-K))\right]}. \tag{30}$$

Arguably, one may consider testing a non-traded model using the traded BF (BF^T) of BS by directly treating the sample-estimated mimicking portfolios as traded factors without adjusting for their uncertainty. BF^N in (30) suggests that such tests could be misleading. In fact, BF^N reduces to BF^T only when the test assets and traded factors perfectly span the non-traded factors, or equivalently $\Sigma_\eta = 0$. Because the portfolio returns poorly span the majority of non-traded factors, studies that do not account for this estimation uncertainty could lead to misleading inferences. Examples of such studies include Adrian, Etula, and Muir (2014), and Ang, Hodrick, Xing, and Zhang (2006), and Vassalou (2003), among many others.

Although the expectations in (30) are not analytically solvable, these can be easily estimated from the Monte Carlo Integration approach of Geweke (1988, 1989). A similar approximation has also been employed by Harvey and Zhou (1990). They accurately estimate high (90) dimensional integrals to test the mean-variance efficiency of a given portfolio. In particular, BF^N can be approximated by

$$BF^N \approx \frac{\sum_{l=1}^{L} |\bar{I}_{W_l}|^T ML_R(\bar{R}, RW_l, T) ML_U(RW_l, 1, T-(N-K))}{\sum_{p=1}^{P} |\bar{I}_{W_p}|^T ML_U(\bar{R}, RW_p, T) ML_U(RW_p, 1, T-(N-K))}, \tag{31}$$

where $\{W_l\}_{l=1}^{L}$ and $\{W_p\}_{p=1}^{P}$ are independent draws from the density of the mimicking portfolio weights given in (29). This procedure can be computationally intensive, as it requires a large number of draws from independent and separate samples to first estimate the numerator, and then the denominator. To minimize the computational burden, I use the separate-ratio estimator, which directly estimates the ratio of the expectations from a large stratified sample.[14] In particular, given a large number of draws $\{W\}_{g=1}^{G}$ from the given in (29), the ratio estimator first splits the sample draws into several strata S_g, estimates the ratio of the means in each strata, and then takes the average of such ratio estimates across all the strata. The estimator can be expressed as

$$BF^N \approx \sum_{Sg=1}^{S} \frac{\sum_{l\in S_g} |\bar{I}_{W_l}|^T ML_R(\bar{R}, RW_l, T) ML_U(RW_l, 1, T-(N-K))}{\sum_{l\in S_g} |\bar{I}_{W_l}|^T ML_U(\bar{R}, RW_l, T) ML_U(RW_l, 1, T-(N-K))}, \tag{32}$$

where each sample draw W_{mg}, belongs to the stratum S_g. Note that the ratio estimate under each S_g, $\frac{\sum_{l\in S_g} |\bar{I}_{W_l}|^T ML_R(\bar{R},RW_l,T) ML_U(RW_l,1,T-(N-K))}{\sum_{l\in S_g} |\bar{I}_{W_p}|^T ML_U(\bar{R},RW_l,T) ML_U(RW_l,1,T-(N-K))}$, is a biased estimate of BF^N due to Jensen's inequality. Although such a bias could be insignificant for a large number of sample draws, the separate ratio estimator, which takes the mean of these estimates, reduces the bias at an even quicker rate and thus makes the estimator more efficient.

I conduct estimation with 100,000 draws by splitting the sample into 100 equal strata with each strata comprising 1000 samples. I find that the estimates are quite precise, where the maximum standard error of BF^N is in the order of 10^{-4}. This result is in the spirit of Harvey and Zhou (1990), who also find that for a large number of sample draws (100,000), Monte-Carlo integration delivers five digits of precision.

10. Comparing Asset Pricing Models with Non-Traded Factors

Given a set of models $\{M_1, M_2, \ldots, M_n\}$, this section develops a general framework to measure each model's relative likelihood in summarizing the cross-section of expected returns compared with the other models. The framework involves computing posterior probabilities of the models, which

[14] Chapter 6 from "The Elementary Survey Sampling" textbook summarizes the properties of the separate-ratio estimator (Scheaffer, III, Ott, and Gerow (2011)). The bias of this estimator diminishes at a higher rate than the conventional Monte-Carlo estimator. The variance of the estimator is available in a closed-form expression and thus can be used to evaluate the precision of the estimation procedure.

are defined by

$$P(M_i|Data) = \frac{P(Data|M_i)P(M_i)}{\sum_{j=1}^{n} P(Data|M_j)P(M_j)} = \frac{ML(i)P(M_i)}{\sum_{j=1}^{N} ML(j)P(M_n)}, \tag{33}$$

where $P(M_i)$ is the prior probability that M_i is the true model, and $ML(i)$ is the marginal likelihood under M_i. Thus, under the prior that the models are equally likely to be true, model comparison is equivalent to ranking models based on their marginal likelihoods.

Comparing multiple models with MLs is not straightforward. As discussed in the previous section, MLs computed using the conventional improper priors are not properly defined when the underlying parameters are not common across the models. In addition, two fundamental issues arise for models containing non-traded factors. First, mimicking portfolios have estimation uncertainty. Second, evaluating a non-traded factor model involves regression of returns on the mimicking portfolios, which themselves belong to the returns subspace.

In what follows, I derive an essential set of results and propose novel, non-informative priors that permit model comparisons with non-traded factors. Importantly, these priors yield a conditional test assets irrelevance result for models containing non-traded factors. This result implies that test assets are required only for identifying mimicking portfolios. Conditional on the mimicking portfolios, test assets would be irrelevant for model comparisons. Barillas and Shanken (2017) have drawn similar conclusions for several frequentist-based model comparison methodologies, including the differences in likelihoods and Sharpe ratios. Using the novel, non-informative priors, I formally prove the conditional test assets irrelevance in a Bayesian framework.

Let $\{M_1, M_2, \ldots, M_n\}$ be the set of models to be compared. Each model M_j in the set can include both traded and non-traded factors. Also, in the spirit of CZZ, I allow for the possibility that the market factor can be excluded from the models. This specification, which differs from the BS framework, is important because models might not always include the market portfolio as one of their factors (e.g., He et al. (2017)). F_j denotes the set of traded factors included in M_j, and F_j^* denotes the set of traded factors excluded from M_j but included in any other model from the set. Let $F = [F_j, F_j^*]$. F_j^N denotes the set of non-traded factors included in M_j, and F_j^{N*} denotes the set of non-traded factors excluded from M_j but included in any other model from the set. Let W be the weights of the true mimicking portfolios F^m of all the non-traded factors ($F^N = \{F_j^N, F_j^{N*}\}$) across the models. F_j^m denotes the set of mimicking portfolios of the included non-traded factors, F_j^N, and F_j^{m*} denotes the mimicking portfolios of the excluded non-traded factors, F_j^{N*}. Finally, let R be the excess returns of N test assets, K^T and K^N be the total number of traded factors and non-traded factors across all the models, respectively.

When M_j is the true asset pricing model, F_j and F_j^m will price R, F_j^*, and F_j^{m*}. This implies,

$$\begin{aligned}&\alpha_{rj}^* = 0 \text{ in the regression, } \{R, F_j^{m*}, F_j^*\} = 1(\alpha_{rj}^*)^T + [F_j, F_j^m]\beta_{rj}^* + \epsilon_j^*, \epsilon_j^* \sim MVN(0, \Sigma_j^* \otimes I),\\ &\alpha_j \neq 0 \text{ in the regression, } \{F_j^m, F_j\} = 1\alpha_j^T + \epsilon_j, \epsilon_j \sim N(0, \Sigma_j \otimes I), \qquad (34)\end{aligned}$$

where the set of mimicking portfolios, $F^m = \{F_j^m, F_j^{m*}\}$, are a linear combination of the test assets and traded factors. In particular, $F^m = [R, F]W$. The weights, W, are proportional to the coefficients in the regression of the non-traded factors on the test assets and traded factors

$$F^N = c + [R, F]W + \eta, \;\; \eta \sim MVN(0, \Sigma_\eta). \qquad (35)$$

Note that the test-asset returns and traded factors together constitute a return space of dimension $N + K^T$. Given a set of K^T traded factors and K^N mimicking portfolios, there are only $N - K^N$ linearly independent test assets. Thus, M_j holds whenever the condition in (34) holds for any $N - K^N$ linearly independent set of test-asset returns, $\bar{R}$. This is equivalent to the conditions,

$$\bar{\alpha}_{rj}^* = 0, \;\text{ where } \{\bar{R}, F_j^{m*}, F_j^*\} = 1(\bar{\alpha}_{rj}^*)^T + [F_j, F_j^m]\bar{\beta}_{rj}^* + \epsilon_j^*, \;\; \epsilon_j^* \sim MVN(0, \Sigma_{rj}^* \otimes I), \qquad (36)$$

$$\alpha_j \neq 0, \;\text{ where } \{F_j^m, F_j\} = 1\alpha_j^T + \epsilon_j, \epsilon_j \sim N(0, \Sigma_j \otimes I). \qquad (37)$$

In the spirit of previous section, I construct priors in such a way that the marginal likelihoods of the models are invariant to the choice of $N - K^N$ returns $\bar{R}$, which are used in the regression (37).

Moreover, the restriction $\bar{\alpha}_{rj}^* = 0$ implies that the intercepts, $\bar{\alpha}_r$, and α_j^*, in the following regressions equal zero.

$$\bar{\alpha}_r = 0 \text{ in the regression of } \bar{R} = 1\bar{\alpha}_r^T + [F_j, F_j^m, F_j^*, F_j^{m*}]\bar{\beta}_r + \bar{\epsilon}_r, \bar{\epsilon}_r \sim MVN(0, \bar{\Sigma}_r \otimes I) \tag{38}$$

$$\alpha_j^* = 0 \;\text{ in the regression of } \{F_j^{m*}, F_j^*\} = 1(\alpha_j^*)^T + [F_j, F_j^m]\beta_j^* + \epsilon_j^*, , \epsilon_j^* \sim MVN(0, \Sigma_j^* \otimes I), \tag{39}$$

$$\alpha_j \neq 0 \text{ in the regression of } \{F_j^m, F_j\} = 1\alpha_j^T + \epsilon_j, \epsilon_j \sim MVN(0, \Sigma_j \otimes I). \tag{40}$$

The marginal likelihoods or posterior probabilities that I derive in this section quantify how likely the models satisfy the conditions in (38), (39) and (40). Because the first regression ($\bar{\alpha}_r = 0$) commonly appears across the models, it drops out in the computation of posterior model probabilities when the true mimicking portfolios are known. Thus, test assets will be irrelevant for comparing models that comprise only traded factors. When the true mimicking portfolios are unknown, test assets do not automatically drop out because the MLs involve integrating over the prior density of unknown mimicking portfolios. Thus, my framework fundamentally differs from BS, where they use the test assets irrelevance to derive the marginal likelihoods.

10a. Priors for Comparing Asset Pricing Models

The total set of parameters under M_j from the regressions (38), (39) and (40) are

$$Parameters_j = \{c, W, \Sigma_\eta, \bar{\beta}_r, \bar{\Sigma}_r, \beta_j^*, \Sigma_j^*, \alpha_j, \beta_j, \Sigma_j\}. \tag{41}$$

Note that the parameters $\{c, W, \Sigma_\eta\}$ commonly appear across the models. Thus, I first specify these parameters with the traditional Jeffreys (1998) priors,

$$P(c, W, \Sigma_\eta) \propto |\Sigma_\eta|^{-(K^N+1)/2}. \tag{42}$$

Even the parameters $\{\bar{\beta}_r, \bar{\Sigma}_r\}$ commonly appear across the models. So, one may wonder whether these parameters could be specified with the traditional Jeffreys (1998) priors. However, recall that these parameters vary with the choice of subset of test assets, $\bar{R}$. Thus, the traditional Jeffreys (1998) priors require modification to ensure that the resultant Bayesian tests are invariant across the subsets of test assets. In particular, the traditional Jeffreys (1998) priors for $\{\bar{\beta}_r, \bar{\Sigma}_r\}$, given F, F^m are

$$P_{Jeffreys}(\bar{\beta}_r, \bar{\Sigma}_r)|F, F^m = \det\left([F, F^m]^T [F, F^m]\right)^{(N-K^N)/2} |\bar{\Sigma}_r|^{-(N-K^N+1)/2}. \tag{43}$$

As in the previous section (19), I modify the exponent of $\bar{\Sigma}_r$ from the number of regressands in (38), $(N - K^N + 1)$, to the total number of test assets and traded factors, $(N + K^T + 1)$. The modified priors take the following form

$$P(\bar{\beta}_r, \bar{\Sigma}_r)|F, F^m = \det\left([F, F^m]^T [F, F^m]\right)^{(N-K^N)/2} |\bar{\Sigma}_r|^{-(N+K^T+1)/2}. \tag{44}$$

The premise of these priors is to first start with the Jeffreys (1998) priors for the covariance matrix of the test assets and traded factors. And then obtain the induced priors for the coefficients in the regression of $\bar{R}$ on $[F, F^m]$, given $[F, F^m]$. I call the set of all nuisance parameters that commonly appear across the models as "global-nuisance" parameters. These are $\{c, W, \Sigma_\eta, \bar{\beta}_r, \bar{\Sigma}_r\}$.

From the "model-specific" parameters, $\{\beta_j^*, \Sigma_j^*, \alpha_j, \beta_j, \Sigma_j\}$, I specify the alphas with the previously used informative priors, as models impose restrictions on them,

$$\alpha_j|\{W, \beta_j, \Sigma_j\} \propto MVN(0, k\Sigma_j). \tag{45}$$

The remaining model-specific nuisance priors are unrestricted and can take any values. So, one might again wonder whether the traditional Jeffreys (1998) priors are natural candidates. Unlike the global-nuisance parameters, the model-specific priors are not common across the models. For example, consider a set of factors $\{Mkt, SMB,\ HML, svar^m\}$, where $svar^m$ is the mimicking portfolio for the non-traded factor stock market variance $(svar)$. The ML computation for the model, $M_1 = \{Mkt, SMB\}$, first requires regression of the excluded factors, $\{HML, svar^m\}$, on the included factors, $\{Mkt, SMB\}$, and then the regression of the included factors, $\{Mkt, SMB\}$, on a constant. Thus, the dimensions of the corresponding parameters are : $\beta_1^* \sim 2 \times 2$; $\Sigma_1^* \sim 2 \times 2$; and $\Sigma_1 \sim 2 \times 2$. However, for a different model $M_2 = \{Mkt, HML, svar^m\}$, $\beta_2^* \sim 1 \times 3$; $\Sigma_2^* \sim 1 \times 1$;

and $\Sigma_2 \sim 3 \times 3$. Recall that MLs are defined only up to arbitrary constants with improper priors. Because the parameters between the models are different, these constants do not drop out while comparing multiple models. Thus, model comparisons with the traditional Jeffreys (1998) yield misleading inferences.

However, BS show that there is a one-to-one mapping from the nuisance parameters of a model to the nuisance parameters of any other. They note that the marginalization constants are preserved under the corresponding change-of-variables. CZZ build on this insight and propose modified improper priors that permit ML-based model comparisons. I use their frameworks to derive the priors for the model-specific nuisance parameters. Recognizing their respective contributions, I call these modified improper priors as "BS-CZZ" priors.

The premise of the BS-CZZ priors is to first specify the parameters of any particular model with the conventional improper priors and then derive the priors for the other models' parameters by applying the corresponding one-to-one mapping formula. The exact computations follow.

First, conditional on the true mimicking portfolio weights, W, I specify the covariance matrix of all the traded factors and mimicking portfolios with the traditional Jeffreys (1998),

$$\begin{aligned}&[F, F^m] = 1\alpha_{F,F^m}^T + \epsilon_{F,F^m}, \epsilon_{F,F^m} \sim MVN(0, \Sigma_{F,F^m} \otimes I),\\&P\left(\Sigma_{F,F^m}|W\right) = |\Sigma_{F,F^m}|^{-(K^N+K^T+1)/2}.\end{aligned} \tag{46}$$

By applying the appropriate one-one-mapping formula to (46), the priors for the parameters in various models could be obtained. In the spirit of CZZ, the model-specific priors are given by

$$\{\beta_j^*, \Sigma_j^*, \beta_j, \Sigma_j\}|W = |\Sigma_j^*|^{-(K^T+K^N+1)/2} \left|\Sigma_j\right|^{-\frac{2K_j-K^T-K^N+1}{2}}. \tag{47}$$

Thus, (42), (44), (45), and (47) summarize the priors for all the required parameters.

Before obtaining the marginal likelihoods, it is worth emphasizing a critical discussion of what priors to use, when, and why, to compare asset pricing models in general. Recall that I *induce* the priors for the parameters under all the models using 3 traditional Jeffreys (1998) that correspond to the following hierarchical regressions (in that order): i) F^N on $\{R, F\}$; ii) R on $\{F, F^m\}$; iii) $\{F, F^m\}$ on a constant.

Alternatively, one may consider inducing the priors using various other specifications. For example, one could induce using a single, joint Jeffreys (1998) that corresponds to the regression of the entire data, $\{F^N, R, F\}$, on a constant. Or one could induce using two traditional Jeffreys (1998) that correspond to the regressions - F^N on $\{R, F\}$, and $\{R, F\}$ on a constant. I prove that such specifications yield intuitively puzzling and misleading inferences. Due to the inherent hierarchy in the regressions, and the fact that the non-traded factors F^N, test assets R, and traded factors F play distinct roles in model comparisons, it is *necessary* to induce the priors from the three independent, hierarchical regressions, as in this paper.

For example, the mimicking portfolios should always be obtained by regressing F^N on $\{R, F\}$, but not, say, $\{R, F\}$ on F^N. Similarly, the set of all traded factors and mimicking portfolios, $\{F, F^m\}$, should always price the test assets R, whereas the test assets need not price the set of all factors. The alternative prior specifications entertain these economically implausible scenarios, thereby yielding paradoxical inferences. To conserve space and not overwhelm readers with technical details, I present these paradoxes in the Internet Appendix-I.[15]

Finally, using the priors in (42), (44), (45), (46), and (47), I obtain the following marginal likelihoods for models containing non-traded factors.

Proposition 3: *The marginal likelihood for the model M_j comprising the traded factors F_j and non-traded factors F_j^n is given by,* $EML_j =$

$$E\left[C_I \times ML_R\left([F_j^*, F_j^{m*}], [F_j, F_j^m]; T\right) \times ML_U\left([F_j, F_j^m], \mathbf{1}; T - (K^T + K^N - K_j)\right)\right], \tag{48}$$

where F_j^{m*}, and F_j^m are the unknown mimicking portfolios for the non-traded factors F_j and F_j^*, respectively. In particular, $F_j^{m*} = [R, F]W_j^*$ and $F_j^m = [R, F]W_j$. F, F^m are total set of traded factors and mimicking portfolios across all the models, respectively. $ML_R(Y, X; v)$ and

[15] My extensive conversations with Jay Shanken and a more recent paper by Barillas and Shanken (2020) have been a significant influence on the development of this section. Special thanks to Jay Shanken for his insightful comments. All errors are mine.

$ML_U(Y, X; v)$ denote the restricted and the unrestricted MLs of BS and CZZ, expressions of which are given in (24). Expectation is taken with respect to the following density of the true mimicking portfolio weights W

$$g(vec(W)) \sim MVN\left(vec(\hat{W}), \Sigma_\eta\right),\ \Sigma_\eta \sim IW\left(S_{F^N}, T - N - K^T - 1\right), \tag{49}$$

where vec denotes the vectorized form of a matrix (stacking up all columns under one column). $\hat{W}$ and S_{F^N} are the OLS estimates of the true mimicking portfolio weights (regression coefficients) and $K \times K$ cross-product of residuals in the regression of non-traded factors F^N on the excess returns R and traded factors F, respectively. MVN and IW are notations for the multivariate normal and Inverse-Wishart densities, respectively. The invariant-constant is given by

$$C_I = det\left([F, F^m]'[F, F^m]\right)^{T/2}, \tag{50}$$

where $det(.)$ denotes the determinant of a square-matrix. Moreover, the product of terms inside the expectation $(E(.))$ is independent of W for models exclusively comprising traded factors. This result is consistent with the intuition that traded factor models would not require uncertainty adjustment, as their mimicking portfolios are completely known.

Proof. See Appendix Xb. □

Thus, the posterior model probability of M_j is given by

$$P(M_j|Data) = \frac{EML_j}{\sum_{j=1}^{N} EML_j}. \tag{51}$$

Because the marginal likelihoods for models containing non-traded factors can be expressed as constant-adjusted expectations of the BS and CZZ MLs, I denote these by "EML"s. EML refers to Expected Marginal Likelihood. Henceforth, I denote the marginal likelihoods based on BS and CZZ by ML. Here, the estimated mimicking portfolios are substituted for the non-traded factors without adjusting for their estimation uncertainty by MLs. To be precise, these MLs are given by

$$ML_j = ML_R\left([F_j^*, F_j^{m*}], [F_j, F_j^m]; T\right) \times ML_U\left([F_j, F_j^m], \mathbf{1}; T - (K^T + K^N - K_j)\right). \tag{52}$$ [16]

In the simulation and empirical sections of the paper, I compare the contrasting results delivered by the ML and EML measures.

As in the previous section, the expectation adjusts for the estimation uncertainty of the mimicking portfolios. The invariant-constant, C_I, ensures invariance with respect to chosen subsets of the test assets, $\bar{R}$. It is evident from (50) that C_I relates to the total variance of the factors and mimicking portfolios. Thus, the EMLs capture the *proportion of explained-variance*, which is invariant to the scale of data and choice of the subset of test assets. Whereas test assets are required to obtain the posterior densities of the weights of the mimicking portfolios, conditional on the mimicking portfolios, $EMLs$ are independent of the test assets. Thus, the *conditional test assets irrelevance* result is accomplished.

Both the "invariance" and "irrelevance" results follow from the novel prior specification developed in this paper and would not be achievable with the existing priors. Rather than using the novel priors, one may consider applying the BS-CZZ framework as it is to compare models containing non-traded factors. Inferences from such a framework could be misleading.

In particular, the BS-CZZ framework specifies the model-specific nuisance parameters with the similar priors used in this paper (47). However, their framework specifies the global-nuisance parameters with the traditional Jeffreys (1998) priors, as in (43). When the models exclusively comprise traded factors, these priors achieve test-assets irrelevance and yield valid model comparisons. However, when the true mimicking portfolios are unknown, such priors yield the following marginal likelihood for M_j containing non-traded factors

$$\begin{aligned} EML_j^{BS\text{-}CZZ} &= E\left[|\bar{I}_W|^T ML_R\left(\bar{R}, [F, F^m]; T - (K^T + K^N)\right)\right. \\ &\quad \left. ML_R\left([F_j^*, F_j^{m*}], [F_j, F_j^m]; 1; T\right) \times ML_U\left([F_j, F_j^m]\right); T - (K^T + K^N - K_j)\right]. \end{aligned} \tag{53}$$

[16] These MLs differ, albeit slightly, from the original expressions derived by BS, because my framework entertains the possibility that the market portfolio need not always be included across all the models.

Given the subset of test assets $\bar{R}$, consistent with CZZ, these marginal likelihoods share the same marginalization constant across the models and thus seem to deliver valid model comparisons. However, the marginal likelihoods would be sensitive to the choice of subset of the test assets $\bar{R}$. As a consequence, these marginal likelihoods favor one model or another in unanticipated ways, depending on the choice of $\bar{R}$. In fact, simulations suggest that the model comparisons with these measures are more misleading than the comparisons with the MLs of BS and CZZ that do not explicitly account for the mimicking portfolios' estimation uncertainty.[17]

11. Comparing Models with Principal Components

This section further generalizes the methodology developed in the previous sections to compare models containing PCs. PCs of a given set of test assets are various weighted sums of excess returns of the test assets. The weights are determined by the true return covariances, which are unknown. Thus, the framework for comparing models with PCs precisely matches the non-traded framework, where the unknown weights of PCs take the role of unknown mimicking portfolio weights. As in the previous section, I show that marginal likelihoods for models with PCs can be expressed as modified marginal likelihoods of BS and CZZ. The modification explicitly adjusts for the estimation error of returns' covariance, as well as ensures invariance.

Let $\{M_1, M_2, \ldots, M_n\}$ be the set of models to be compared. The models could comprise traded, non-traded factors and up to the first P PCs of a given set of test assets, R. I borrow the standard notations from the previous section. F_j (F_j^*) are the set of included (excluded) factors in (from) M_j. F_j^N (F_j^{N*}) are the set of included (excluded) non-traded factors, whose mimicking portfolios are F_j^m (F_j^{m*}). Let F_j^P (F_j^P) denote the set of PCs included in M_j. The PCs and mimicking portfolios can be expressed as linear combinations of the test assets, R, and traded factors, F (=$[F_j, F_j^*]$). Thus, $F^P = RW^P$,[18] and $F^m = [R, F]W$, where W^P and W are the unknown, true weights of the PCs and mimicking portfolios. Denote the included (excluded) mimicking portfolio weights by

[17] To conserve space, I do not report these results in the paper. However, the results are available upon request.

[18] This specification is consistent with the standard practice of extracting the PCs exclusively from the test-asset returns, excluding the traded factors. One could easily entertain PCs to use information from the traded factors as well, in which case $F^P = [R, F]W^P$. The generalization is straightforward.

W_j (W_j^*) and PCs by W_j^P (W_j^{P*}). And let K^T (K^N) be the total number of traded (non-traded) factors F.

When M_j is the true model, alphas in the regression of the test assets on the set of all factors and the regression of the excluded factors on the included factors must equal zero. This implies,

$$\alpha_{rj}^* = 0 \ \text{ where } \{R, F_j^*, F_j^{m*}, F_j^{P*}\} = \alpha_{rj}^* + \beta_{rj}^*[F_j, F_j^m, F_j^P] + \epsilon_{rj}^*, , \epsilon_{rj}^* \sim N(0, \Sigma_{rj}^*), \tag{54}$$

$$\alpha_j \neq 0, \ \text{where } \{F_j^P, F_j\} = \mathbf{1}\alpha_j^T + \epsilon_j, \epsilon_j \sim N(0, \Sigma_j). \tag{55}$$

Because the returns are regressed on their own sub-space, as in the previous section, the following conditions should hold for any $(N - K^N - P)$ linearly independent test assets, $\bar{R}$.

$$\bar{\alpha}_r = 0, \text{ where } \bar{R} = 1\bar{\alpha}_r^T + [F_j, F_j^m, F_j^P, F_j^*, F_j^{m*}, F_j^{P*}]\bar{\beta}_r + \bar{\epsilon}_r, \bar{\epsilon}_r \sim MVN(0, \bar{\Sigma}_r \otimes I), / \tag{56}$$

$$\alpha_j^* = 0 \ , \text{ where } \{F_j^*, F_j^{m*} F_j^{P*}\} = 1(\alpha_j^*)^T + [F_j, F_j^m, F_j^P]\beta_j^* + \epsilon_j^*, \epsilon_j^* \sim MVN(0, \Sigma_j^* \otimes I), \tag{57}$$

$$\alpha_j \neq 0, \text{ where } \{F_j, F_j^m, F_j^P\} = \mathbf{1}\alpha_j^T + \epsilon_j, \epsilon_j \sim MVN(0, \Sigma_j \otimes I). \tag{58}$$

Given the exact resemblance of these equations with the excluded and included regressions examined in (38), (39) and (40) to compare models with non-traded factors, priors for the parameters in (56), (57) and (58) can be easily obtained, as in the previous specification. Before laying out the priors, a special case of comparing multiple models that exclusively comprise PCs as factors deserves emphasis.

Consider a simple example of comparing the model comprising the first three PCs (PC1-3) with the model containing the first five PCs (PC1-5). Because the PCs 4-5 are orthogonal to the PCs 1-3 by definition, betas in the regression of the PCs 4-5 on the PCs 1-3 are identically equal zero. Thus, testing for the zero-alphas in this regression is equivalent to merely testing that the PCs 4-5 have zero means. While this could instinctively appear a little puzzling, I provide a simple and economically meaningful explanation to demonstrate the result. When PC1-3 is the true model, the risk-premia of the PCs 4-5 should be proportional to their corresponding exposures to the PCs 1-3. Because these exposures are zero by definition, the risk-premia would be zero. PCs are tradable portfolios, and thus zero risk-premia would imply that the PCs 4-5 have zero means. My methodology holds even under this particular sub-category of comparing models that exclusively comprise PCs.

11a. Priors for Comparing Asset Pricing Models with Principal Components

Given the true weights of the mimicking portfolios (W) and PCs (W^P), I specify the parameters in (56), (57) and (58) with the following priors to ensure invariant model comparisons.

$$P(\bar{\beta}_r, \bar{\Sigma}_r)|F, F^m, F^p, W, W^P = \det\left([F, F^m, F^p]^T[F, F^m, F^p]\right)^{(N-K^N-P)/2}|\bar{\Sigma}_r|^{-(N+K^T+1)/2}, \tag{59}$$

$$\{\beta_j^*, \Sigma_j^*, \beta_j, \Sigma_j\}|W, W^P = |\Sigma_j^*|^{-(K^T+K^N+P+1)/2}\left|\Sigma_j\right|^{-\frac{2K_j-K^T-K^N-P+1}{2}}, \tag{60}$$

$$\alpha_j|\{W, W^P, \beta_j, \Sigma_j\} = MVN(0, k\Sigma_j), \tag{61}$$

where K_j is the total number of included traded, non-traded factors and PCs in M_j.

The mimicking portfolios weights (W) are specified with the traditional Jeffreys (1998) priors, as in (42). To obtain the weights of the PCs, I consider the following specification for the return covariance

$$R - R_{avg} = MVN(0, \Sigma), \tag{62}$$

where R_{avg} denotes the matrix of time-series sample averages of the assets' excess returns. Because Σ commonly appears across the models, I specify it with the conventional Inverse-Wishart improper prior,

$$\Sigma \propto |\Sigma|^{-\frac{N+1}{2}}. \tag{63}$$

This prior automatically governs the priors for the unknown weights of PCs (W^P), as the weights could be expressed as various functions of Σ.

Under the priors in (42), (59), (60), (61) and (63), I obtain the following proposition.

Proposition 4: *The marginal likelihood for M_j comprising the traded factors F_j, non-traded factors*

F_j^n *and PCs* F_j^P *is given by,* $EML_j =$

$$E\left[C_I \times ML_R\left([F_j^*, F_j^{m*}, F_j^{P*}], [F_j, F_j^m, F_j^P]; T\right) \times ML_U\left([F_j, F_j^m, F_j^P], \mathbf{1}; T - (K^T + K^N + P - K_j)\right)\right], \tag{64}$$

where F_j^{m*} (F_j^{P*}), and F_j^m (F_j^P) are the unknown mimicking portfolios of the non-traded factors (PCs) F_j (F_j^P) and F_j^* (F_j^{P*}), respectively. In particular, $F_j^{m*} = [R, F]W_j^*$ and $F_j^m = [R, F]W_j$, $F_j^{P*} = RW_j^{P*}$ and $F_j^m = RW_j^P$. F, F^m, F^P are the total set of traded factors, mimicking portfolios and PCs across all the models, respectively. $ML_R(Y, X; v)$ and $ML_U(Y, X; v)$ denote the restricted and unrestricted MLs of BS and CZZ, expressions of which are given in (24). Expectation is taken with respect to the of the true mimicking portfolio weights W given in (49), and the following density of the covariance matrix of test assets

$$\Sigma \sim IW(S_R, T), \text{ where, } S_R = (R - \bar{R})'(R - \bar{R}), \tag{65}$$

where $IW(S_R, T)$ is the Inverse-Wishart distribution with the scale-matrix S_R, and degrees of freedom T. The invariant constant is given by

$$C_I = det\left([F, F^m, F^P]'[F, F^m, F^P]\right)^{T/2}, \tag{66}$$

where *det* refers to the determinant of a matrix.

Proof. See Appendix Xc. □

12. Marginal Likelihoods and Out-of-Sample Predictions

Before analyzing the empirical data, it is worth highlighting the equivalence between my methodology and ranking models based on their predictive performances. Inferences based on out-of-sample predictive performances of models have recently gained attention because in-sample comparisons favor overfit models. Recent studies, including Kozak et al. (2019) and Lettau and Pelger (2020), among many others, compare competing models based on their estimated out-of-sample R^2s and Sharpe ratios. Such measures often rely on in-sample estimates of various parameters, including alphas, betas, return covariance and means and variances of traded factors that are known to be imprecise measurements of their corresponding true parameters. As a consequence, the point out-of-sample R^2 estimates would be imprecise and may not provide informative inferences.[19] Obtaining exact (asymptotic) distributions that entail parameter estimation uncertainty to conduct valid inferences using these point estimates is a highly complex problem.

An advantage with a Bayesian framework is that it naturally inherits an out-of-sample interpretation. In particular, this paper's procedure provides exact out-of sample predictive distributions of returns under each model, after adjusting for the estimation uncertainties of both the parameters, as well as the mimicking portfolios and principal components. Of course, these predictive distributions entail priors that impose bounds on the maximum Sharpe ratios implied by each model. Importantly, I show that the top-model with the highest marginal likelihood also has the best prediction record in predicting the out-of-sample returns. Thus, model selection based on this paper's marginal likelihoods and out-of-sample prediction records is equivalent.

This fundamental result has not been well-illustrated in prominent Bayesian model comparison studies, including Avramov and Chao (2006) and Barillas and Shanken (2018), among many others. In fact, I have explored this result after the following question raised by several readers of the paper $-$ "Does the Bayesian methodology simply favors a model with a good in-sample but poor out-of-sample performance?" I argue that the methodology rather favors the model with the (relatively) best out-of-sample prediction record. The details follow.

Consider a cross-section of returns $R_1, R_2, \ldots R_T$, over T time-periods. Given a set of models $\{M_j\}$, the predictive density of subsequent returns under M_j is given by

$$P(r_{T+1}|R_1, R_2, \ldots, R_T, M_j) = \int_\Theta P(r_{T+1}|R_1, R_2, \ldots, R_T, \Theta, M_j) P(\Theta|R_1, R_2, \ldots, R_T, M_j), \quad (67)$$

where Θ denotes the parameters under M_j. When R_{T+1} are the true realized returns at period $T+1$,

[19] See Kan et al. (2013) for many examples of modes with large in-sample R^2s that are statistically insignificant.

the model with the highest predictive density evaluated at R_{t+1}, $P(R_{T+1}|R_1, R_2, \ldots, R_T, M_j))$, predicts R_{t+1} better than the other models. Thus, (Bayesian) out-of-sample comparisons involve evaluating predictive densities of models.

Note that the predictive densities are obtained by integrating the parameters over their *posterior* density given the past data, $P(\Theta|R_1, R_2, \ldots, R_T, M_j)$. In contrast, recall that marginal likelihoods are obtained by integrating the parameters over their *prior* density. Interestingly, there is a formal link between predictive densities and the marginal likelihoods. In particular, marginal likelihood of M_j can be decomposed as below

$$ML(R_1, R_2, \ldots, R_T|M_j) = P(R_1, R_2, \ldots, R_T|M_j) = \Pi_{t=2}^{T} P(R_t|R_1, \ldots, R_{t-1}, M_j). \qquad (68)$$

Thus, marginal likelihood can be summarized as *the product of predictive densities*, each of which is obtained by updating the prior based on the past data and predicting subsequent returns, at the end of each time-period. As a consequence, model with the highest marginal likelihood also has the best prediction record, which is an aggregate measure of the predictive performance of the model over the entire period. This result was first recognized by Geisel (1973). It was later documented by Geweke (2005), who called marginal likelihoods as "out-of-sample prediction records". He interprets this result to Milton Friedman's (Friedman (1953)) recommendation on evaluating a model (theory) − "Theory is to be judged by its predictive power .... The only relevant test of the validity of a hypothesis is comparison of predictions with experience".

Thus, model comparisons with marginal likelihoods are naturally related to evaluating models based on their predictive performances.

13. Simulation Evidence

In this section, I use Monte-Carlo simulations to show that this paper's methodology dominates existing methodologies in terms of identifying the true simulated null model with a significant selection-rate and posterior probability. The existing procedures include model comparisons based on their OOS-R^2s, GLS-R^2s, and the MLs of BS and CZZ by treating sample estimated mimicking portfolios as traded factors without adjusting for their estimation uncertainty. To obtain the OOS-R^2 and GLS-R^2 measures, I first estimate underlying parameters of models, including alphas, betas, and residual covariance from the first half of the simulated sample. Then, I compute the out-of-sample measures by comparing the return predictions based on the estimated parameters with the true ex-post realized returns, over the last half of the simulated sample.

Because the estimated parameters are known to be highly imprecise, the point estimates of the OOS-R^2 and GLS-R^2 would be imprecise as well, thus rendering model comparisons with these measures misleading. The ML procedure of BS and CZZ alleviates this concern by shrinking the parameters toward economically motivated priors and thus expected to deliver better inferences. Applying the ML procedure to compare models containing non-traded factors could be misleading, as the sample estimates of mimicking portfolios are known to be imprecise. By shrinking the sample mimicking portfolios toward economically meaningful priors, my methodology, EML, adjusts for their estimation error and outperforms the ML. Consistent with the intuition, the simulation study illustrates that the methodologies' performance can be ranked as below

$$\text{OOS-}R^2 \sim \text{GLS-}R^2 < ML < EML \tag{69}$$

I simulate returns and factors under 8 null models. I calibrate the models' parameters to capture the empirical properties of prominent traded and non-traded factor models. The traded models are BS-6, FF-5, HXZ, and SY. The non-traded models are Petkova (2006), CV, CGP, and CGPT. Test assets are calibrated to match the traditional 25 book-to-market and investment portfolios. With the simulated data, I perform comparisons across the 8 models using each methodology. An ideal methodology should select the true null as the best model more number of times (selection-rate), assign the null with an average rank close to one, and yield large average probabilities of the null (with small standard deviations), across the simulations. I show that the EML dominates the other methodologies across all these metrics. To conserve space, I detail the exact procedure of simulating returns and factors under various versions of the null models in the Internet Appendix-II.

To examine the impact of estimation uncertainty on the model comparison methodologies, I calibrate parameters so that the true Sharpe ratios of the simulated returns and traded factors under the null models equal various multiples of the sample Sharpe ratio of the market portfolio. Higher the Sharpe multiple, lower the volatility of returns, and thus lower would be the (parameter) estimation uncertainty. In the "benchmark scenario", the Sharpe multiple equals 4.5, whereas, in the "high-uncertainty" scenario, it is 3.5. These scenarios are more likely to be realistic. For example, sample Sharpe ratios of prominent models, including FF-5, FF-6, BS-6, HXZ, and SY, are between 3 to 5 times the Sharpe ratio of the market. In the "low-uncertainty" scenario, the multiple equals 6. Due to the market-efficiency forces, the low-uncertainty scenario is highly implausible. As a frame of reference, the sample Sharpe ratio of the entire 25 book-to-market/investment portfolios and set of all traded factors that I use in the study is nearly 6 times the Sharpe ratio of the market. Nevertheless, I examine whether the performances of the existing methodologies improve in this unrealistic scenario.

I conduct 100 iterations of Monte-Carlo simulations under each null model. Each iteration comprises data over 600 time periods, typically the size of standard monthly data sets. Table (13) presents the results when the null models include non-traded factors. Aggregating across the null models in the benchmark scenario, I find that my methodology, EML, on an average, selects the true null as the best model 37.2% times. This selection rate is at least 100% higher than those of the OOS-R^2 and GLS-R^2 and ML procedures. Note that the selection-rate is a binary measure, which only measures whether or not the null emerges as the best model. Thus, inferences from such a measure may not provide adequate information about the performance of the methodologies. For example, consider a simple Bayesian procedure that unconditionally assigns an equal posterior probability to all the models. By construction, this procedure has a selection-rate of 100% because the null (along with the other models) always emerges as the best model.

To draw more informative inferences, I examine the average ranks assigned by the methodologies to each null. For the Bayesian procedures, I additionally report an average and standard deviation of each null model's posterior probabilities over the simulation iterations. On average, I find that the OOS-R^2, GLS-R^2, ML, and EML procedures assign ranks of 5.7, 6.18, 3.98, and

2.41 to the null, respectively. In addition, on average, EML increases the average of each null's posterior probabilities by at least 40%, and decreases the standard deviation of the null probabilities by an impressive 15% over the simulation iterations, compared with those of the ML. Thus, these results confirm that my methodology outperforms the existing methodologies in identifying the mull models with non-traded factors.

The results are qualitatively similar in the high-uncertainty scenario as well. The average selection-rate of the EML procedure is at least 100% higher than those of the existing methodologies. The average and standard deviation of a null model's probabilities are 50% higher and 12% lower than those of the ML, respectively. Consistent with the intuition, the ML and EML methodologies perform significantly better in the low-uncertainty scenario. Nevertheless, the EML procedure dominates the ML by increasing the average null probability by 28% and simultaneously decreasing its standard deviation by 11%. Note that the performance of the procedures that rely on the point estimates of OLS-R^2 and GLS-R^2 do not improve significantly even in this unrealistic, low-uncertainty scenario, thereby highlighting how imprecise such point estimate measures are.

Table (14) presents the results when the null models exclusively comprise traded factors. I find that the ML and EML procedures significantly dominate the OOS-R^2 and GLS-R^2 procedures. Note that the performances of the ML and EML are identical in terms of all the metrics. Recall that the marginal likelihoods based on the ML and EML would be equal for models exclusively comprising traded factors. Of course, the ML and EML procedures yield different marginal likelihoods for models involving non-traded factors. The EML, in a way, uses the economically shrunk mimicking portfolio weights rather than the sample estimated weights. Thus, the relative odds and posterior probabilities of the ML and EML procedures could differ even when the null exclusively comprises traded factors. However, table (14) confirms that shrinking the mimicking portfolio weights do not necessarily increase the non-traded models' probabilities when the null is a traded-factor model.

The average probabilities in tables (13) and (14) could be loosely interpreted as "power" and "size" in frequentist-based hypothesis testing under the null that a traded factor model is true. The probabilities in table (14) suggest that the ML and EML are "well-sized" by consistently assigning lower probabilities to the non-traded models when the null is a traded factor model. However, table (13) suggests that, by increasing the null probability by at least 40%, the EML procedure is more "powerful" than the ML procedure in identifying a non-traded factor model when it is indeed the true model.

Overall, the simulations confirm that the EML dominates the existing methodologies in identifying the true null models.

14. Empirical Comparison of Prominent Asset Pricing Models

This section examines the relative performances of 16 prominent asset pricing models that include traded and non-traded factors and principal components. The traded factor models comprise the traditional CAPM; FF-3, FF-5; the investment-based model, HXZ; the six-factor model, BS-6; and the more recent mispricing model, SY. Among the models with non-traded factors, I consider the liquidity risk model of Pastor and Stambaugh (2003); the ICAPMs of Hahn and Lee (2006), Petkova (2006), the good-beta and bad-beta ICAPM, CV; the more recent ICAPMs, CGP and CGPT; and the intermediary capital factor model of He et al. (2017). Lastly, I consider two models that include the first five and six principal components of the excess returns of various test assets and traded factors, respectively.

I mainly use two sets of test assets for comparing models with monthly data. The first one is a comprehensive set of 52 portfolios containing anomalies (52-anom), which also were examined by Kozak et al. (2019)[20]. These portfolios, obtained by sorting the stocks on various characteristics, possess a strong factor structure. Because asset pricing tests based on such test assets could be misleading (Lewellen et al. (2010)), I also consider a second set of test assets that adds excess returns of 10 industry portfolios to the 52-anom set (52-anom+10-ind). The industry portfolios are available from Kenneth French's website[21]. The monthly data of returns and factors span from

[20] I thank Serhiy Kozak for making the anomaly return data publicly available. Although Kozak et al. (2019) examine 50 anomalies, in the most updated version of Kozak's website, monthly data are available for 52 anomalies during the time-span January, 1974 to December, 2016. So, I use all the available 52 anomalies.

[21] I thank Kenneth French, Yu Yuan, and Lu Zhang for making the factors and portfolio return data publicly available.

January, 1974 to December, 2016. The internet appendix lists the construction of all the factors and details the cross-sectional restriction induced by each asset pricing model.

14a. Models with Traded vs Non-Traded Factors

I begin by comparing the 14 prominent asset pricing models that include traded and non-traded factors but exclude the principal components. Tables (15) and (16) present the posterior model probabilities when the test assets include the 52-anom and the 52-anom+10-ind set, respectively. Between both the test asset sets, SY and BS-6 have the highest posterior model probabilities, which are succeeded by the ICAPMs CGPT and CGP and CV.

CV constitutes first-order innovations to the aggregate excess market returns, term-structure, price-earnings ratio and value spread. CGP adds first-order innovations to the default spread to the CV. CGPT adds first-order innovations to the aggregate stock market variance to CGP and replaces the innovations in the default-spread with the innovations in the risk-free rate. The premise of these ICAPMs is to explain expected returns using a long-term investor's equilibrium portfolio choice in the presence of time-varying investment opportunities. Factors in ICAPMs represent various state variables that reflect time-varying investment opportunities. For example, in the more recent CGPT ICAPM, (innovations to) the stock market variance proxies for the time-varying volatility of the stock returns, which could deteriorate investment opportunities.

Under the prior expectation that the portfolio comprising traded factors and test assets has a maximum Sharpe ratio of 1.5 times the market portfolio's sample Sharpe ratio ($p_m = 1.5 \times Sh_{mkt}$), the posterior SY model probabilities are 30.17% and 34.35% for the 52-anom and 52-anom+10-ind test assets, respectively. The BS-6 model probabilities are 24.7% and 28.85%, CGPT are 14.48% and 10.65%, CGP are 12.75% and 10.65% for the 52-anom and 52-anom+10-ind sets, respectively.

When the test assets are 52-anom, the odds in favor of the more recent CGPT ICAPM is at least thrice more likely than the investment-based HXZ model, and 28 times more likely than the traditional five factor of FF-5. The same holds true for the 52-anom+10-ind set, in which case the probability of CGPT is at least 1.5 and 12 times larger than HXZ and FF-5, respectively. These findings are consistent with Campbell et al. (2018), who document that their ICAPM, which includes the stock return volatility risk factor, explains the cross-section of expected returns better than the FF-5. Their inference is based on relatively lower alphas achieved by each anomaly individually. Nevertheless, this paper suggests that CGPT jointly explains all anomalies significantly better than traditional models with traded factors, after adjusting for the uncertainties related to the macroeconomic factors.

I find that even the former ICAPMs, CGP and CV, outperform FF-5 and HXZ. These findings are robust across other prior multiples of $p_m = 1.25$ and $p_m = 1.75$ that allow for relatively lower and higher prior expectations of mispricing. Given that the factors of HXZ and FF-5 are directly constructed from the sorted portfolios of various anomalies, the superior performance of theoretically motivated ICAPMs makes them even more impressive.

I also report the posterior model probabilities based on the Bayesian procedure of BS and CZZ, which is denoted by ML. In this approach, estimated mimicking portfolios are substituted for the non-traded factors without adjusting for their uncertainty. I find that the ML procedure significantly underestimates the probabilities of the ICAPMs. For example, when the test assets are 52-anom and $p_m = 1.5$, the ML yields posterior model probabilities of 9.1% and 9.48% for the CGP and CGPT models. These are 28% and 35% lower than those implied by this paper's methodology (EML), which adjusts for the estimation uncertainty in the mimicking portfolios. Similarly, the ML overestimates the probabilities of the BS-6 and SY by 17%.

When the test assets are 52-anom+10-ind, the ML underestimates the probabilities of the CGP and CGPT models by 14% and 27%, respectively, and overestimates the BS-6 and SY by 6%, respectively. Moreover, the ML suggests that the HXZ and CGP models are equally likely to summarize the expected returns. In contrast, this paper's EML methodology suggests that the CGP model is 1.23 times more likely to generate the cross-section of expected returns compared with the HXZ model. Thus, adjusting for the estimation uncertainty of mimicking portfolios could substantially affect the model comparison inferences.

The differences in results between the ML and EML procedures are consistent with the simulation results and intuition that the sample-estimated mimicking portfolios often underestimate the explanatory "power" of true non-traded factor models because of large sampling errors. By shrinking the weights of mimicking portfolios toward the economically motivated priors, which impose prior restrictions on the maximum Sharpe ratio attainable by such mimicking portfolios, EML reduces the influence of sampling errors. Thus, the EML methodology captures the explanatory power of true non-traded models better than the existing procedures.

Interestingly, the difference between the ML and EML posterior model probabilities also has an important investment application. For example, consider a risk-averse investor who seeks an optimal portfolio of stocks that maximizes his mean-variance utility. The true means, variances and covariances of the stocks are unknown. Thus, the investor faces uncertainty in determining both the true factor model that might have generated the returns, as well as the underlying parameters (betas and residual covariances) of the models. In addition, the investor is uncertain about the true mimicking portfolios of non-traded factors. In such a scenario, the optimal portfolio would be a function of the EML posterior model probabilities (Pástor and Stambaugh (2000)). Thus, the investor deems the portfolio based on the ML probabilities sub-optimal, which leads to a loss in mean-variance utility. This loss measures the potential economic gains of the investor for using the EML over ML probabilities, which can be easily quantified from the predictive distributions of returns under the EML and ML probabilities, respectively, in the spirit of Kandel and Stambaugh (1996).

Examining this investment application is beyond the scope of this paper, as it requires computation of predictive distributions of returns under each model in addition to the posterior model probabilities. I leave that extension for future research.

Lastly, consistent with the theory developed in the previous sections, the ML and EML procedures yield identical relative-odds among the traded factor models. For example, with the 52-anom test assets and prior multiple equaling 1.5, the ratio of posterior SY and BS-6 model probabilities equals 1.22, based on both the ML and EML procedures. Because the uncertainty adjustment is not required for models with traded factors, their relative odds based on the ML and EML marginal likelihoods would be identical.

14b. Are these Non-Traded Factors Spurious?

Recall from the previous section that the economically motivated ICAPMs of CV, CGP and CGPT outperform the traditionally traded factor models of Fama and French (2015) and Hou et al. (2015). This section examines whether that result is driven by non-traded factors, which are possibly spurious and not correlated with the cross-section of stock returns. Table (17) presents the R^2s and the adjusted R^2s in the regression of non-traded factors on the test assets and traded factors. I find that both measures are significantly high. In particular, the non-traded factors from the CGP and CGPT models, which are innovations in the term spread, default spread, risk-free rate, aggregate price-to-earnings ratio, aggregate stock-return variance and value spreads have R^2s of 22%, 18%, 19%, 36%, 35% and 49%, respectively. The adjusted R^2s are 11%, 6%, 7%, 27%, 25% and 42%, which suggest that the factors are less likely to be spurious.

I also check whether this paper's methodology is susceptible to the spurious factors using Monte-Carlo simulations and find that it is not. In particular, I artificially simulate non-traded factors such that they have the same means and variances as the original non-traded factors but are completely uncorrelated with the returns of test assets and traded factors. Then, I repeat the model comparison exercise substituting the original non-traded factors with the artificially simulated ones. The data for traded factors remain the same.

Tables (18) and (19) present the average model probabilities from 100 such simulations. The test assets are 52-anom and 52-anom+10-ind, respectively. Between the test assets, I find that the models with artificially simulated non-traded factors have low posterior model probabilities, which are dominated by the BS-6, SY and HXZ models. These three models constitute more than 90% posterior probability across all the priors.

Thus, the results confirm that the methodology is not sensitive to spurious factors. As argued previously, the informative priors on the models' mispricing, α and α_{mim}, shrink the posterior mean and Sharpe ratios of mimicking portfolios of spurious non-traded factors toward zero and the sample Sharpe ratio of the market portfolio, respectively. The shrinkage shields against not well-behaved sample Sharpe ratio estimates of spurious non-traded factors' mimicking portfolios.

14c. Traded versus Non-Traded versus Principal Components

This section compares the performance of asset pricing models that include all traded, non-traded factors and principal components. To the 14 prominent asset pricing models studied in the previous section, I add two models that include the first five (PC1-5) and six (PC1-6) principal components of the test assets and traded factors, respectively. PC1-5 was earlier examined by Kozak et al. (2018). Because various models, including BS-6 and CGPT, constitute six factors, I also examine the six factor model, PC1-6, to make a fair assessment.

Table (20) first presents the results for the 52-anom test assets. I find that SY and BS dominate other asset pricing models. These models are succeeded by PC1-6, the macroeconomic CGPT, CGP and CV models and PC1-5. In particular, when the prior multiple p_m equals 1.5, SY and BS-6 have posterior model probabilities of 21.75% and 17.76%, respectively. The probabilities of PC1-6, CGPT, CGP, CV and PC1-5 are 16.16%, 12.79%, 10.32%, 8.12% and 7.78%, respectively.

When the PCs and mimicking portfolios are directly used as traded factors without adjusting for their uncertainty, the ML marginal likelihoods suggest that the previously examined model by Kozak et al. (2018), PC1-5, dominates the economically motivated CV, CGP and CGPT ICAPMs. For example, the ML procedure suggests that odds in favor of PC1-5 are 1.18 times more than the CGPT model. In contrast, the EML marginal likelihoods indicate that the CV, CGP and CGPT ICAPMs dominate PC1-5. In particular, CGPT is 1.64 times more likely than PC1-5 to summarize the cross-section of expected returns. As in the previous section, this result is consistent with the intuition that, the EML, by shrinking the weights toward economically motivated priors, reduces the influence of sampling errors and better captures the explanatory power of true non-traded models.

Note that the EML reports an improvement in the relative-odds in favor of PC1-5 and PC1-6 against the traded factor models, compared with those of the ML. For example, when $p_m = 1.5$, the ML suggests that BS-6 is 1.19 times more likely to summarize the expected anomaly returns than PC1-6, whereas the EML indicates that it is only 1.09. Again, this result is consistent with the previous intuition that, the EML reduces influence of sampling errors of the PCs as well, and thus better captures the explanatory power of true models that comprise PCs. However, unlike the substantial improvements in the relative odds for the CV, CGP, and CGPT ICAPMs, these improvements are modest for PC1-5 and PC1-6. Thus, the sampling errors in the weights of the mimicking portfolios influence model comparisons more than the PCs' weights. This result is expected, given that the non-traded factors are not well-spanned by the test assets, and thus their estimated mimicking portfolios have relatively more significant sampling errors.

To alleviate any misleading inferences that might arise from using the characteristic-sorted 52-anom test assets, which possess a strong factor structure (Lewellen et al. (2010)), I also consider the 52-anom+10-ind test assets. Table (21) presents the results using these test assets. I find that SY and BS-6 significantly dominate the other prominent models. Their posterior probabilities are 34.49% and 28.16%, respectively. Interestingly, I find that the CGPT, CGP, and CV ICAPMs significantly dominate PC1-5 and PC1-6. For example, CGPT, with a posterior probability of 10.58%, is almost thrice more likely than PC1-6 to summarize the expected returns of the 52-anom+10-ind test assets and traded factors.

As a frame of reference, PC1-6 (PC1-5) explains 76.63% (73%) and 75.47% (71.91%) of the total time-series variation in the 52-anom and 52-anom+10-ind test assets, respectively. Thus, the performances of models with the leading PCs do not significantly change between the 52-anom and 52-anom+10-ind test assets, in terms of explaining the time-series variation. However, their performances relative to the other prominent models vary enormously between the test assets when explaining the cross-sectional variation in the expected returns. This result is in the spirit of Lewellen et al. (2010), who noted similar sensitivities in the individual performances of various models when the industry portfolios are added to characteristic-sorted portfolios, albeit in a cross-sectional framework.

Overall, the results from tables (20) and (20) suggest that sparse factor models, including SY, BS-6, and CGPT explain the cross-section of the expected anomaly returns as adequately as models with the first few PCs, if not better. These results partially contrast with Kozak et al. (2019), who compare models based on their OOS-R^2s. They argue that traditional models with sparse characteristic-based factors would not summarize the cross-section of expected anomaly returns as adequately as the first few PCs. However, the previous section's simulation results indicate that inferences based on the point estimates of out-of-sample R^2s would be highly misleading. In this paper, I statistically establish dominance or comparable performances of various sparse-factor models, including SY, BS-6, and CGPT over models with the first few PCs. Note that BS-6 is a sparse characteristics-based factor model. I also document the supremacy of theoretically motivated ICAPMs over the benchmark HXZ and FF-6 models. These results have not been formally established in the literature before.

Of course, the majority of these traditional factor models, including SY and BS-6, consists of factors that are empirically motivated and directly constructed from various sorted portfolios. Thus, the influence of data mining in the construction of these empirically motivated factors still needs to be explored.

In addition, the test assets used in this paper suffer from an ex-post bias. Recall that these test assets represent various portfolio returns that are anomalous to the standard models. Thus, by construction, it is likely that a random linear combination of these test assets outperforms the benchmark models. However, the spurious-factor simulation results in the previous section show that such random combinations yield lower probabilities than the BS-6, HXZ, and SY models, thereby confirming that my results are robust to inherent data-mining in the test assets. Also, recall that the mimicking portfolios and PCs are various linear combinations of the test-asset returns. Other than the theory that relates non-traded factors to mimicking portfolios, there is no economic reason to expect why the combinations of test-asset returns that correspond to the mimicking portfolios should dominate the other linear combinations, including the ones that correspond to the PCs.

As mentioned in the introduction, my results are not directly comparable with Kozak et al. (2019). Whereas this paper considers models with the first few PCs, they use a different SDF (model) involving higher-order PCs.[22] Mathematically, their SDF comprises a few PCs with the highest mean returns, which differ from the conventional leading PCs that instead have the highest

[22] The SDF of Kozak et al. (2019) is motivated by a prior specification that differs from mine. Footnote 12 describes how both the priors differ.

returns' variance. Because the true cross-sectional means of PCs are unknown, comparing models with these sophisticated factors requires an additional uncertainty adjustment. Also, Kozak et al. (2019) use data at a daily level. Recall that I derive marginal likelihoods for models with PCs under the Inverse-Wishart prior specification and the assumption that returns are uncorrelated in the time series. Kozak et al. (2019) made the same assumptions. However, daily returns are known to be non-synchronous, and thus it is unclear whether these assumptions suit the dynamics of daily data (Shanken (1987)). A methodology that accommodates potential temporal dependence of the daily returns and adjusts for the uncertainty of PCs' mean returns could be developed in a future work.

In summary, my results suggest that SY and BS-6 dominate the other prominent asset pricing models in adequately summarizing the cross-section of various anomalies' expected monthly returns. The macroeconomic models such as CGPT, CGP and CV dominate the benchmark PC1-5 model. Although the six-factor PC1-6 performs considerably well on the monthly 52-anom test assets, its performance considerably deteriorates on the expanded 52-anom+10-ind test assets. The macroeconomic models, including CGPT, CGP, and CV dominate the models with the leading PCs on these expanded test assets.

15. Conclusion

Comparing models with non-traded factors and principal components introduces two fundamental challenges. First, regressions of test assets on such factors are not linearly independent, rendering model comparisons with the recent Bayesian procedures sensitive to subsets of the test assets. Second, true mimicking portfolios of non-traded factors and return covariances for principal components are unknown and thereby can only be estimated. I address both of these challenges in a Bayesian framework and develop novel, non-informative priors for invariant comparisons of models containing non-traded factors and PCs.

This is the first Bayesian paper that permits model comparison with traded, non-traded factors and principal components under one framework. My methodology is equivalent to ranking models based on their out-of-sample prediction records, after adjusting for the associated estimation uncertainties. After extensive simulations, I find that the methodology enhances the selection rate of a true non-traded model by more than 100%, compared with the existing model comparison procedures, and increases the posterior probability of a true non-traded model by more than 40% compared with the other Bayesian procedures.

I find that the models of Stambaugh and Yuan (2017) and Barillas and Shanken (2018) dominate the other prominent models in summarizing the cross-section of various anomalies' expected monthly returns. Additionally, the economically motivated ICAPMs of Campbell et al. (2018), Campbell et al. (2013), and Campbell and Vuolteenaho (2004) outperform several benchmark models including, Fama and French (2015) and Hou et al. (2015) and the model with the first five principal components. Given that the factors of these benchmark models are directly constructed from sorted portfolios of various anomalies, the superior performance of ICAPMs makes them even more impressive.

16. Appendix

16a. Bayes Factors for Absolute Tests

Proposition 2:

Proof. Marginal Likelihood under the null hypothesis H_0 is defined as below :

$$ML^N(R, F^N|H_0) = \int P(F^N, R|parameters, H_0)P(parameter|H_0)d(parameters), \tag{70}$$

where $P(parameters|H_0)$ is the prior density of parameters under H_0 and $P(F^N, R|parameters)$ is the density of data under the hypothesis H_0, given the parameters. The above equation further reduces to :

$$\int P(F^N|parameters, H_i)P(R|parameters, H_0)P(parameter|H_0)d(parameters), \tag{71}$$

Note that $P(F^N|parameters, H_0)$ can be directly obtained from equation (6). In particular,

$$P(F^N|R, c, W, H_0) = |2\pi\Sigma_\eta|^{-T/2} \exp\left(-\frac{1}{2}tr\left[(F^N - 1c^T - RW)^T(F^N - 1c^T - RW)\Sigma_\eta^{-1}\right]\right) \tag{72}$$

However, it is not straightforward to compute $P(R|parameters, H_0)$. I first obtain the joint density of the considered $N-K$ test assets $\bar{R}$ and the mimicking portfolios, $P(\bar{R}, RW)$, and then use the transformation matrix $\bar{I}_W$, to obtain the required likelihood, $P(R|parameters, H_0)$. In particular,

$$P(R|parameters, H_0) \propto |\bar{I}_W|^T P(\bar{R}, RW|parameters) \tag{73}$$

This is because, $[\bar{R}, RW] = R\bar{I}_W$. Thus, the likelihood of $P(R|parameters)$ equals

$$\begin{aligned} P(R|parameters, H_0) &= P(\bar{R}, RW|parameters)|det(\frac{\partial vec(R\bar{I_W})}{\partial vec(R)})| \\ &= P(\bar{R}, RW|parameters)|\bar{I}_W|^T \end{aligned}$$

Thus, from equations (9) and (10), I obtain that $P(R|parameters, H_0)$ is proportional to :

$$\begin{aligned} P(R|W, par, H_0) =& |\bar{I}_W|^T ||2\pi\bar{\Sigma}_\epsilon|^{-T/2} exp\left(\frac{-1}{2} tr\left[(\bar{R} - 1\alpha^T - RW\bar{B})^T(\bar{R} - 1\alpha^T - RW_m\bar{B})\bar{\Sigma}_\epsilon^{-1}\right]\right) \\ &\times|2\pi\Sigma_{mim}|^{-T/2} exp\left(\frac{-1}{2}tr\left[(RW - 1\alpha_{mim})^T(RW - 1\alpha_{mim}^T)\Sigma_{mim}^{-1}\right]\right) \end{aligned} \tag{74}$$

Thus, using the likelihood functions from the equations (72) and (74), and the prior specifications from (6), (7) and (8), I derive the marginal likelihoods.

I then use derivations of Barillas and Shanken (2018, 2020) marginal likelihoods to obtain the following results. First, integrating the former expression of (74) with respect to the prior $P(\bar{B}, \bar{\Sigma}_\epsilon)$, I get :

$$\begin{aligned} &\int |2\pi\bar{\Sigma}_\epsilon|^{-T/2} exp\left(\frac{-1}{2} tr\left[(\bar{R} - 1\alpha^T - RW\bar{B})^T(\bar{R} - 1\alpha^T - RW_m\bar{B})\bar{\Sigma}_\epsilon^{-1}\right]\right) \times P(\bar{B}, \bar{\Sigma}_\epsilon) d\bar{B} d\bar{\Sigma}_\epsilon \\ &= |\bar{I}_W|^T \int |2\pi\bar{\Sigma}_\epsilon|^{-T/2} exp\left(\frac{-1}{2} tr\left[(\bar{R} - 1\alpha^T - RW\bar{B})^T(\bar{R} - 1\alpha^T - RW_m\bar{B})\bar{\Sigma}_\epsilon^{-1}\right]\right) \times |\bar{\Sigma}_\epsilon|^{-(N+1)/2} d\bar{B}\bar{\Sigma}_\epsilon \\ &= |\bar{I}_W|^T \left(\frac{1}{2\pi}\right)^{(N-K)(T)/2} 2^{(N-K)(T)/2}\Gamma_{N-K}\left(T/2\right)|S^R_{\bar{R},RW}|^{-(T)/2}|(RW)^T(RW)|^{-(N-K)/2} \\ &= |\bar{I}_W|^T ML_R(\bar{R}, RW; T) \end{aligned}$$

where $|S^R_{\bar{R},RW}|$ is the determinant of the residual sum of square matrix in the restricted (no intercept) regression of $\bar{R}$ on the mimicking portfolios. Similarly, integrating the second expression of (74), with respect to the prior density $P(\alpha_{mim}, \Sigma_{mim})$, I get :

$$\begin{aligned} &\int |2\pi\Sigma_{mim}|^{-T/2} exp\left(\frac{-1}{2}tr\left[(RW - 1\alpha_{mim})^T(RW - 1\alpha_{mim}^T)\Sigma_{mim}^{-1}\right]\right) \times P(\alpha_{mim}, \Sigma_{mim}) d\alpha d\Sigma_{mim} \\ &= \left(\frac{1}{2\pi}\right)^{KT/2} 2^{K(T-(N-K))/2}\Gamma_K\left((T-(N-K))/2\right)|S^U_{RW}|^{-(T-(N-K))/2} Q_{RW} \\ &= ML_U(RW, 1; T-(N-K)), \end{aligned}$$

where $|S^U_{RW}|$ is the determinant of the residual sum of squares matrix in the regression of mimicking portfolios on the unit vector, Q_{RW} is analogous to the scalar derived by Barillas and Shanken (2018).

$$Q_{RW} = (1 + \frac{1}{1+k} V/T)^{-(T-(N-K))/2} (1+k)^{-N/2}, \tag{75}$$

where $V = T\hat{\alpha}_{mim}^T \hat{\Sigma}_{mim}^{-1} \hat{\alpha}_{mim}$, the expressions under hat, $\hat{}$ are the OLS estimates of corresponding parameters in the regression of mimicking portfolios on the unit vector. Thus, the marginal likelihood of data under the null reduces to

$$ML^N(R, F^N|H_0) = \int P(F^N, R|c, W, \Sigma_\eta, H_0) ML_R(\bar{R}, RW) ML_U(RW) P(c, W, \Sigma_\eta|H_0) d(c, W, \Sigma_\eta) \tag{76}$$

Note that $P(F^N, R|c, W, \Sigma_\eta, H_0)P(c, W, \Sigma_\eta|H_0)d(c, W, \Sigma_\eta)$ is proportional to the posterior density $P(c, W, \Sigma_\eta|F^N, R)$, where

$$\begin{aligned} P(c, W, \Sigma_\eta|F^N, R) &= P(c, W|\Sigma_\eta, F^N, R) \times P(\Sigma_\eta|F^N, R), \\ P(vec([c, W])|\Sigma_\eta, F^N, R) &= MVN\left(vec(\hat{c}, \hat{W}), \Sigma_\eta\right) \\ P(\Sigma_\eta|F^N, R) &= IW\left(S_{F^N}, T\right), \end{aligned} \tag{77}$$

where vec denotes the vectorized form of a matrix (stacking up all columns under one column), $\hat{c}, \hat{W}$ are the OLS estimates of the regression coefficients c, W, and S_{F^N} is the $K \times K$ cross-product of OLS residuals, in the regression of non-traded factors F^N on the excess returns R. MVN and IW are notations for the Multivariate normal and Inverse-Wishart densities respectively. Thus,

$$ML^N(R, F^N|H_0) = E\left[|\bar{I}_W|^T ML_R(\bar{R}, RW; T) ML_U(RW, 1; T-(N-K))\right], \tag{78}$$

where the expectation is taken with respect to the posterior density given in equation (77).

Similarly, marginal likelihood under the alternative hypothesis reduces to

$$ML^N(R, F^N|H_1) = E\left[|\bar{I}_W|^T ML_U(\bar{R}, RW; T) ML_U(RW, 1; T-(N-K))\right], \tag{79}$$

where

$$
\begin{aligned}
ML_U(\bar{R}, RW; T) = \left(\frac{1}{2\pi}\right)^{(N-K)(T)/2} & 2^{(N-K)(T)/2}\Gamma_{N-K}\left((T)/2\right) \\
& |S^U_{\bar{R},RW}|^{-(T)/2}|(RW)^T(RW)|^{-(N-K)/2}Q_{\bar{R},RW},
\end{aligned}
$$

$|S^U_{\bar{R},RW}|$ is the determinant of the residual sum of square matrix in the unrestricted (intercept included) regression of $\bar{R}$ on the mimicking portfolios, $Q_{\bar{R},RW}$ is analogous to the scalar derived by Barillas and Shanken (2018).

$$
Q_{\bar{R},RW} = (1 + \frac{a}{a+k}\bar{V}/T)^{-(T)/2}(1 + \frac{k}{a})^{-(N-K)/2}, \tag{80}
$$

where $a = (1 + Sh(RW)^2)/T$, $Sh(RW)^2$ is the squared Sharpe-Ratio of the mimicking portfolio,

$$
\bar{V} = T\frac{\hat{\bar{\alpha}}^T\bar{\Sigma}_\epsilon^{-1}\hat{\bar{\alpha}}}{1 + Sh(RW)^2} \tag{81}
$$

□

16b. *Invariance to the Choice of Test Assets and Scaling of Mimicking Portfolio Weights*

In this section, I show that the above marginal likelihood is invariant to the choice of subset of test assets, $\bar{R}$, and also to the scale of mimicking portfolio weights. To show that $ML^N(R, F^N|H_0)$ is invariant to $\bar{R}$, it is enough to show that $ML_R(\bar{R}, RW)$ is invariant to $\bar{R}$. This is because, all the other terms do not depend on $\bar{R}$. Further, collecting the terms of $ML_R(\bar{R}, RW)$, it is enough to show that $|\bar{I}_W|^T|S^R_{\bar{R},RW}|^{-(T)/2}$ is invariant to $\bar{R}$. This can be easily showed using the determinant lemma for the partitioned matrices. In particular, note that

$$
\begin{aligned}
& |\left[\bar{R}\ RW\right]^T\left[\bar{R}\ RW\right]| = |S^R_{\bar{R},RW}| \times |(RW)^T(RW)| \\
\implies & |R^TR||\bar{I}_W|^2 = |S^R_{\bar{R},RW}| \times |(RW)^T(RW)| \\
\implies & |S^R_{\bar{R},RW}|^{-(T)/2}|\bar{I}_W|^{(T)} = |(RW)^T(RW)|^{(T)/2}(R^TR)^{-(T)/2}
\end{aligned} \tag{82}
$$

Thus, $|\bar{I}_W|^T|S^R_{\bar{R},RW}|^{-(T)/2}$ is invariant to the subset of test assets $\bar{R}$, where it depends only on the mimicking portfolios (RW). Consequently, $ML^N(R, F^N|H_0)$ is invariant to $\bar{R}$.

Similarly, it follows that $ML_U(\bar{R}, RW)$ is also invariant to the choice of test assets $\bar{R}$. Also, the GRS-test statistic in equation (81) is also invariant to the choice of $\bar{R}$. Thus, $ML^N(R, F^N|H_1)$ is also invariant to the choice of $\bar{R}$.

To show that $ML^N(F^N, R|H_0)$ is independent to the scale of the mimicking portfolio weights, note that using equation (82), I can express it as

$$
\begin{aligned}
&ML^N(F^N, R|H_0) = E\left(ML_R(\bar{R}, RW)ML_U(RW)|H_0\right)\\
&= \text{constant} \times E\left(|(RW)^T(RW)|^{(T)/2}(R^TR)^{-(T)/2}|(RW)^T(RW)|^{-(N-K)/2}|S^U_{RW}|^{-(T-(N-K))/2}Q_{RW}|H_0\right)\\
&= \text{constant}_1 E\left(|(RW)^T(RW)|^{(T)/2}|(RW)^T(RW)|^{-(N-K)/2}|(RW\ 1)^T(RW\ 1)|^{-(T-(N-K))/2}Q_{RW}|H_0\right)
\end{aligned}
$$

Note that Q_{RW} is invariant to the scale of W. Moreover, all the exponents sum to zero in the above expression. Thus, $ML^N(F^N, R|H_0)$ is invariant to the scale of W. Similarly, it can be shown that $ML^N(F^N, R|H_1)$ is also invariant to the scale of W.

16c. Bayes Factors for Model Comparisons

Proposition 3:

Proof. Marginal Likelihood of model M_j equals

$$
\begin{aligned}
ML_j^N(F^N, R, F) &= \int_{par} P(F^N, F, R|par, M_j)P(par|M_j)\\
&= \int_{par} P(F^N|F, R, par, M_j)P(F, R|par, M_j)P(par|M_j) \qquad (83)
\end{aligned}
$$

Note that the conditional density of non-traded factor given the traded factors and the mimicking portfolios can be obtained from (35). Similar to the previous section obtaining $P(R, F|par)$ is not straight forward. However, the conditional densities $P(\bar{R}|F, F^m, par)$, $P(F_j^*, F_j^{m*}|F_j, F_j^m, par)$ and

$P(F_j, F_j^m|par)$ can be obtained from the equations (38) and (40). Moreover,

$$P(R, F|par) \propto |\bar{I}_W|^T P(\bar{R}, F_j^*, F_j^{m*}, F_j, F_j^m|par), \tag{84}$$

where $[\bar{R}, F_j^*, F_j^{m*}, F_j, F_j^m] = [\bar{R}, F, F^m] = R\bar{I}_W$. Thus, I have

$$\begin{aligned} ML_j^N(F^N, R, F) = \int_{par} & |\bar{I}_W|^T P(F^N|F, R, par, M_j) P(\bar{R}|F, F^m, par) \\ & P(F_j^*, F_j^{m*}|F_j, F_j^m, par) P(F_j, F_j^m|par) P(par|M_j) \end{aligned} \tag{85}$$

Under the prior specifications in (42), (44), (45), (46), (47), I have,

$$P(c, W, \Sigma_\eta) \propto |\Sigma_\eta|^{-(K^N+1)/2} \tag{86}$$

$$P(\bar{\beta}_r, \bar{\Sigma}_r)|F, F^m = \det\left([F, F^m]^T [F, F^m]\right)^{(N-K^N)/2} |\bar{\Sigma}_r|^{-(N+K^T+1)/2} \tag{87}$$

$$\alpha_j|\{W, \beta_j, \Sigma_j\} \propto MVN(0, k\Sigma_j) \tag{88}$$

$$\{\beta_j^*, \Sigma_j^*, \beta_j, \Sigma_j\}|W = |\Sigma_j^*|^{-(K^T+K^N+1)/2} \left|\Sigma_j\right|^{-\frac{2K_j - K^T - K^N + 1}{2}}, \tag{89}$$

First, conditional on W, integrating out the parameters $\bar{\beta}_r, \bar{\Sigma}_r$, I get the below term

$$T1 = \left|S^R_{\bar{R},[F,F^m]}\right|^{(-T/2)}, \tag{90}$$

where $S^R_{\bar{R},[F,F^m]}$ is the restricted residual sum of squares in the regression of test assets $\bar{R}$ on the mimicking portfolios and traded factors, $[F, F^m]$.

Next conditional on W, I integrate out the $\{\beta_j^*, \Sigma_j^*\}$, $\{\alpha_j, \beta_j, \Sigma_j\}$ parameters sequentially, which yields

$$\begin{aligned} T2 &= ML_R\left([F_j^*, F_j^{m*}], [F_j, F_j^m]; T\right) \\ T3 &= ML_U\left([F_j, F_j^m], \mathbf{1}; T - (K^T + K^N - K_j)\right) \end{aligned} \tag{91}$$

Thus, the marginal likelihood of model M_j can be expressed as

$$ML_j^N(F^N, R, F) = \int_{par} |\bar{I}_W|^T P(F^N|F, R, c, W, \Sigma_\eta, M_j)$$
$$T_1 \times T_2 \times T_3 \times P(c, W, \Sigma_\eta|M_j) \tag{92}$$

Moreover, notice that

$$\left|[[R, F]\bar{I}_W]^T\ [[R, F]\bar{I}_W]\right| = \left|[R, F]^T[R, F]\right|\ |\bar{I}_W||^2 \tag{93}$$

Moreover, by the determinant lemma for the partitioned matrices, also notice that,

$$\left|[[R, F]\bar{I}_W]^T\ [[R, F]\bar{I}_W]\right| = \left|[F, F^m]^T[F, F^m]\right| \left|S^R_{R,[F,F^m]}\right| \tag{94}$$

Therefore, from (93) and (94), I have

$$T_1 \times |\bar{I}_W|^T = \left|[F, F^m]^T[F, F^m]\right|^{T/2} \tag{95}$$

Thus, the marginal likelihood reduces to :

$$\begin{aligned} ML_j^N(F^N, R, F) &= \int_{par} \left|[F, F^m]^T[F, F^m]\right|^{T/2} T_2 T_3 P(F^N|F, R, c, W, \Sigma_\eta, M_j) P(c, W, \Sigma_\eta|M_j) \\ &= E\left[\left|[F, F^m]^T[F, F^m]\right|^{T/2} ML_R\left([F_j^*, F_j^{m*}], [F_j, F_j^m]; T\right) ML_U\left([F_j, F_j^m], \mathbf{1}; T - (K^T + K^N - K_j)\right)\right], \end{aligned} \tag{96}$$

where the expectation is taken with respect to the posterior density (29).

□

Corollary 2 : The derived marginal likelihoods are invariant to the scale of the factors.

Proof. This is straightforward by noting that the exponents of each factor, whether included or excluded from the model, sums to zero. For example, excluded factors has an exponent of $T/2$ from the first term of (96), and $-T/2$ from the second term of (96). Thus, the sum of exponents is zero. Similarly, included factors has an exponent of $T/2$ from the first term of, $\left(-(K^T + K^N - K_j)\right)/2$ from the second term, whereas $-\left(T - (K^T + K^N - K_j)\right)/2$ from the last term of (96). Thus, the exponents again sum to zero. □

16d. Model Comparisons with Principal Components

Proposition 4 :

Proof. The proof is similar to the proof of proposition 3. □

References

Adrian, Tobias, Erkko Etula, and Tyler Muir, 2014, Financial Intermediaries and the Cross-Section of Asset Returns, *The Journal of Finance* 69, 2557–2596.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The Cross-Section of Volatility and Expected Returns, *The Journal of Finance* 61, 259–299.

Avramov, Doron, and John C. Chao, 2006, An Exact Bayes Test of Asset Pricing Models with Application to International Markets, *The Journal of Business* 79, 293–324.

Barillas, Francisco, Raymond Kan, Cesare Robotti, and Jay Shanken, 2019, Model Comparison with Sharpe Ratios, *Journal of Financial and Quantitative Analysis* 1–35.

Barillas, Francisco, and Jay Shanken, 2017, Which Alpha?, *The Review of Financial Studies* 30, 1316–1338.

Barillas, Francisco, and Jay Shanken, 2018, Comparing Asset Pricing Models, *The Journal of Finance* 73, 715–754.

Barillas, Francisco, and Jay Shanken, 2020, Comparing Priors for Comparing Asset Pricing Models, *Forthcoming, Comments and Rejoinder, The Journal of Finance* .

Breeden, Douglas T., 1979, An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* 7, 265–296.

Britten-Jones, Mark, 1999, The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights, *The Journal of Finance* 54, 655–671.

Bryzgalova, Svetlana, Jintao Huang, and Christian Julliard, 2020, Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models, *Working Paper, SSRN* .

Campbell, John Y., Stefano Giglio, and Christopher Polk, 2013, Hard Times, *The Review of Asset Pricing Studies* 3, 95–132, Publisher: Oxford Academic.

Campbell, John Y., Stefano Giglio, Christopher Polk, and Robert Turley, 2018, An intertemporal CAPM with stochastic volatility, *Journal of Financial Economics* 128, 207–233.

Campbell, John Y., and Tuomo Vuolteenaho, 2004, Bad Beta, Good Beta, *American Economic Review* 94, 1249–1275.

Carhart, Mark M., 1997, On Persistence in Mutual Fund Performance, *The Journal of Finance* 52, 57–82.

Chib, Siddhartha, Dashan Huang, Lingxiao Zhao, and Guofu Zhou, 2020a, Winners from Winners: A Tale of Risk Factors, SSRN Scholarly Paper ID 3478223, Social Science Research Network, Rochester, NY.

Chib, Siddhartha, Xiaming Zeng, and Lingxiao Zhao, 2020b, On Comparing Asset Pricing Models, *The Journal of Finance* 75, 551–577.

Cochrane, John H., and Jesus Saa-Requejo, 2000, Beyond Arbitrage: Good-Deal Asset Price Bounds in Incomplete Markets, *Journal of Political Economy* 108, 79–119.

Fama, Eugene F., and Kenneth R. French, 1992, The Cross-Section of Expected Stock Returns, *The Journal of Finance* 47, 427–465.

Fama, Eugene F., and Kenneth R. French, 2010, Luck versus Skill in the Cross-Section of Mutual Fund Returns, *The Journal of Finance* 65, 1915–1947.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F., and Kenneth R. French, 2016, Dissecting Anomalies with a Five-Factor Model, *The Review of Financial Studies* 29, 69–103.

Fama, Eugene F., and Kenneth R. French, 2018, Choosing factors, *Journal of Financial Economics* 128, 234–252.

Friedman, Milton, 1953, *Essays in Positive Economics* (University of Chicago Press).

Geisel, Martin S., 1973, Bayesian Comparisons of Simple Macroeconomic Models, *Journal of Money, Credit and Banking* 5, 751–772.

Geweke, John, 1988, Antithetic acceleration of Monte Carlo integration in Bayesian inference, *Journal of Econometrics* 38, 73–89.

Geweke, John, 1989, Bayesian Inference in Econometric Models Using Monte Carlo Integration, *Econometrica* 57, 1317–1339.

Geweke, John, 2005, Contemporary Bayesian Econometrics and Statistics.

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken, 1989, A Test of the Efficiency of a Given Portfolio, *Econometrica* 57, 1121–1152.

Hahn, Jaehoon, and Hangyong Lee, 2006, Yield Spreads as Alternative Risk Factors for Size and Book-to-Market, *The Journal of Financial and Quantitative Analysis* 41, 245–269.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, ... and the Cross-Section of Expected Returns, *The Review of Financial Studies* 29, 5–68.

Harvey, Campbell R, and Guofu Zhou, 1990, Bayesian inference in asset pricing tests, *Journal of Financial Economics* 26, 221–254.

He, Zhiguo, Bryan Kelly, and Asaf Manela, 2017, Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1–35.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting Anomalies: An Investment Approach, *The Review of Financial Studies* 28, 650–705.

Huberman, Gur, Shmuel Kandel, and Robert F. Stambaugh, 1987, Mimicking Portfolios and Exact Arbitrage Pricing, *The Journal of Finance* 42, 1–9.

Jeffreys, Harold, 1998, *The Theory of Probability* (OUP Oxford).

Jensen, Michael C., 1968, The Performance of Mutual Funds in the Period 1945-1964, *The Journal of Finance* 23, 389–416.

Jensen, Michael C., Fischer Black, and Myron S. Scholes, 1972, The Capital Asset Pricing Model: Some Empirical Tests, SSRN Scholarly Paper ID 908569, Social Science Research Network, Rochester, NY.

Kan, Raymond, Cesare Robotti, and Jay Shanken, 2013, Pricing Model Performance and the Two-Pass Cross-Sectional Regression Methodology, *The Journal of Finance* 68, 2617–2649.

Kan, Raymond, Xiaolu Wang, and Xinghua Zheng, 2019, In-Sample and Out-of-Sample Sharpe Ratios of Multi-Factor Asset Pricing Models, SSRN Scholarly Paper ID 3454628, Social Science Research Network, Rochester, NY.

Kan, Raymond, and Chu Zhang, 1999, Two-Pass Tests of Asset Pricing Models with Useless Factors, *The Journal of Finance* 54, 203–235, Publisher: [American Finance Association, Wiley].

Kan, Raymond, and Guofu Zhou, 2007, Optimal Portfolio Choice with Parameter Uncertainty, *Journal of Financial and Quantitative Analysis* 42, 621–656.

Kandel, Shmuel, and Robert F. Stambaugh, 1996, On the Predictability of Stock Returns: An Asset-Allocation Perspective, *The Journal of Finance* 51, 385–424.

Koijen, Ralph S. J., Hanno Lustig, and Stijn Van Nieuwerburgh, 2017, The cross-section and time series of stock and bond returns, *Journal of Monetary Economics* 88, 50–69.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting Factor Models, *The Journal of Finance* 73, 1183–1223.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2019, Shrinking the cross-section, *Journal of Financial Economics* .

Lettau, Martin, and Markus Pelger, 2020, Factors That Fit the Time Series and Cross-Section of Stock Returns, *The Review of Financial Studies* 33, 2274–2325.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics* 96, 175–194.

MacKinlay, A. Craig, 1995, Multifactor models do not explain deviations from the CAPM, *Journal of Financial Economics* 38, 3–28.

Maio, Paulo F., 2019, Comparing asset pricing models with traded and macro risk factors, SSRN Scholarly Paper ID 2535572, Social Science Research Network, Rochester, NY.

Pastor, Lobos, and Robert F. Stambaugh, 2003, Liquidity Risk and Expected Stock Returns, *Journal of Political Economy* 111, 642–685.

Petkova, Ralitsa, 2006, Do the Fama-French Factors Proxy for Innovations in Predictive Variables?, *The Journal of Finance* 61, 581–612.

Pástor, Ľuboš, and Robert F. Stambaugh, 2000, Comparing asset pricing models: an investment perspective, *Journal of Financial Economics* 56, 335–381.

Scheaffer, Richard L., William Mendenhall III, R. Lyman Ott, and Kenneth G. Gerow, 2011, *Elementary Survey Sampling* (Cengage Learning).

Shanken, Jay, 1987, Nonsynchronous Data and the Covariance-Factor Structure of Returns, *The Journal of Finance* 42, 221–231.

Stambaugh, Robert F., and Yu Yuan, 2017, Mispricing Factors, *The Review of Financial Studies* 30, 1270–1315.

Vassalou, Maria, 2003, News related to future GDP growth as a risk factor in equity returns, *Journal of Financial Economics* 68, 47–73.

Table 13

Performance of the Proposed Methodology: Simulation Evidence

This table compares the performance of this paper's methodology with the existing procedures in identifying the true simulated-asset pricing model. The data comprise 100 independent simulations of 600 monthly returns and factors. The column "Simulated Null Model" presents the null model under which data are simulated. The columns under "Selection-Rate" present the number of times each methodology picks the simulated null model as the best model, out of 100 simulations. The columns under "Average Null Rank" present the average rank of the null model across 100 simulations. The columns under "Avg Null Prob", "Std Null Prob" present the average and standard deviation of posterior model probabilities of the null across 100 simulations, respectively. The OOS-R^2 and GLS-R^2 columns present model comparisons based on the out-of sample R^2 and Sharpe ratio procedures, respectively. The out-of-sample measures are estimated using the last 300 observations from the parameters, which are estimated using the first 300 observations. The "ML" columns report model comparisons based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors without adjusting for their uncertainty. "EML" presents model comparisons based on this paper's methodology, which adjusts for the estimation error in the mimicking portfolios. The prior multiple (p_m) equals 1.5 for both the "ML" and "EML" procedures. This implies $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

	Selection-Rate of Null				Avg Null Rank				Avg Null Prob		Std Null Prob	
Null model	OOS-R^2	GLS-R^2	ML	EML	OOS-R^2	GLS-R^2	ML	EML	ML	EML	ML	EML
Panel A: Benchmark Scenario - True Sharpe Ratio under Null= 4.5 × Sharpe ratio of Market												
Campbell et al. (2013)	3	9	5	7	6.15	6.38	6.12	3.29	5.81%	10.36%	13.10%	11.77%
Petkova (2006)	11	13	22	70	5.87	5.91	2.21	1.58	44.43%	54.53%	36.85%	31.80%
Campbell and Vuolteenaho (2004)	24	17	4	4	5.47	5.5	4.98	3.37	8.68%	11.54%	14.62%	12.53%
Campbell et al. (2018)	19	11	36	68	5.32	6.92	2.62	1.41	31.59%	52.32%	29.57%	26.34%
Average	14.25	12.5	16.75	37.25	5.70	6.18	3.98	2.41	22.63%	32.19%	23.53%	20.61%
Panel B: High-Uncertainty Scenario - True Sharpe Ratio under Null= 3.5 × Sharpe ratio of Market												
Campbell et al. (2013)	10	4	2	3	6.11	6.17	6.54	3.95	5.25%	9.56%	9.16%	10.73%
Petkova (2006)	11	10	14	24	6.22	6.09	5.21	3.62	14.76%	17.31%	20.69%	17.54%
Campbell and Vuolteenaho (2004)	8	8	3	0	6.55	5.78	5.91	4.84	6.68%	8.62%	7.84%	6.41%
Campbell et al. (2018)	20	7	23	58	6.16	6.69	3.64	1.74	20.78%	32.18%	20.81%	17.52%
Average	12.25	7.25	10.5	21.25	6.26	6.18	5.33	3.54	11.87%	16.92%	14.62%	13.05%
Panel C: Low-Uncertainty Scenario - True Sharpe Ratio under Null= 6 × Sharpe ratio of Market												
Campbell et al. (2013)	9	13	6	18	6.1	6.02	3.78	2.32	7.92%	20.44%	18.31%	24.96%
Petkova (2006)	12	20	76	84	5.71	5.17	1.6	1.32	74.09%	81.45%	37.66%	29.44%
Campbell and Vuolteenaho (2004)	11	14	6	0	5.56	5.52	5.07	3.95	4.55%	5.43%	10.59%	7.60%
Campbell et al. (2018)	12	7	41	62	6.12	6.26	2.59	1.6	37.30%	51.89%	37.12%	31.11%
Average	11	13.5	32.25	41	5.87	5.74	3.26	2.30	30.97%	39.80%	25.92%	23.27%

Table 14

Performance of the Proposed Methodology: Simulation Evidence

This table compares the performance of this paper's methodology with the existing procedures in identifying the true simulated-asset pricing model. The data comprise 100 independent simulations of 600 monthly returns and factors. The column "Simulated Null Model" presents the null model under which data are simulated. The columns under "Selection-Rate" present the number of times each methodology picks the simulated null model as the best model, out of 100 simulations. The columns under "Average Null Rank" present the average rank of the null model across 100 simulations. The columns under "Avg Null Prob", "Std Null Prob" present the average and standard deviation of posterior model probabilities of the null across 100 simulations, respectively. The OOS-R^2 and GLS-R^2 columns present model comparisons based on the out-of sample R^2 and Sharpe ratio procedures, respectively. The out-of-sample measures are estimated using the last 300 observations from the parameters, which are estimated using the first 300 observations. The "ML" columns report model comparisons based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors without adjusting for their uncertainty. "EML" presents model comparisons based on this paper's methodology, which adjusts for the estimation error in the mimicking portfolios. The prior multiple (p_m) equals 1.5 for both the "ML" and "EML" procedures. This implies $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

	Selection Rate of Null				Avg Rank of Null				Avg Null Prob		Std Null Prob	
Null model	OOS-R^2	GLS-R^2	ML	EML	OOS-R^2	GLS-R^2	ML	EML	ML	EML	ML	EML
	Panel A: Benchmark Scenario - True Sharpe Ratio under null= 4.5 × Sharpe ratio of Market											
Fama and French (2015)	27	26	100	100	4.76	5.74	1	1	99.99%	99.99%	0.06%	0.06%
Barillas and Shanken (2018)	36	3	100	100	4.65	9.2	1	1	95.78%	95.78%	4.37%	4.37%
Hou et al. (2015)	26	12	72	72	6	8.31	1.28	1.28	54.80%	54.80%	10.78%	10.78%
Stambaugh and Yuan (2017)	32	4	100	100	4.52	7.93	1	1	99.42%	99.42%	1.65%	1.65%
Average	30.25	11.25	93.00	93.00	4.98	7.80	1.07	1.07	87.50%	87.50%	4.22%	4.22%
	Panel B: High-Uncertainty Scenario - True Sharpe Ratio under Null= 3.5 × Sharpe ratio of Market											
Fama and French (2015)	15	4	98	98	6.56	9.59	1.03	1.03	81.21%	81.13%	15.72%	15.89%
Barillas and Shanken (2018)	0	0	95	96	10.01	9.84	1.05	1.04	63.96%	71.48%	14.88%	15.41%
Hou et al. (2015)	23	1	73	73	7.43	9.95	1.28	1.28	48.34%	48.24%	10.14%	10.15%
Stambaugh and Yuan (2017)	25	5	99	99	4.9	9.04	1.02	1.02	92.72%	92.72%	12.25%	12.26%
Average	15.75	2.50	91.25	91.50	7.23	9.61	1.10	1.0925	71.56%	73.39%	13.25%	13.43%
	Panel C: Low-Uncertainty Scenario - True Sharpe Ratio under Null= 6 × Sharpe ratio of Market											
Fama and French (2015)	27	26	100	100	4.76	5.74	1	1	99.99%	99.99%	0.06%	0.06%
Barillas and Shanken (2018)	32	9	100	100	4.72	7.7	1	1	98.38%	98.38%	4.07%	4.07%
Hou et al. (2015)	23	6	82	82	4.78	8.71	1.18	1.18	61.78%	61.78%	13.96%	13.96%
Stambaugh and Yuan (2017)	24	2	100	100	5.38	9.14	1	1	94.79.%	94.77%	8.59%	8.58%
Average	26.5	10.75	95.50	95.50	4.91	7.82	1.05	1.05	88.73%	88.73%	6.67%	6.67%

Table 15

Traded vs Non-Traded: Model Probabilities with 52 Test Assets of Kozak et al. (2019)

This table compares the posterior model probabilities of 14 prominent asset pricing models that include both traded and non-traded factors. The prior multiple, p_m, denotes the prior expectation of the relative increase in the Sharpe ratio when the factors are added to the market portfolio. The "ML" columns report the posterior model probabilities based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors. These probabilities do not account for the estimation error in the mimicking portfolios. The "EML" columns report the posterior model probabilities using this paper's methodology, which adjusts for the estimation uncertainty in mimicking portfolios. Test assets are monthly returns of 52 characteristic-sorted anomaly portfolios, which are used in Kozak et al. (2019). $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

Test Assets : 52-anom	Prior (p_m) =1.25		Prior (p_m) =1.5		Prior (p_m) =1.75	
Model	ML	EML	ML	EML	ML	EML
CAPM	0.03%	0.05%	0.01%	0.01%	0.00%	0.00%
Fama and French (1992)	0.09%	0.12%	0.02%	0.02%	0.00%	0.00%
Barillas and Shanken (2018)	26.13%	23.49%	28.83%	24.70%	31.60%	21.07%
Fama and French (2015)	1.05%	1.23%	0.52%	0.45%	0.19%	0.12%
FF-6	1.83%	2.06%	1.04%	0.89%	0.45%	0.30%
Hou et al. (2015)	7.00%	7.02%	5.57%	4.78%	3.87%	2.58%
Stambaugh and Yuan (2017)	30.58%	27.10%	35.21%	30.17%	41.12%	27.42%
Pastor and Stambaugh (2003)	0.59%	0.83%	0.26%	0.31%	0.08%	0.14%
He et al. (2017)	0.03%	0.09%	0.01%	0.01%	0.00%	0.00%
Petkova (2006)	0.69%	1.42%	0.31%	0.86%	0.09%	0.52%
Campbell and Vuolteenaho (2004)	10.84%	10.67%	9.63%	10.58%	7.80%	12.16%
Campbell et al. (2013)	10.38%	12.13%	9.10%	12.75%	7.22%	15.81%
Campbell et al. (2018)	10.74%	13.74%	9.48%	14.48%	7.58%	19.87%
He et al. (2017)	0.03%	0.04%	0.01%	0.00%	0.00%	0.00%

Table 16
Traded vs Non-Traded: Model Probabilities with 52 Anomalies + 10 Industry portfolios
This table compares the posterior model probabilities of 14 prominent asset pricing models that include both traded and non-traded factors. The prior multiple, p_m, denotes the prior expectation of the relative increase in the Sharpe ratio when the factors are added to the market portfolio. The "ML" columns report the posterior model probabilities based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors. These probabilities do not account for the estimation error in the mimicking portfolios. The "EML" columns report the posterior model probabilities using this paper's methodology, which adjusts for the estimation uncertainty in mimicking portfolios. Test assets are monthly returns of 52 characteristic-sorted anomaly portfolios, which are used in Kozak et al. (2019), plus the excess returns of 10 industry portfolios. $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

Test assets : 52-anom+10-ind	prior (p_m) =1.25		prior (p_m) =1.5		prior (p_m) =1.75	
Model	ML	EML	ML	EML	ML	EML
CAPM	0.12%	0.11%	0.02%	0.02%	0.00%	0.00%
Fama and French (1992)	0.26%	0.26%	0.06%	0.05%	0.01%	0.01%
Barillas and Shanken (2018)	25.98%	25.56%	30.61%	28.85%	34.02%	28.69%
Fama and French (2015)	1.96%	1.93%	0.89%	0.83%	0.35%	0.30%
FF-6	3.08%	3.03%	1.64%	1.54%	0.78%	0.65%
Hou et al. (2015)	9.01%	8.87%	7.17%	6.76%	5.25%	4.43%
Stambaugh and Yuan (2017)	29.42%	28.94%	36.44%	34.35%	42.84%	36.13%
Pastor and Stambaugh (2003)	1.72%	1.56%	0.74%	0.98%	0.28%	0.41%
Hahn and Lee (2006)	0.11%	0.17%	0.02%	0.03%	0.00%	0.01%
Petkova (2006)	0.49%	0.76%	0.13%	0.31%	0.03%	0.13%
Campbell and Vuolteenaho (2004)	9.14%	8.45%	7.32%	7.23%	5.39%	7.32%
Campbell et al. (2013)	9.06%	9.49%	7.21%	8.38%	5.28%	9.45%
Campbell et al. (2018)	9.56%	10.80%	7.75%	10.65%	5.77%	12.47%
He et al. (2017)	0.09%	0.09%	0.01%	0.01%	0.00%	0.00%

Table 17

Correlations of Macroeconomic Factors with the Cross-Section of Stock Returns

This table presents the correlations of non-traded factors with the cross section of stock returns. The first column presents the set of all non-traded factors from the 14 asset pricing models considered in the paper. The second and third columns present the R^2 and adjusted R^2 in the time series regression of each non-traded factor on the test assets. The last six columns present the correlations of each non-traded factor's mimicking portfolio with the five factors of Fama and French (2015) and the momentum factor UMD, respectively. Panel A presents the results when the test assets are 52 anomaly portfolios of Kozak et al. (2019). Panel B presents the results when the test assets are 52 anomaly portfolios of Kozak et al. (2019), plus excess returns of 10 industry portfolios.

Panel A: Test Assets are 52-Anomalies

Non-Traded Regression R^2			Correlation of Mimicking Portfolios with FF-5					
Non-Traded Factor	R^2	Adj R^2	Mkt	SMB	HML	RMW	CMA	UMD
liq	0.30	0.20	0.56	0.15	-0.12	-0.07	-0.22	-0.08
ts	0.22	0.11	-0.29	-0.13	-0.14	0.27	-0.11	0.31
ds	0.18	0.06	0.17	0.22	0.00	-0.10	-0.08	0.02
rf	0.19	0.07	-0.33	-0.21	0.00	0.05	0.02	0.09
dy	0.59	0.53	0.82	0.34	-0.16	-0.08	-0.27	-0.19
pe	0.36	0.27	0.11	-0.04	-0.05	0.02	-0.04	0.00
svar	0.35	0.25	0.54	0.35	-0.02	-0.21	-0.12	0.06
vs	0.49	0.42	-0.31	-0.25	0.87	0.26	0.62	-0.25
interm	0.78	0.75	0.85	0.11	0.02	-0.20	-0.22	-0.29

Panel B: Test Assets are 52-Anomalies + 10 Industry Portfolios

Non-Traded Regression R^2			Correlation of Mimicking Portfolios with FF-5					
Non-Traded Factor	R^2	Adj R^2	Mkt	SMB	HML	RMW	CMA	UMD
liq	0.32	0.21	0.54	0.14	-0.11	-0.07	-0.21	-0.08
ts	0.24	0.11	-0.28	-0.12	-0.13	0.26	-0.10	0.30
ds	0.20	0.07	0.16	0.20	0.00	-0.09	-0.08	0.02
rf	0.20	0.07	-0.32	-0.20	0.00	0.05	0.02	0.08
dy	0.62	0.55	0.80	0.33	-0.16	-0.08	-0.26	-0.19
pe	0.39	0.29	0.10	-0.04	-0.05	0.01	-0.04	0.00
svar	0.36	0.25	0.53	0.35	-0.02	-0.20	-0.12	0.06
vs	0.50	0.41	-0.31	-0.25	0.86	0.26	0.62	-0.25
interm	0.80	0.77	0.84	0.11	0.02	-0.19	-0.22	-0.29

Table 18

Traded vs Non-Traded : Model Probabilities with Spurious Factors

This table compares the posterior model probabilities of traded factor asset pricing models with non-traded models. The monthly factors data span from January 1974 to December 2016. The non-traded factors are substituted with artificially simulated spurious factors, which are uncorrelated with the test assets and traded factors. The data for traded factors remain the same. The "ML" columns report the posterior model probabilities based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors. These probabilities do not account for the estimation error in the mimicking portfolios. The "EML" columns report the average posterior model probabilities using this paper's methodology, which adjusts for the estimation uncertainty in mimicking portfolios. Posterior model probabilities are averaged across 100 independent simulations of spurious non-traded factors. Test assets are monthly returns of 52 characteristic-sorted anomaly portfolios, which are used in Kozak et al. (2019). $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

Test assets : 52-anom	Prior (p_m) =1.25		Prior (p_m) =1.5		Prior (p_m) =1.75	
Model	ML	EML	ML	EML	ML	EML
CAPM	0.11%	0.10%	0.01%	0.01%	0.00%	0.00%
Fama and French (1992)	0.25%	0.25%	0.03%	0.03%	0.00%	0.00%
Barillas and Shanken (2018)	34.80%	33.95%	39.77%	37.75%	40.58%	37.94%
Fama and French (2015)	2.19%	2.14%	0.68%	0.65%	0.22%	0.21%
FF-6	3.55%	3.46%	1.37%	1.30%	0.55%	0.51%
Hou et al. (2015)	11.21%	10.93%	7.52%	7.14%	4.84%	4.52%
Stambaugh and Yuan (2017)	39.77%	38.80%	48.72%	46.24%	53.01%	49.55%
Pastor and Stambaugh (2003)	1.62%	1.16%	0.06%	0.12%	0.04%	0.12%
Hahn and Lee (2006)	0.39%	0.52%	0.10%	0.25%	0.03%	0.18%
Petkova (2006)	1.19%	1.93%	0.38%	1.45%	0.21%	1.80%
Campbell and Vuolteenaho (2004)	1.05%	1.20%	0.23%	0.57%	0.05%	0.57%
Campbell et al. (2013)	1.43%	1.98%	0.37%	1.27%	0.16%	1.27%
Campbell et al. (2018)	2.33%	3.44%	0.75%	3.18%	0.31%	3.30%
He et al (2018)	0.11%	0.14%	0.02%	0.05%	0.01%	0.01%

Table 19

Traded vs Non-Traded : Model Probabilities with Spurious Factors

This table compares the posterior model probabilities of traded factor asset pricing models with non-traded models. The monthly factors data span from January 1974 to December 2016. The non-traded factors are substituted with artificially simulated spurious factors, which are uncorrelated with the test assets and traded factors. The data for traded factors remain the same. The "ML" columns report the posterior model probabilities based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors. These probabilities do not account for the estimation error in the mimicking portfolios. The "EML" columns report the average posterior model probabilities using this paper's methodology, which adjusts for the estimation uncertainty in mimicking portfolios. Posterior model probabilities are averaged across 100 independent simulations of spurious non-traded factors. Test assets are monthly returns of 52 characteristic-sorted anomaly portfolios, which are used in Kozak et al. (2019), plus excess returns of 10 industry portfolios. $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

Test assets : 52-anom+10-ind	Prior (p_m) =1.25		Prior (p_m) =1.5		Prior (p_m) =1.75	
Model	ML	EML	ML	EML	ML	EML
CAPM	0.07%	0.07%	0.01%	0.01%	0.00%	0.00%
Fama and French (1992)	0.18%	0.18%	0.03%	0.03%	0.00%	0.00%
Barillas and Shanken (2018)	37.19%	36.40%	38.59%	37.97%	40.39%	39.30%
Fama and French (2015)	1.87%	1.84%	0.66%	0.65%	0.22%	0.22%
Carhart (1997)	3.15%	3.08%	1.33%	1.31%	0.54%	0.53%
Hou et al. (2015)	10.93%	10.70%	7.30%	7.18%	4.81%	4.68%
Stambaugh and Yuan (2017)	43.00%	42.09%	47.27%	46.51%	52.76%	51.34%
Pastor and Stambaugh (2003)	0.32%	0.39%	0.14%	0.20%	0.04%	0.17%
Hahn and Lee (2006)	0.22%	0.32%	0.27%	0.33%	0.02%	0.15%
Petkova (2006)	0.84%	1.26%	0.78%	1.45%	0.86%	1.43%
Campbell and Vuolteenaho (2004)	0.40%	0.64%	0.40%	0.54%	0.02%	0.16%
Campbell et al. (2013)	0.66%	1.08%	1.52%	1.49%	0.11%	0.63%
Campbell et al. (2018)	1.09%	1.85%	1.69%	2.31%	0.22%	1.39%
He et al. (2017)	0.07%	0.10%	0.01%	0.02%	0.00%	0.00%

Table 20
Traded versus Non-Traded versus Principal Components: Posterior Probabilities with 52 Anomalies

This table compares the posterior model probabilities of 16 prominent asset pricing models that include all traded and non-traded factors and principal components. The "ML" columns report the posterior model probabilities based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors. These probabilities do not account for the estimation error in the mimicking portfolios and PCs. The "EML" columns report the posterior model probabilities using this paper's methodology, which adjusts for the estimation uncertainty in mimicking portfolios. Test assets are monthly returns of 52 characteristic-sorted anomaly portfolios, which are used in Kozak et al. (2019). The last two rows "PC 1-5" and "PC 1-6" present the probabilities of the models, whose factors are the first five and six principal components, respectively. $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

Test assets : 52-anom	prior (p_m) =1.25		prior (p_m) =1.5		prior (p_m) =1.75	
Model	ML	EML	ML	EML	ML	EML
CAPM	0.04%	0.04%	0.00%	0.00%	0.00%	0.00%
Fama and French (1992)	0.10%	0.09%	0.02%	0.01%	0.00%	0.00%
Barillas and Shanken (2018)	18.62%	17.54%	21.27%	17.76%	23.42%	16.72%
Fama and French (2015)	0.97%	0.92%	0.38%	0.32%	0.14%	0.10%
FF-6	1.62%	1.53%	0.76%	0.64%	0.33%	0.24%
Hou et al. (2015)	5.56%	5.24%	4.11%	3.43%	2.87%	2.05%
Stambaugh and Yuan (2017)	21.53%	20.28%	26.05%	21.75%	30.58%	21.84%
Pastor and Stambaugh (2003)	0.58%	0.62%	0.19%	0.22%	0.06%	0.12%
Hahn and Lee (2006)	0.04%	0.06%	0.00%	0.01%	0.00%	0.00%
Petkova (2006)	0.66%	1.14%	0.23%	0.68%	0.07%	0.48%
Campbell and Vuolteenaho (2004)	8.32%	8.14%	7.11%	8.12%	5.79%	8.58%
Campbell et al. (2013)	7.98%	9.39%	6.71%	10.32%	5.36%	12.42%
Campbell et al. (2018)	8.23%	10.43%	6.98%	12.79%	5.61%	15.53%
He et al. (2017)	0.03%	0.03%	0.00%	0.00%	0.00%	0.00%
PC 1-5	9.28%	8.87%	8.24%	7.78%	6.97%	6.37%
PC 1-6	16.43%	15.70%	17.93%	16.16%	18.80%	15.55%

Table 21

Traded vs Non-Traded vs Principal Components: Posterior Probabilities with 52 Anomalies + 10 Industry portfolios

This table compares the posterior model probabilities of 16 prominent asset pricing models that include all traded and non-traded factors and principal components. The "ML" columns report the posterior model probabilities based on the marginal likelihood procedure of BS and CZZ, where estimated mimicking portfolios are substituted for the non-traded factors. These probabilities do not account for the estimation error in the mimicking portfolios and PCs. The "EML" columns report the posterior model probabilities using this paper's methodology, which adjusts for the estimation uncertainty in mimicking portfolios. Test assets are monthly returns of 52 characteristic-sorted anomaly portfolios, which are used in Kozak et al. (2019), plus excess returns of 10 industry portfolios. The last two rows "PC 1-5" and "PC 1-6" present the probabilities of the models, whose factors are the first five and six principal components, respectively. $k = \frac{p_m^2 \times 0.115^2}{N}$, where N equals the total number of test assets and traded factors.

Test assets : 52-anom+10-ind	Prior (p_m) =1.25		Prior (p_m) =1.5		Prior (p_m) =1.75	
Model	ML	EML	ML	EML	ML	EML
CAPM	0.12%	0.11%	0.01%	0.01%	0.00%	0.00%
Fama and French (1992)	0.26%	0.25%	0.03%	0.02%	0.01%	0.01%
Barillas and Shanken (2018)	24.63%	23.97%	31.76%	28.16%	33.38%	28.01%
Fama and French (2015)	1.92%	1.87%	0.57%	0.51%	0.37%	0.31%
FF-6	2.99%	2.91%	1.14%	1.01%	0.80%	0.67%
Hou et al. (2015)	8.67%	8.44%	6.14%	5.45%	5.28%	4.43%
Stambaugh and Yuan (2017)	27.88%	27.14%	38.90%	34.49%	42.03%	35.27%
Pastor and Stambaugh (2003)	1.69%	1.50%	0.47%	0.52%	0.29%	0.44%
Hahn and Lee (2006)	0.11%	0.17%	0.01%	0.01%	0.00%	0.01%
Petkova (2006)	0.48%	0.76%	0.07%	0.22%	0.03%	0.13%
Campbell and Vuolteenaho (2004)	8.79%	7.77%	6.28%	7.09%	5.42%	7.13%
Campbell et al. (2013)	8.71%	8.94%	6.17%	8.28%	5.30%	9.20%
Campbell et al. (2018)	9.17%	9.92%	6.67%	10.58%	5.78%	11.43%
He et al. (2017)	0.09%	0.09%	0.00%	0.00%	0.00%	0.00%
PC 1-5	0.66%	0.73%	0.11%	0.11%	0.06%	0.06%
PC 1-6	3.83%	5.41%	1.69%	3.53%	1.23%	2.88%

True Liquidity and Fundamental Prices: US Tick Size Pilot

Rohit Allena[*] Tarun Chordia [†‡]

March 31, 2021

Abstract

We develop a big-data methodology to estimate fundamental prices and true liquidity measures, explicitly considering the rounding specification due to the minimum tick size. Evaluation of the tick size pilot (TSP), which increased the tick size for some randomly chosen stocks, requires the impact of rounding. Our true liquidity measures capture the TSP-driven decreased inventory costs of market-makers, whereas traditional measures without the rounding adjustment cannot. We find that the TSP increases market-maker profits, but does not improve liquidity and price efficiency. This result contrasts existing empirical studies but is consistent with recent theoretical studies that account for rounding.

Keywords: True Liquidity, Fundamental Prices, True bid-ask spreads, US Tick Size Pilot, Machine Learning for Structural Estimation, Variational Inference, High-Frequency Data, Scalable Algorithms

[*] Goizueta Business School, Emory University, Atlanta GA 30322, email:rohit.allena@emory.edu

[†] Goizueta Business School, Emory University, Atlanta GA 30322, email:tarun.chordia@emory.edu

[‡] We thank Jonathan Brogaard, Kee Chung, Terry Hendershott, Ryan Riordan, Dominique Roesch, Gideon Saar, Xin Wang, Chen Yao, and Mao Ye for helpful comments and suggestions.

17. Introduction

The efficient incorporation of information into prices or price discovery is of primary concern to market participants and regulators. Chordia, Roll, and Subrahmanyam (2008) have shown that price discovery and efficiency is aided by market liquidity, which is also important for (i) understanding the variation in cross-sectional expected returns (Amihud and Mendelson (1986)), (ii) the investigation of the impact of high frequency trading (HFT, O'Hara (2015)), and (iii) evaluating policies such as the recent tick size pilot (TSP, Rindi and Werner (2017), Albuquerque, Song, and Yao (2020), FINRA Report (2018)), among other fundamental questions. Because stocks trade at prices rounded to the grid determined by the minimum tick size, observed prices and quoted bid-ask spreads do not represent their corresponding true values, the fundamental prices and the true spreads, that would exist in a market with no minimum tick size. Thus, the traditional liquidity and price efficiency measures that are not adjusted for rounding and are computed using quoted spreads and observed prices or the mid-point of the quoted spreads could be biased. Rounding-adjusted liquidity and price efficiency measures are extremely challenging to estimate, given the massive millisecond-level Trade and Quote (TAQ) data, and the discretization-induced non-gaussian observed prices and quotes.

Using recent advances in machine learning, we develop a novel methodology that explicitly incorporates the rounding feature to measure each stock's true liquidity and fundamental prices at the millisecond-level. The method scales to big TAQ data and structurally estimates the fundamental prices, true bid-ask spreads and its components, true effective spreads, and price discovery measures. Simulations show that our methodology recovers the fundamental simulated prices and true spreads with statistically insignificant biases, whereas the traditional measures without the rounding adjustment are grossly biased. For instance, the mean square error when naively estimating the spread can be 16 times larger.

Importantly, we demonstrate that the true liquidity and price efficiency measures resolve the seemingly contrasting conclusions drawn by the recent theoretical and empirical studies on the TSP. The Securities and Exchange Commission (SEC) conducted the TSP over the period October 2016 through September 2018, increasing the tick size from 1 cent to 5 cents for a randomly selected stock sample. The TSP was primarily designed to test whether an increased tick size i) would enhance market-makers' profits (*mmp*), thus encouraging their participation, and thereby ii) improve price discovery and liquidity in the treated stocks.[1] Existing empirical studies of the TSP (e.g., Rindi and Werner (2017), Chung, Lee, and Rösch (2019), Comerton-Forde, Grégoire, and Zhong (2019)) that have evaluated these two hypothesis use quoted spreads as a proxy for the market-makers' profits; effective spreads calculated using the mid-point of bid-ask quotes as a proxy for the transactions costs borne by the liquidity takers; and price efficiency measures computed using transaction prices or the mid-point of the quoted bid and ask prices. These measures ignore the rounding specification induced by the tick size. Thus, we argue that the existing empirical inferences lead to results that contrast with theoretical studies of TSP that explicitly incorporate the rounding specification (e.g., Li, Wang, and Ye (2020)).

In particular, market-makers charge a bid-ask spread for facilitating trades. In a world without a minimum tick, theoretical bid-ask spread models, including Demsetz (1968), Stoll (1978), Glosten and Milgrom (1985), Kyle (1985) and Roll (1984), argue that these true spreads precisely compensate for the inventory holding, order processing and adverse selection costs faced by competitive market-makers, resulting in a zero net profit in equilibrium. However, the true bid (ask) prices are rounded down (up) to the nearest tick grid. Thus, market makers earn a net profit of quoted spreads minus true spreads for each share traded. As a result, quoted spreads would not proxy for market-maker profits (mmp), whereas quoted minus true spreads would.[2] Thus, measuring mmp using quoted spreads without subtracting true spreads could lead to biased inferences.

Analogously, the traditional effective spread measures would be biased proxies for investors' transaction costs. Fundamentally, the effective spread equals twice the absolute difference between

[1] Congress passed the Jumpstart Our Business Startups Act ("Jobs Act") in 2012 with the goal of increasing the number of initial public offerings (IPOs) in the US markets with the idea that increased access to capital would lead to job creation by the smaller companies. The Jobs Act directed the SEC to conduct a study on how decimalization impacted the number of IPOs as well as the liquidity and trading of small-capitalization company stocks. The SEC decided to conduct a randomized trial to assess the impact of higher tick sizes on small firm stock liquidity, which can be particularly important for small firms as a number of papers including Amihud and Mendelson (1986), Brennan, Chordia, and Subrahmanyam (1998), and Brennan, Chordia, Subrahmanyam, and Tong (2012) have shown that higher liquidity leads to a lower cost of capital.

[2] More specifically, this is the rent earned by liquidity suppliers due to rounding.

the transaction price and the true fundamental price that would arise in a world without a minimum tick (Bessembinder (2003)). Since fundamental prices are unknown, existing studies use the quote mid-points and this biases the traditional effective spread measures. Hagstromer (2020) documents that this bias can be as high as 96% for low priced stocks. Similarly, measures of price efficiency and price impact computed using the mid-point of the quote also suffer from discretization driven biases.

To extract the rounding-adjusted liquidity and price measures, we adapt the models of Ball and Chordia (2001) and Huang and Stoll (1997) to the current HFT environment. Observed transaction prices are modeled as the discretized (rounded) sum of i) the unobserved fundamental price that evolves as a random walk, subject to information shocks and to price discovery through the market and limit orders, and ii) the impact of trading frictions due to inventory and order processing costs. Thus, over short horizons, the observed price is a discrete version of the sum of a permanent informational component and the transient components arising out of the trading mechanism. The true spread, which equals the continuous spread that would exist in the absence of the tick, is modeled as a transform of a Gaussian autoregressive process associated with the fundamental price and other structural variables such as the time of day, time between trades, and the size and depth of the prior trade. Consistent with previous work (Hasbrouck (1999a,b)), the quoted ask equals the true ask rounded up to the nearest grid point and the quoted bid is the true bid rounded down.

Estimating the unknown fundamental price and true spread values from this model is not straightforward due to two highly complex challenges. First, although the resultant model takes a bivariate state space form, the rounding destroys the Gaussian structure and the time series independence of errors, rendering standard frequentist methods such as Kalman Filter inapplicable. To address this concern, Ball and Chordia (2001) set up the problem in a Bayesian framework, and they use the Gibbs sampling procedure to estimate the posterior densities of the hidden state variables. However, this method is computationally infeasible on the millisecond TAQ data, as it requires repeated sampling of a large number of hidden state variables. Although, the method could be employed by aggregating the millisecond data to the second level, Holden and Jacobsen (2014) demonstrate that such an aggregation yields distorted liquidity measures.

Our methodology tackles both, (i) the problem of non-gaussian errors and (ii) scalability to big data. We set up the problem in a Bayesian state-space framework and instead of drawing a massive number of repeated samples, we directly approximate the posterior densities of the fundamental prices and true spreads using a known distribution function. This procedure is known as "Variational Inference." In particular, to approximate the posterior density of the latent state variables, i.e., fundamental prices and true spreads, we consider a family of densities $\mathcal{Q}$ over the hidden state variables. Each density q ($\in \mathcal{Q}$) in the family is a candidate approximation for the true posterior. The basic idea is to find the best density in the family, q^* ($\in \mathcal{Q}$), that is statistically closest (Kullback-Leibler (KL) divergence) to the true posterior density. We then use the obtained optimal density, q^*, as an approximation for the true posterior density of prices and spreads. Variational Inference, thus, turns the sampling problem into an optimization problem and the optimal density q^* is obtained by iteratively solving the first order conditions.

Hoffman, Blei, Wang, Paisley, and Edu (2013), and Blei, Kucukelbir, and McAuliffe (2017) have proposed algorithms for obtaining the first order conditions of Variational Inference problems. These algorithms are widely used in the applications of topic modeling, especially for identifying and classifying millions of words in documents into different topics. However, such algorithms do not directly apply in our framework because of the rounding specification of the errors. Noting that the conditional posterior density of hidden variables given the observed variables and other parameters could be expressed as a truncated multivariate distribution, we explicitly derive the first order conditions to approximate the posteriors. All first order conditions are obtained as closed form expressions, allowing for quick estimation.

We use the millisecond TAQ data over the two months of the non-pilot (September 2016 and November 2018) and pilot periods (November 2016 and September 2018), across nearly 2400 treated and control stocks that were part of the TSP to extract the true *mmp*, liquidity, and price efficiency measures. The pilot securities have been divided into one control group, C, of nearly 1200 stocks and three test groups of 400 stocks each - G_1, G_2 and G_3. G_1 stocks continue to trade at a one cent tick but are allowed to quote only in five cent increments, G_2 stocks are allowed to both trade and quote only in five cent increments with a few exception,[3] G_3 stocks are quoted and traded in five cent increments and are subject to a Trade-at-Prohibition rule, which generally prevents price matching by a trading center that is not displaying the best price unless an exception applies. Our methodology incorporates appropriate rounding rules for each group. The observed prices of G_1 stocks are rounded to one cent but quoted asks and bids are rounded to five cents. For G_2 and G_3 stocks both transaction prices and quotes are rounded to five cents, except that for G_2 stocks when we see transaction prices that are not on the five cent grid we round prices to one cent. For the C stocks all trades and quotes are rounded to one cent.

We first note that rounding has a large impact on the quoted spreads, which are orders of magnitude larger than the true spreads for the constrained (as compared to the unconstrained) stocks, especially during the pilot period.[4] The difference-in-differences analysis shows that the TSP leads to an increase in the quoted spreads for the constrained stocks due to the binding tick size driven discretization. Interestingly, true spreads decrease and this decrease is larger for the unconstrained stocks. The increase in the quoted spreads for the constrained stocks combined with a decline in the true spreads for constrained as well as unconstrained stocks leads to an increase in mmp across all the treated stocks. The profits per trade are ordered as $G_1 < G_2 < G_3$ and are consistent with appropriate quoting and trading restrictions imposed on each group.

Our results differ from Rindi and Werner (2017) and Chung et al. (2019), who use quoted spreads (rather than quoted minus true spreads) as proxies for mmp to argue that the TSP does not increase and may even decrease market maker profits / revenues for the unconstrained stocks. In contrast, we find that the TSP significantly increases mmp for both, the constrained and unconstrained stocks. For the unconstrained stocks, even though the quoted spreads do not increase during the pilot regime, the costs incurred by the market-makers to facilitate trades, the true spreads, decrease and this results in an increase in mmp.

We document that TSP causes a decrease in both, the adverse selection as well as the inventory

[3] Exceptions that permit executions in one cent increments are the (1) midpoint between the national or protected best bid and the national or best protected offer, (2) retail investor orders with price improvement of at least \$0.005 per share, and (3) negotiated trades.

[4] Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP.

(plus order processing) cost components of the true spreads, with a larger decline in the inventory costs. This result is consistent with the significant decrease in end-of-day inventory held by aggregate market-makers during the tick size pilot (as reported by the SEC website), which is indicative of lower inventory costs borne by the market-makers (Comerton-Forde, Hendershott, Jones, Moulton, and Seasholes (2010), Chordia and Subrahmanyam (2004), Muravyev (2016)). In contrast, the inventory cost components of the spread that are estimated using the non-rounded Huang and Stoll (1997) methodology from quoted spreads do not capture the TSP-induced decreased inventory costs reported by the SEC. In fact, the non-rounded inventory cost component increase significantly for the constrained stocks even though there is a decline in aggregate market-makers' inventory holdings.

The decline in true spreads during the TSP is consistent with the theoretical model of Li et al. (2020). They argue that, under discrete pricing and depending on mmp per trade, high-frequency traders (HFTs) and informed algorithmic traders, who are not HFTs, endogenously choose to provide liquidity. In constrained stocks, due to higher mmp per trade, HFTs compete on speed to provide liquidity. As a result, non-HFT informed traders, who are slower than HFTs and are likely to be crowded out, submit more market orders than limit orders. Yao and Ye (2018) call this "queueing equilibrium (competition)." However, informed traders are relatively more likely to submit price improving limit orders (or their limit orders are more likely to be successfully consummated) in unconstrained stocks. As a result, TSP results in lower adverse selection costs, particularly for trading unconstrained stocks. In addition, inventory costs are also lowered, as non-HFT informed traders do not holding inventory that has to be laid off. This leads to relatively lower true spreads in the unconstrained stocks during the TSP.

We also compute the realized profits per share traded, which equals the transaction price less the true ask if the trade is a customer buy, and the true bid minus the transaction price if the trade is a customer sell. The realized profits are less than half the *mmp* suggesting that, on average, liquidity demanders trade more at the quoted ask when the difference between the quoted and the true ask is lower than the difference between the true and the quoted bid and vice versa. Thus, (some) liquidity demanders are able to trade at prices where the impact of rounding is the lowest suggesting that they employ sophisticated algorithms that allow them to ascertain the rounding cost. We also provide evidence that (some) liquidity-supplying traders understand the true price and quote process. Thus, in order to back out the impact of the sophisticated trading strategies, we, as econometricians, should also use sophisticated big-data methods to estimate the true spreads and fundamental prices. This motivates our use of sophisticated big-data methods to estimate true spreads and fundamental prices.

We proxy investors' transaction costs by the effective spread, which equals twice the absolute difference between the transaction price and the true fundamental. TSP increases transaction costs for the constrained stocks. Further, our transaction cost estimates align with the appropriate restrictions imposed on stocks in each treated group. For example, traders incur lower transaction costs per trade for demanding liquidity in G_1 constrained stocks as compared to the G_2 constrained stocks because investors are allowed to trade these stocks at one cent levels. Similarly, investors incur relatively lower transaction cost per trade for stocks in G_2 compared to that of G_3 because G_2 stocks can trade at quotes that are within the five cent tick presumably when trading against quotes from other venues (including dark venues), whereas G_3 stocks cannot. When the transaction costs are estimated using the quote mid-point or a depth weighted quote mid-point as suggested by Hagstromer (2020), the effective spreads do not align with the restrictions.

Our model also allows us to estimate the proportion of price discovery directly through the new information, and indirectly through the limit and market orders. We find that the TSP does not improve, and in fact decreases the proportion of price discovery through new information, for a majority of treated stocks. This indicates that prices are proportionally more responsive to previous trades and quotes rather than to the new information, and thus are less efficient (Hendershott, Jones, and Menkveld (2011), Chordia, Green, and Kottimukkalur (2018)). Further, we follow the methodology of Chordia and Swaminathan (2000) and Hou and Moskowitz (2005) to conduct tests to ascertain the impact of TSP on the speed of incorporation of information into prices. Based on the estimated fundamental prices we find that, in general, the TSP decreases (increases) the speed of price discovery for the constrained (unconstrained) stocks. This result contrasts with Comerton-Forde et al. (2019), who document an increased speed of adjustment for the treated stocks. Since this analysis is conducted at a lower frequency of one minute, it relates more to the speed of price discovery of fundamental information, suggesting that, for the constrained stocks, TSP decreases the speed of incorporation of fundamental information into prices.

In summary, despite the TSP achieving its first objective of increasing *mmp* across the treated stocks, it does not increase liquidity or reduce trading costs across all the treated stocks. Price efficiency declines. The speed of incorporation of information into prices generally improves for the unconstrained stocks but it deteriorates for the constrained stocks. These findings are consistent with the theoretical predictions of Li et al. (2020) as well as the insights of practitioners.[5] Li et al. (2020) have argued that large TSP-induced discretization rents (i.e., *mmp*), especially on constrained stocks, promotes queue competition rather than price competition, thereby decreasing the proportion of price discovery through new information, and also delaying price discovery.

Our paper relates to the literature on estimating various liquidity measures including the components of bid-ask spreads. Whereas Hagstromer (2020) focuses only on effective spreads, and estimates it using a depth weighted quote mid-point or the micro-price (see Stoikov (2018)), our methodology provides a general framework to estimate all the fundamental market-microstructure measures including *mmp*, fundamental prices, true spreads, and price discovery. Muravyev (2016) generalizes the Huang and Stoll (1997) approach to estimate the inventory cost component of bid-ask spreads, and finds that it has a first-order effect on option prices. Bharath, Pasquariello, and Wu (2009) aggregate various measures of the adverse selection component including Roll (1984) and Huang and Stoll (1997) to construct an information asymmetry index for each firm, and find that this index impacts a firm's capital structure decisions. Our results suggest that such adverse selection and inventory cost measures could be biased in the presence of rounding.

This paper also relates to the growing literature on the applications of machine learning meth-

[5] A recent report from Mesirow Financial Equity Management argues that - "Some observations might indicate that the winners appear to be high-frequency traders, who are able to take advantage of the mandated larger spreads by capturing the difference as profit, similar to the market-makers of the pre-decimalization era, albeit on a smaller scale of volume and price. If this is the case, the pilot program is probably meeting its original objective of incentivizing market making in small stocks. The caveat is that today's market makers are no longer the research-producing institutional brokers who were compensated with hefty commissions and spreads for the risks of making markets. Instead, they are electronic traders who sometimes fight for fractions of a cent on order sizes of a much smaller magnitude." See also Chordia and Subrahmanyam (1995).

ods in finance. Our Variational Inference methodology is general and could be applied to address other non-trivial problems with big datasets in economics and finance. Whereas studies such as Gu, Kelly, and Xiu (2018), Chinco, Clark-Joseph, and Ye (2019) and Chen, Pelger, and Zhu (2019) apply machine learning techniques to empirically identify the best models or predictors, we conduct a structural estimation of a well established model. Thus, our inferences are economically interpretable, which is usually difficult with machine learning algorithms. Note also that unlike machine learning applications to predict returns at a daily, weekly, or monthly frequency, we are operating at the transaction level where predictability may not be completely arbitraged away.

18. Model

The model is a generalization of the Ball and Chordia (2001) and Huang and Stoll (1997) models and is designed to accomodate the current high frequency trading environment. It accommodates different rounding rules imposed by the tick-size pilot across the stocks in the different test groups (G_1, G_2, G_3) and the control group (C).

The observed transaction price P_t is modeled as

$$P_t \equiv [p_t^{NR}]_{Round} = [m_t + (1-\lambda)s_t Q_t/2]_{Round}, \tag{1}$$

where p_t^{NR} is the nonrounded price at time t; m_t is the fundamental price of the security at time t, immediately after a trade; Q_t is a trade indicator for buyer/seller classification of trades and is +1 if the trade is buyer initiated, -1 if the trade is seller initiated, and 0 if we are unable to sign the traded; λ is the adverse selection component of the spread; and s_t is the true spread that would obtain in a market with continuous prices i.e., a zero tick size. The notation $[.]_{Round}$ indicates rounding onto the tick grid. Thus, the observed transaction price is a result of rounding or discretization of the sum of the fundamental price and the inventory and order processing component of the spread. Note that in the presence of rounding, the disturbances in observed price changes are not Gaussian. Most market microstructure models ignore rounding, and, thus are unlikely to be correctly specified, especially if the rounding is severe.

The fundamental price (m_t) updates the past price (m_{t-1}) by incorporating any new information contained in the market orders, limit orders and other sources of publicly available information. We assume that m_t evolves as follows:

$$m_t = m_{t-1} + \overbrace{\lambda \frac{s_t Q_t}{2}}^{\text{price discovery through market orders}} + \overbrace{\epsilon_t}^{\text{price discovery through public info}} + \underbrace{\lambda_1 (\Delta A_t)\, \mathrm{D}_t^A + \lambda_2 (\Delta B_t)\, \mathrm{D}_t^B + \lambda_3\, (\Delta D_t^A)\, I_{\Delta A=0}\, Q_t + \lambda_4 (\Delta D_t^B)\, I_{\Delta B=0}\, Q_t}_{\text{price discovery through limit orders}}, \tag{2}$$

where $\{\epsilon_t, t = 1, 2, \ldots, T\}$ are *i.i.d* $N(0, \sigma_\epsilon^2)$ and represent information shocks. The second term in equation (2) is the half fraction of spread attributable to adverse selection and represents price discovery through the market orders. The final term is the contribution of limit orders to the price discovery, where A_t (B_t) are the NBBO ask (bid) quotes and D_t^A (D_t^B) are the corresponding depths at time t. $\Delta(X_t)$ denotes the first order difference $X_t - X_{t-1}$ and I is an indicator variable for when $\Delta A = 0$ or $\Delta B = 0$.

A key distinction of our model is that we allow for price discovery through limit orders. Ball and Chordia (2001) and Huang and Stoll (1997) allow for price discovery only through the adverse selection component of the market orders and through information shocks. Our specification is consistent with the recent empirical evidence of Brogaard, Hendershott, and Riordan (2019) who, due to the presence of HFTs, attribute a majority of the price discovery to limit orders. We use publicly available information including the best ask and bid quotes, corresponding depths, and their first order differences to capture price discovery through limit orders.

In a world without ticks, let a_t (b_t) be the true ask (bid) price and so $s_t = a_t - b_t$ is the true spread that market-makers charge for facilitating a trade at time t. Theoretical models on bid-ask spreads argue that these spreads compensate market makers for the inventory holding, order processing and adverse selection costs faced by competitive market makers, resulting a zero net profit in equilibrium. However, in the presence of a positive tick size, market makers round up (down) the true ask (bid) price, a_t (b_t), to the nearest grid and quote a higher (lower) ask (bid) price, A_t, (B_t). Since the quoted spreads ($A_t - B_t$) are greater than the true spreads (s_t) incurred by the market markets, they earn a net profit of $A_t - B_t - s_t$ for facilitating a trade.[6]

Under this rationale, SEC anticipates that the tick size pilot program (TSP) would increase market maker profits and encourage them to provide more liquidity. Studies that evaluate TSP including Rindi and Werner (2017) and Chung et al. (2019) use the quoted spread ($A_t - B_t$) as a proxy for profits per trade without deducting the true spreads (s_t). This could lead to biased inferences as the true spreads (s_t) are known to vary with the tick-size. For example, Goettler, Parlour, and Rajan (2009) and more recently Werner, Wen, Rindi, and Buti (2019), and Riccó, Rindi, and Seppi (2018) show that an increase in tick-size could lead to informed traders switching their orders from market to limit orders, resulting in lower adverse selection costs and, thus, lower true spreads.

Given a tick-size regime, we model the dynamics of true spreads as a first order logarithmic auto-regression with additional structural variables as in Ball and Chordia (2001) :

$$ln(s_t) = \alpha + \beta ln(s_{t-1}) + \delta ln\frac{V_{t-1}}{D_{t-1}} + \tau Time_{t-1} + d_1 BEG_t + d_2 END_t + e_t, \tag{3}$$

where $\{e_t, t = 1, 2, \ldots, T\}$ are $i.i.d$ $N(0, \sigma_e^2)$, V_{t-1} is the volume of stock transacted at the previous trade, D_{t-1} is the corresponding bid or ask depth, $Time_{t-1}$ is the time, in seconds between the last trade and the one before it and BEG_t (END_t) is an indicator variable denoting the first (last) hour of the trading day.

The regression specification is consistent with the empirical evidence in Chordia, Roll, and Subrahmanyam (2001) that shows how the relative size of trade to depth on the previous transaction possibly impacts the ensuing spread at the current transaction. The dummy variables capture the intraday seasonalities and the use of lagged time between trades is motivated by Easley and O'Hara (1992), who suggest that absence of trades may provide information about the occurence of information events. Note that, due to rounding, the quoted spread cannot be modeled as an

[6] To be precise, the realized profit from each trade is the transaction price less the true ask if the trade is a customer buy (market-maker sell) and the true bid minus the transaction price if the trade is a customer sell (market-maker buy). In the empirical section, we also estimate these realized profits for a given trade.

auto-regressive process with Gaussian errors. However, the (log) true spread lies on the real line and is modeled as in equation (3).

We denote $x_t = m_t - \lambda s_t Q_t/2$ and $\gamma_t = log(s_t)$ for algebraic convenience. Combining the above equations we have the following econometric model:

$$P_t \equiv [p_t^{NR}]_{Round} = [x_t + s_t Q_t/2]_{Round}, \tag{4}$$

$$x_t = x_{t-1} + \lambda \frac{s_{t-1}Q_{t-1}}{2} + +l_1 L_{1t} + l_2 L_{2t} + l_3 L_{3t} + l_4 L_{4t} + \epsilon_t, \tag{5}$$

$$\gamma_t = \alpha + \beta\gamma_{t-1} + d_1 D_{1t} + d_2 D_{2t} + d_3 D_{3t} + d_4 D_{4t} + e_t, \tag{6}$$

where the regression dependent variables, $L_{1t} = (\Delta A_t)\,\mathrm{D}_t^A$, $L_{2t} = (\Delta B_t)\,\mathrm{D}_t^B$, $L_{3t} = (\Delta D^A)$ $I_{\Delta A_t=0}\,Q_t$, $L_{4t} = (\Delta D^B)\,I_{\Delta B_t=0}\,Q_t$, $D_{1t} = ln\frac{V_{t-1}}{D_{t-1}}$, $D_{2t} = Time_{t-1}$, $D_{3t} = BEG_t$ and $D_{4t} = END_t$. Denoting $z_t = \{x_t, \gamma_t\}$, the system is expressed as the following first order vector autoregression (VAR(1)) model:[7]

$$z_t = \mu_t + A_t z_{t-1} + \epsilon_t. \tag{7}$$

The true spreads and fundamental prices (z_t) are not observable but the transaction prices and quoted spreads that are discretized to the nearest grid are available. We adhere to the following discretization process for the observed transaction prices and quoted spreads:

1. If the observed price, P_t, is at the ask (bid) then we assume that the nonrounded price, p_t^{NR}, has been rounded up (down) to the nearest tick. Furthermore, the bid (ask) price is assumed to have been rounded down (up). Thus, for a trade at the ask, $x_t + s_t/2 \in [P_t - tick, P_t]$ and $x_t - s_t/2 \in [B_t, B_t + tick]$. Similarly, for a trade at the bid, $x_t - s_t/2 \in [P_t, P_t + tick]$ and $x_t + s_t/2 \in [A_t - tick, A_t]$.

[7] Note that our VAR model uses only lagged information.

2. If the trade is a customer buy, $Q_t = +1$, and the price is not the same as the ask, $P_t \neq A_t$, then $x_t + s_t/2 \in [P_t - tick, P_t]$ and $x_t - s_t/2 \in [B_t, B_t + tick]$.

3. If the trade is a customer sell, $Q_t = -1$, and the price is not the same as the bid, $P_t \neq B_t$, then $x_t - s_t/2 \in [P_t, P_t + tick]$ and $x_t + s_t/2 \in [A_t - tick, A_t]$.

Thus, at each time point t, we have the following information

$$\begin{aligned} x_t + s_t/2 &\in I_{1t}, \\ x_t - s_t/2 &\in I_{2t}, \end{aligned} \tag{8}$$

where, I_{1t} and I_{2t} indicate the intervals of length the tick size, that each linear functional of the state variable must lie within. In other words, the observed information places the adjusted fundamental price plus the half-spread in one interval of length tick size and places the adjusted fundamental price minus the half-spread in another length of tick size. We use the above specification with a uniform *tick* of 1 *cent* across all the treated stocks in groups G_1, G_2, G_3 and the control group C during the non-pilot regime.

For the periods considered during the TSP, we adapt the following tick rule:

1. Considering that the SEC restricts stocks in group G_1 to quote only in 5 *cents* but are allowed to trade in 1 *cent*, we use a *tick* of 5 *cents* for rounding the quoted bid-ask spreads, and a *tick* of 1 cent for the transaction prices. This is also equivalent to rounding all the trades at the ask or the bid to 5 *cents*, and rounding the remaining trades with transaction prices between the bid-ask quotes to 1 *cent*.

2. We use a *tick* of 5 *cents* for stocks in the groups G_2, G_3 since these are restricted to quote and trade only in 5 *cents*. However, for G_2 stocks when we observe transaction prices that are not on the five cent tick grid, we round to one cent.

3. Lastly, we use a *tick* of 1 *cent* for stocks in the control group (C) because they continue to both trade and quote in 1 *cent*.

Summarizing the set of observed values at time period t as $Y_t = \{P_t, A_t, B_t, tick\}$, our interest lies in estimating the hidden state variables (z_t) that includes true spreads and fundamental prices, and the set of parameters $\Theta = \{\lambda, \beta, l_1, l_2, l_3, l_4, d_1, d_2, d_3, d_4, \sigma_\epsilon^2, \sigma_\eta^2\}$. We cast our econometric model in equations (7) and (8) into the state space framework with equation (7) as the transition equation and (8) as the measurement equation. The rounding mechanism embedded in the measurement equation destroys the Gaussian structure and the time series independence of the errors, rendering standard estimation methods (Kalman Filtering, Kalman (1960)) invalid.

Ball and Chordia (2001) employ a Bayesian procedure that accommodates discreteness of measurement errors, and involves estimating the hidden state variables by drawing a large number of simulated samples (Gibbs Sampling). Subsequently, they extract fundamental prices and true spreads of seven large and mid-cap stocks using second level data restricted to a maximum of 14,000 transactions in the sample period. This methodology is computationally infeasible on a large cross-section of stocks with the millisecond level TAQ data since it requires repeated simulations of a large number of latent variables. In particular, Ball and Chordia (2001) extract fundamental prices and true spreads for only seven large and mid-cap stocks using second-level transactions data, restricted to a maximum of 14,000 transactions in the sample period. Given that an average stock with four months of millisecond level TAQ data has over 6×10^5 transactions, conducting a tick size pilot study across 2400 stocks requires drawing simulations of about 29×10^8 ($\sim 2 \times 2400 \times 6 \times 10^5$) latent variables. Furthermore, these simulations would have to be repeated a large number of times (10,000 times in Ball and Chordia (2001)) until all the latent variables and parameters converge.

Recognizing the above challenges in existing approaches, we develop a computationally scalable Variational Inference methodology that incorporates rounding when estimating these large number of latent variables and parameters. We describe the methodology in the following section.

19. Methodology: Variational Bayesian Inference

The goal is to estimate the state variables $z_t = \{x_t, \gamma_t\}$ and the parameters Θ, having observed $\{Y_t\}$ and other explanatory variables such as $\{Q_t\}$ and $\{L_{it}, D_{it}\}_{i=1}^{4}$. We set up the estimation problem in a Bayesian framework due to the discreteness and rounding structure of the observed variables. The premise of a Bayesian framework is to place priors on the latent variables and parameters, and estimate the posterior density of latent variables and parameters given the priors and the observed data,

$$P(z_{t=1}^{T}, \Theta | Y_{t=1}^{T}) \propto P(Y_{t=1}^{T} | z_{t=1}^{T}, \Theta) P(z_{t=1}^{T} | \Theta) P(\Theta), \tag{9}$$

where $P(\Theta)$ is the prior density of the parameters, $P(Y_{t=1}^{T} | z_{t=1}^{T}, \theta)$ is the conditional likelihood of $Y_{t=1}^{T}$ given $z_{t=1}^{T}$ and Θ, and $P(z_{t=1}^{T} | \Theta)$ is the conditional likelihood of $z_{t=1}^{T}$ given Θ.

The calculation of joint posterior is generally intractable. Considering the Markovian structure of latent variables and the rounding specification of our model, the conditional posterior densities of each variable or parameter given the other parameters, variables and the data can be easily derived and expressed as known density functions. Building on this insight, Ball and Chordia (2001) obtain samples from the conditional densities $P(\Theta | z_{t=1}^{T}, Y_{t=1}^{T})$ and $P(z_t | z_{\sim t}, \Theta, Y_{t=1}^{T})$, for $t = 1, 2, \ldots T$, where $z_{\sim t}$ is the set of all latent variables excluding z_t. For a large number of draws, statistical theories (Geman and Geman (1993)) show that such samples from the conditional densities represent samples from the joint posterior density $P(z_{t=1}^{T}, \Theta | Y_{t=1}^{T})$. Subsequently, parameters and latent variables are then estimated by computing averages and standard deviations of the samples drawn from the joint posterior density. This procedure is computationally infeasible on large data sets that include the millisecond level TAQ data, since it requires drawing samples of large number of latent variables and repeating the simulation exercise many times until all the parameters converge.

Our Variational Inference based methodology directly approximates the posterior density by solving a simple optimization problem and bypasses the challenge of drawing large number of repeated samples. In particular, to approximate the posterior density $P(z_{t=1}^{T}, \Theta | Y_{t=1}^{T})$, we consider a family of known densities $\mathcal{Q}$ over the latent variables $(z_{t=1}^{T})$ and the parameters Θ. Each density q $(\in \mathcal{Q})$ in the family is a candidate approximation for the true posterior. The premise of our methodology is to find the best density in the family, q^{*} $(\in \mathcal{Q})$, that is (statistically) closest to the true posterior density in terms of Kullback-Leibler (KL) divergence,

$$\begin{aligned} q^{*}(\{z_t\}_{t=1}^{T}, \Theta) &= \arg \min_{q(\{z_t\}_{t=1}^{T}, \Theta) \in \mathcal{Q}} KL\left[q(\{z_t\}_{t=1}^{T}, \Theta) || P\left(\{z_t\}_{t=1}^{T}, \Theta | Y_{t=1}^{T}\right)\right] \\ &= \arg \min_{q(\{z_t\}_{t=1}^{T}, \Theta) \in \mathcal{Q}} E\left[\log\left(q(\{z_t\}_{t=1}^{T}, \Theta)\right)\right] - E\left[\log\left(P\left(\{z_t\}_{t=1}^{T}, \Theta | Y_{t=1}^{T}\right)\right)\right], \end{aligned} \tag{10}$$

where Kullback-Leibler (KL) distance between two densities quantifies how much the second density is different from the first, and all expectations are taken with respect to the considered density over the latent variables and parameters, $q(.)$. Finally, we approximate the posterior with the optimized member of the family $q^{*}(.)$. Variational inference thus turns the sampling problem into an optimization problem. The key is to consider a generous family of densities $\mathcal{Q}$ such that a member of the family closely approximates the true posterior, but simple enough for solving the optimization problem.

Several studies in the asset pricing literature (e.g.,Backus, Chernov, and Zin (2014)) use the KL metric to evaluate and compare prominent theoretical models such as the consumption capital asset pricing and habit models that explain the time series and cross-section of stock returns. We use this metric to find a good approximation q^{*} for the true posterior density of true spreads, fundamental prices and the parameters (Θ), $P(z_{t=1}^{T}, \Theta | Y_{t=1}^{T})$.

Minimizing the KL objective appears to be not possible since the true posterior density, $P(\{z_{t=1}^{T}, \Theta | Y_{t=1}^{T}\})$ is not known. However, a useful decomposition of the second term in equation (10) shows that this minimization objective is solvable despite the absence of true posterior density. The decomposition is as below:

$$\log\left(P(z_{t=1}^{T}, \Theta | Y_{t=1}^{T})\right) = \log\left(P(z_{t=1}^{T}, \Theta, Y_{t=1}^{T})\right) - \log\left(P(Y_{t=1}^{T})\right), \tag{11}$$

where the first term of equation (11) is the joint density of the latent variables, parameters and the observed data that can be computed using the priors on the parameters and the latent variables; and the likelihood of the data given the parameters and the latent variables. The second term is the marginal likelihood of the observed data that involves integrating the likelihood function with respect to the priors on parameters and the latent variables. Although not computable, this term is free of parameters and the latent variables, and thus is a constant with respect to any density $q(.)$, over the latent variables and parameters. Therefore, the minimization objective involves only the known priors and likelihood functions, which are solvable. In particular it is equivalent to maximizing the popularly known objective, Evidence Lower Bound ($ELBO$), which is defined below:

$$ELBO(q) = E\left[\log\left(P\left(\{z_t\}_{t=1}^{T}, \Theta, Y_{t=1}^{T}\right)\right)\right] - E\left[\log\left(q(\{z_t\}_{t=1}^{T}, \Theta)\right)\right]. \tag{12}$$

Blei et al. (2017) outline a procedure that addresses two key questions for a set of specialized models that belong to the exponential family such as topic modeling : **i**) what family of densities to consider for the approximation of the latent variables? **ii**) how to obtain the optimal density in the family that best approximates the true posterior? We generalize their theory and derive a procedure for approximating the posterior in the context of discreteness and rounding specification using two vital results:

1. The likelihood function of observed prices and spreads given true spreads, prices and parameters is a truncated bivariate normal density.

2. Truncated bivariate normal densities belong to exponential set of family.

In what follows, we lay out the procedure for choosing a family of densities, $\mathcal{Q}$ to approximate the joint posterior of the latent variables and parameters, and obtaining the optimal density $q^* \in \mathcal{Q}$ that best approximates the true posterior.

19a. Family of Densities for Approximation

Our idea is to approximate the posterior density of latent variables given observed variables by solving an optimization problem. We use a family of densities over the latent variables, parametrized by “variational parameters”. The optimization finds the member of this family, that is, the setting of “variational parameters”, which is closest to the true posterior density of latent variables.

We consider a specific family of densities, where the latent variables and parameters are independent. These are popularly known as mean-field densities and each of its candidate density is of the form:

$$q(z_{t=1}^T, \Theta|\Phi) = \Pi_{t=1}^T q(z_t; \Phi_{z_t}) \Pi_{i=1}^4 q(l_i; \Phi_{l_i}) q(d_i; \Phi_{d_i}) q(\lambda; \Phi_\lambda) q(\beta; \Phi_\beta) q(\alpha; \Phi_\alpha), \tag{13}$$

where $\Phi = \{\{\Phi_{z_t},\}_{t=1}^T, \{\Phi_{l_i}, \Phi_{d_i}\}_{i=1}^4, \Phi_\lambda, \Phi_\beta, \Phi_\alpha\}$ are the "variational parameters" governing approximate densities $q(.|\Phi)$. Our goal is then to find the optimal density $q^*(.|\Phi^*)$, or the variational parameters Φ^*, such that the density $q^*(.|\Phi^*)$ is closest to the true posterior $P(z_{t=1}^T, \Theta|y_{t=1}^T)$. Equivalently, q^* is the solution to the below optimization problem:

$$\begin{aligned} q^*(z_{t=1}^T, \Theta|\Phi^*) &= \Pi_{t=1}^T q^*(z_t; \Phi_{z_t}^*) \Pi_{i=1}^4 q^*(l_i; \Phi_{l_i}^*) q^*(d_i; \Phi_{d_i}^*) q^*(\lambda; \Phi_\lambda^*) q^*(\beta; \Phi_\beta^*) q^*(\alpha; \Phi_\alpha^*) \\ &= \arg\min KL(q(z_{t=1}^T, \Theta)||P(z_{t=1}^T, \Theta|y_{t=1}^T)) \end{aligned}$$

Before describing the procedure for obtaining the optimal density q^*, it is worth highlighting few properties of the mean-field family of densities. Note that the candidate densities $q \in \mathcal{Q}$, assumes that all the parameters and hidden-states are time-independent. However, the state variables and the parameters in the true posterior are not time- independent. Given that we are minimizing the Kullback-Leibler distance, this simple approximation, however works well in approximating the true marginal posteriors, and thus means and variances of the individual state variables. However, it may not be appropriate to consider the mean-field family to estimate the covariance between the hidden-state variables and parameters. Since we estimate fundamental prices, true spreads and parameters (Θ) using means and standard deviations of marginal posterior densities of respective variables and parameters, mean-field approximations aptly serves our purpose.

We demonstrate the success of the mean-field approximations in our context using Monte-Carlo simulations. We find that the approximation is quite precise, and recovers the true simulated values. More recently, Wang and Blei (2018) conclude that Variational Inference with mean-field families is a theoretically sound approximate inference procedure for the marginal densities even though it is not appropriate for the true joint posterior as it does not account for the covariance structure of the state variables and the parameters.

19b. Estimating the Optimal Density Function

Estimating the optimal mean-field density, $q^*(.|\Phi^*)$ that is closest to the true posterior is not straight forward and does not have a closed form solution. However, Blei et al. (2017) derive a useful result, which shows that the optimal density of an individual latent variable given the optimal densities of all other latent variables are easily obtained. Therefore, $q^*(.|\Phi^*)$ can be estimated by starting with some initial guesses on the optimal densities and recursively updating the optimal density of each variable given the optimal densities of other variables. We state the fundamental result on mean-field approximations due to Blei et al. (2017):

Proposition 1. *Given observations $\{Y_{t=1}^T\}$, the optimal mean-field density of a hidden variable $z_t, q^*\left(z_t|\Phi^*_{z_t}\right)$ given the optimal densities of other hidden variables $z_{\sim t}, q^*\left(z_{\sim t}|\Phi^*_{z_{\sim t}}\right)$ and the parameters Θ, $q^*\left(\Theta|\Phi^*\right)$ is proportional to*

$$q^*\left(z_t|\Phi^*_{z_t}\right) \propto \exp\left(E_{q^*\left(z_{\sim t}|\Phi^*_{z_{\sim t}}\right)q^*(\Theta|\Phi^*)} \log\left[f\left(Y_{t=1}^T|z_{t=1}^T,\Theta\right)f(z_{t=1}^T|\Theta)f(\Theta)\right]\right), \tag{14}$$

where $f\left(Y_{t=1}^T|z_{t=1}^T,\Theta\right)$ is the conditional density of the observed variables $Y_{t=1}^T$ given the hidden state variables $z_{t=1}^T$ and the parameters Θ; $f(z_{t=1}^T|\Theta)$ is the conditional density of hidden state-variables given the parameters Θ; and $f(\Theta)$ is the prior density of the parameters. Note that the expectation in the above equation is taken with respect to the optimal densities of excluded hidden variables $q^*(z_{\sim t}|\Phi^*_{z_{\sim t}})$ and the parameters $q^*(\Theta|\Phi^*)$. Similarly, the optimal density of parameters Θ, $q^*(\Theta|\Phi^*)$ is given by

$$q^*\left(\Theta|\Phi^*\right) \propto \exp\left(E_{q^*\left(z_{t=1}^T;\Phi^*_{z_{t=1}^T}\right)} \log\left[f\left(Y_{t=1}^T|z_{t=1}^T,\Theta\right)f(z_{t=1}^T|\Theta)f(\Theta)\right]\right). \tag{15}$$

Here, the expectation is taken with respect to the optimal densities of state variables, $q^*\left(z_{t=1}^T;\Phi^*_{z_{t=1}^T}\right)$.

Starting with some initial values for the "variational parameters" $\{\{\Phi^*_{z_t}\}_{t=1}^T,\Phi^*\}$, we can recursively update the "variational parameters" or the optimal densities given the other variational parameters using both the above equations (14), (15) until convergence. Before deriving the update equations, it is worth pointing out the similarities of this methodology with the Gibbs Sampling approach of Ball and Chordia (2001). While a large number of samples are recursively drawn in Gibbs Sampling from the conditional posteriors of a variable given other variables and parameters, Variational Inference directly updates the moments of marginal posteriors, and bypasses the challenge of drawing a large number of samples.

19c. Derivations of Updates

We derive the equations for updating the optimal density of a parameter or a hidden variable given other variables and parameters in the appendix. The final updates are given in equations equations (27), (29), (31), (32), (33). Then, we estimate true posteriors of the fundamental prices, true spreads and other parameters including the adverse selection component (λ), by recursively updating variational densities of these equations, until all the densities converge. Overall, the algorithm for approximating the true posterior is given below :

1. Set initial values of variational parameters Φ, Φ_{z_t} for approximating densities of parameters, Θ and state-variables, $z_{t=1}^{T}$ respectively.

2. Update variational density of parameters ($q^*(\Theta)$) using equations (27), (29), (31), (32).

3. Update variational density of state variables ($q^*(z_t)$) using equation (33).

4. Compute Evidence Lower Bound ($ELBO$) using equation (12), and steps 2 and 3.

5. Repeat steps 2 to 4 until $ELBO$ convergence.

6. Approximate the true posterior with $q^*(z_{t=1}^{T}, \Theta)$.

20. Performance of the Proposed Methodology

In this section, we compare our methodology to the existing procedures and show that it performs well in terms of both accurately estimating the true parameters and scaling to the Big Data of millisecond level trades and quotes (TAQ).

20a. Accuracy of the methodology: Monte-Carlo Evidence

Using Monte-Carlo simulations, we first evaluate the performance of our methodology by examining whether the estimated parameters and variables that include true spreads, effective spreads and market-makers profits are close to the respective true simulated values. We then compare our methodology with the existing procedures that naively estimate these variables with the observed transaction prices and quoted spreads, without adjusting for rounding.

In particular, given a set of regression coefficients $\{\lambda, \alpha, \beta, l_1, l_2, l_3, l_4, d_1, d_2, d_3, d_4\}$ and other explanatory variables $\{Q_t, L_{1t}, L_{2t}, L_{3t}, L_{4t}, D_{1t}, D_{2t}, D_{3t}, D_{4t}\}$, we first simulate fundamental prices (x_t) and true spreads (s_t) using the VAR(1) specification in equations (5) and (6). The regression coefficients $\{\lambda, \alpha, \beta, \ldots, d_4\}$, and the explanatory variables $\{Q_t, L_{1t}, \ldots, D_{4t}\}$ are calibrated to match their empirical counterparts for a given stock. For example, to simulate fundamental prices and true spreads for the stock of Florida Community Bank (FCB), we sign the trades during the sample period as buys and sells to obtain Q_t; we use the product of change in its best ask quote and its depth at the ask $(\Delta A_t \times D_t^A)$ as L_{1t}, and similarly for other variables $\{L_{2t}, L_{3t}, \ldots, D_{4t}\}$. For the parameters $\{\lambda, \alpha, \beta, \ldots, d_4\}$, we use respective sample estimates (OLS) of coefficients in the VAR(1) specification of (5) and (6), with mid-quotes as x_t, and half spreads $((Ask_t - Bid_t)/2)$ as s_t.

After simulating the fundamental prices and true spreads, we use the rounding rule (8) to obtain the observed ask quotes of $A_t = [x_t + s_t/2]^{up}$, and the bid quotes of $B_t = [x_t - s_t/2]^{down}$. Without loss of generality, for a grid size of 5 cents, we specify that the rounded prices and quotes are exact multiples of 5 cents. For example, if the true simulated ask quote of a stock $(x_t + s_t/2)$ is \$32.3245 cents, the observed ask quote (A_t) is rounded up to 32.35. Similarly, if the true bid quote $(x_t - s_t/2)$ is \$31.3212, then B_t is rounded down to \$31.30. Using these simulated rounded values of $\{A_t, B_t, P_t\}$, and the explanatory variables $\{q_t, L_{1t}, L_{2t}, L_{3t}, L_{4t}, D_{1t}, D_{2t}, D_{3t}, D_{4t}\}$, we implement our methodology to estimate the latent variables $\{x_t, \gamma_t\}$ and the parameters $\{\lambda, \alpha, \beta, l_1, l_2, l_3, l_4, d_1, d_2, d_3, d_4\}$, and check whether the methodology recovers the true simulated variables and parameters.

Table 31 shows the performance of our methodology, where we perform independent Monte-Carlo simulations calibrated to three randomly selected stocks, Florida Community Bank (FCB), Boston Beer Company (SAM) and AMC Entertainment Holdings (AMC). The first column represents the true simulated values, while the second and third are the estimated values using our methodology and existing procedures, respectively. We find that our methodology performs well by noting that the difference between average simulated fundamental prices and estimated fundamental prices is 0.07, 0.11 and 0.12 cents and the difference between the average simulated true effective spreads and estimated true effective spreads is 0.02, 0.05, and 0.01 cents for FCB, SAM, and AMC, respectively. We further establish the success of our methodology by showing that other variables such as price impact (λ), market maker profits (mmp) and the posterior variance estimates also closely match the corresponding true simulated values.

Rindi and Werner (2017), Chung et al. (2019) and the official report submitted by FINRA to NMS (FINRA Report (2018)) use quoted spreads as a proxy for mmp and the effective spreads measured using mid-quotes as a proxy for the transaction costs borne by the traders. Simulation evidence shows that such estimation procedures are biased. For example, when the fundamental prices and true spreads are simulated under parameters calibrated to match those of SAM, the effective spread is biased by 20% and the mmp by 363%. Also, relative to our estimation, the squared estimation error $\left[\sum_t (True_t - Estimated_t)^2\right]$ when estimating the true spread increases 11- and 16-fold for FCB and AMC respectively, when the true spreads are naively set equal to the quoted spreads. Overall, the simulation evidence not only validates our methodology but also underlines the biases in existing procedures in estimating fundamental microstructure variables.

We also assess the scalability of our methodology by examining the number of iterations required for the algorithm to converge. Figure 1 shows that for AMC, with nearly 4×10^5 transactions, the algorithm requires only 6 iterations (epochs - in machine learning parlance) for all the parameters to converge. For each stock, the algorithm with 6 iterations takes less than two minutes on a standard personal computer with $I7-(4790\ CPU, 3.6GHz)$ processor and 16 GB RAM.

21. Data

Our sample consists of daily millisecond level trade and quote data (TAQ) across all stocks included in the TSP. The data spans over two different time periods: one month before and after the TSP conclusion date (October 1, 2018), and another month before and after the beginning date of TSP (October 1, 2016). We conduct independent analysis across both the above sub-samples. The non-pilot period for the first sub-sample is November 1- November 30, 2018, whereas the pilot period is September 1- September 30, 2018. Similarly, the non-pilot period for the second sub-sample is September 1- September 30, 2016, and the pilot period is November 1- November 30, 2016. We follow Holden and Jacobsen (2014) in cleaning and matching the TAQ data. We drop all stocks without daily TAQ data, and transactions with negative or zero quoted spreads from the sample. We also filter out the stocks that are removed by the SEC from the test or control groups due to various reasons such as price decline below $1. Our final sample has 1007 (899), 391 (331), 383 (311) and 388 (317) stocks in the control, G_1, G_2, and G_3 groups, respectively over the time period September 1-30 and November 1-30, 2016 (2018) .

In the subsequent sections, we use the proposed methodology to empirically study the impact of the TSP. In particular, we test i) whether the TSP increases market-makers profits from providing liquidity and ii) whether these profits result in improved price discovery, higher liquidity, and lower transaction costs for the liquidity demanding investors.

22. Quoted Spreads, True Spreads and Market-Makers' Profits

Panel A (B) of Table 32 presents the average estimates of the quoted spreads, true spreads and market maker profits, mmp, during the non-pilot and pilot periods in 2018 (2016). Columns 2-3 present average quoted spreads across all stocks under each group in the non-pilot regime, November 1- November 30, 2018 (September 1- September 30, 2016), and the pilot regime, September 1-

September 30, 2018 (November 1- November 30, 2016) respectively. Column 4 presents the difference between the average quoted spreads under the pilot and the non-pilot regime, whereas column 5 presents the difference in differences, DD, estimator of quoted spreads under pilot and non-pilot regime with respect to the control group. Analogously, columns 6-9 (10-13) present the average true spreads (*mmp*) under the non-pilot and the pilot regime, difference, and DD for true spreads (*mmp*). All the tables will follow the same pattern in terms of reporting the average values for the non-pilot period, then the pilot period, then the difference between the pilot and the non-pilot period and finally the DD estimate.

The DD estimator is equivalent to estimating the parameter β_3 in the following regression:

$$y_{it} = \beta_0 + \beta_1 I_i^{treated} + \beta_2 I_t^{Pilot} + \beta_3 I_i^{treated} \times I_t^{pilot} + \epsilon_{it}, \tag{16}$$

where y_{it} denotes the dependent variables - quoted spreads, true spreads or *mmp* of stock i at time t; the indicator variable $I_i^{treated} = 1$ if stock i is in a treated group and 0 otherwise; $I_t^{Pilot} = 1$ if the t is in a pilot period regime and 0 otherwise. We do not include any control variables such as the trade volume in the above regression specification because we have already incorporated these at a high frequency in the dynamics of fundamental prices and true spreads (equations (5) and (6)). Our specification is consistent with the rationale that the rounding or discretization component of the spread depends only on the tick size and does not depend on other control variables, whereas the true spreads and fundamental prices are affected by other control variables. The numbers in parenthesis in Table 32 denote the standard errors of estimators. Since we extract fundamental prices and true spreads independently for each stock and under each pilot/ non-pilot regime, we cluster the standard errors with respect to stock and pilot/non-pilot regime. Unlike quoted spreads that are observed, true spreads and *mmp* are estimated using our methodology. So, we additionally adjust the standard errors of these variables for the estimation error by adding corresponding posterior variances of the estimates.[8]

[8] This is obtained as follows: $Var(X) = E(Var(X|Y)) + Var(E(X|Y))$, where $E(Var(X|Y))$ denotes the traditional standard estimators as if the true spreads and prices are known and $Var(E(X|Y))$ is the posterior variance of estimator that corrects for the estimation error. Also, we use the same adjusted variance in our simulation analysis which shows that this variance estimator is close to the true simulated variance.

The TSP increases average quoted spreads (in DD terms) for the constrained stocks but does not significantly change quoted spreads for the unconstrained stocks. It is not surprising that the quoted spreads have increased for the constrained stocks due to the increase in the tick size. The increase in quoted spreads is so large that during the pilot period in September 2018 (November 2016) the quoted spreads are 29 (12) times larger than the true spreads for the G_3 constrained stocks while during the non-pilot period in September 2016 and November 2018 the quoted spreads are only three times larger. Thus, rounding has a large impact on the observed quotes and prices and it is important to account for rounding when evaluating the impact of TSP.

Rindi and Werner (2017) have argued that the TSP has failed to achieve the intended objective of increasing *mmp* (so as to facilitate the provision of liquidity) in unconstrained stocks. However, after deducting the costs incurred by market makers to facilitate trades - the true spreads (which decrease for all treated stocks) from the quoted spreads, we find that the TSP significantly increases *mmp* across all the treated groups, for both constrained and unconstrained stocks. More specifically, around the tick-size conclusion period, we find that the TSP increases (in DD terms) the average *mmp* by 3.67 cents, 3.75 cents and 4.1 cents per share for providing liquidity to stocks in groups G_1, G_2, and G_3, respectively. Profits in the constrained (unconstrained) stocks also increase by 3.91 (2.94) cents, 3.99 (3.42) cents and 4.18 (3.84) cents for the groups G_1, G_2 and G_3, respectively. The monotonic increase in *mmp* from the G_1 to G_3 stocks is consistent with the restrictions imposed on each group. For instance, market makers earn comparatively lower profits on G_1 stocks since traders are allowed to trade these stocks at a one cent tick, possibly against limit orders on other authorized trading venues.

The *mmp* are higher for the constrained stocks because with a higher tick size a larger fraction of the quoted spread is attributable to rounding as evidenced by the fact that the difference between the quoted and true spreads during the pilot for the constrained stocks is far larger than for the unconstrained stocks. Moreover, with a larger tick size that constrains the quoted spreads, market makers face less adverse selection and inventory risk (as we document in the next section). Since price improvements are constrained, liquidity providers (who in current trading environment are mainly HFTs as suggested by Menkveld, 2013) compete to obtain time priority in the limit order book. Yao and Ye (2018) call this "queue competition" since HFTs compete on speed to reduce latency so as to be the first in the queue as per the time priority rules when orders are executed.

In the case of unconstrained stocks, market makers earn positive profits for providing liquidity despite no change in quoted spreads because of the decrease in the true spreads of these stocks. In the next section, we show that true spreads decrease because both the adverse selection and inventory (plus the order processing) costs decrease for the unconstrained stocks as well. The general decline in true spreads during both pilot periods is consistent with the theoretical model of Li et al. (2020). With an increase in the quoted spread, informed traders submit more limit orders, in the constrained as well as the unconstrained stocks, as they attempt to control trading costs; this is referred to as the "undercutting equilibrium" by Li et al. (2020). However, in the constrained stocks, due to relatively more speed competition amongst the liquidity suppliers during the pilot period, non-HFT informed traders are likely to be crowded out and are less likely to successfully consummate trades by submitting limit orders, since, with a binding tick size, the limit orders cannot be price improving. Thus, informed traders are relatively more likely to submit price improving limit orders, or their limit orders are more likely to be successfully consummated in unconstrained stocks. As a consequence, market orders are exposed to less information asymmetry, leading to lower adverse selection costs in the unconstrained stocks. Also, inventory costs are lowered, as these non-HFT informed traders are not holding inventory that has to be laid off. This leads to lower true spreads in the unconstrained stocks.

Note that an increase in market makers profits per share may not translate into an increase in total profits if the number of shares traded decrease during the TSP. Panel A (B) of Table 33 presents the average dollar value (in '000s) of costs and revenues for providing liquidity to a stock in each group around the conclusion (start) of the TSP. This is obtained by first multiplying quoted spreads, true spreads and market-maker profits by the number of shares traded for a given stock and then averaging across all stocks in a given group.

We find that the TSP increases total aggregate profits for providing liquidity to each stock across all the treated groups. In particular, during the sample period in 2018, TSP increases average profits for market-makers (in DD terms) by \$177.61, \$141.32, and \$157.03 (in '000s) for each stock in groups G_1, G_2 and G_3 respectively. Aggregate profits in constrained (unconstrained) stocks also increase by \$412.06 (\$55.69), \$303.38 (\$66.71) and \$296.38 (\$62.93) (in '000s) over the same period. As compared to *mmp* per share traded, the aggregate dollar profits for the unconstrained firms are far lower than for the constrained firms suggesting lower trading volumes in the unconstrained firms, possibly due to the higher trading costs as proxied by quoted and true spreads and, as we will show later, by effective spreads. Quoted and true spreads are significantly higher for the unconstrained stocks. For instance, the G_1 stocks constrained stocks have average quoted spreads of 1.81 (5.57) cents while the G_1 unconstrained stocks have quoted spreads of 14.63 (12.41) cents during the non-pilot (pilot) period in November (September) 2018.

Panel B of Tables 32 and 33 repeats the same analysis for the sample period in September (non-pilot) and November (pilot) 2016. We find similar results even in this sub-sample with the TSP leading to a significant increase in the quoted spreads across the constrained stocks. However, the true spreads decrease during the tick pilot, thus resulting in an increase in *mmp* across all the treated stocks, both constrained and unconstrained.

Recall that *mmp*, which equal the quoted spreads minus the true spreads, would not precisely represent market-makers' realized profits per trade. In particular, the realized profit for each share traded equals the transaction price less the true ask if the trade is a customer buy (market-maker sell), and the true bid minus the transaction price if the trade is a customer sell (market-maker buy). Unlike *mmp*, which measure profits from both sides of a trade, realized profits capture gains only from a single side of the trade, depending on whether the transaction is a buy or a sell. Panel A (B) of Table 34 presents market-makers' realized profits per share and average dollar value (in '000s) of realized profits for providing liquidity to a stock in each group around the conclusion (start) of the TSP. As before, we find that the TSP leads to a significant increase in the realized gains per share and aggregate realized profits in dollar value for providing liquidity to each stock across all the treated groups.

As with *mmp*, while the realized profits also increase monotonically from G_1 to G_3 stocks, they are much smaller relative to *mmp* from Tables 32 and 33. In fact, the realized profits are less than half the *mmp* suggesting that, on average, liquidity demanders trade more at the quoted ask when the difference between the quoted and the true ask is lower than the difference between the true and the quoted bid. Similarly, liquidity demanders trade more at the quoted bid when the difference between the quoted and the true ask is higher than the difference between the true and the quoted bid. Thus, supplying liquidity is not as profitable as expected by naively comparing quoted and true spreads because the liquidity demanders are able to trade at prices where the impact of rounding is the lowest. Given that a large fraction of trades (including liquidity demanding trades) involve HFTs, our results are consistent with Hagstromer (2020) (see also Muravyev and Pearson, 2020) who argues that HFTs have sophisticated algorithms that allow them to ascertain the true cost of trading. This also motivates our use of sophisticated big-data methods to estimate true spreads and fundamental prices.

While the realized profits of Table 34 suggest that (some) liquidity demanding traders understand the underlying true prices and spreads, we now provide evidence that even the liquidity suppliers (likely the HFTs) understand the true price and spread process. In the internet appendix Table 40, we show that liquidity suppliers are willing to supply more depth when their profits are higher, possibly due to queue competition, as suggested by Li et al. (2020). More specifically, we divide each day for each stock into two types of transactions, those that have high *mmp* and those that have low *mmp*. We find that transactions with high *mmp* have higher depth (as measured by the average of the bid and ask depth) than transactions with low *mmp*. We do not see the same pattern, except for the G_3 stocks, when we classify transactions based on high and low quoted spreads because, as we have argued, quoted spreads do not readily translate to *mmp*. Thus, (some) liquidity-supplying and liquidity-demanding traders understand the true price process and this, once again, motivates our use of sophisticated big-data methods to estimate true spreads and fundamental prices.

Overall, our results indicate that the TSP has achieved its first objective of increasing market-maker profits.

22a. True Spreads and Its Components

In this section, we examine the components of the bid-ask spread. The literature classifies true bid-ask spreads into three main components: order processing costs, inventory and adverse selection costs. Our framework allows us to separately identify adverse selection costs, and sum of order processing and inventory costs. Since order-processing costs such as computer costs, labor costs and informational service costs are largely fixed, we can assume that TSP does not impact these costs. As a result, we attribute the impact of TSP on true spreads to changes in the adverse selection and inventory components of the spread. The adverse selection and inventory component given the true spread s_t, are λs_t and $(1-\lambda)s_t$, respectively.

Panel A (B) of Table 35 shows the adverse selection and inventory cost components of the true (quoted) spread. The DD estimator of the adverse selection component from Panel A of Table 35 shows that TSP leads to a decrease in adverse selection costs across all treated stocks, more so for the unconstrained stocks. As we argued earlier, the larger decline for the unconstrained stocks is consistent with informed traders becoming more likely to relatively switch their orders from market to limit orders or their price improved limit orders are more likely to be executed. We also provide additional support to the above argument in Table 38, where we document a significant increase in the relative contribution of limit orders to the price discovery.

Panel A of Table 35 also shows that the TSP significantly decreases the inventory component of the spread across all unconstrained stocks. Recall that the inventory cost component of the spread represents the costs or risks borne by the market-makers for holding inventory to facilitate trades. We use aggregate market-maker participation data, which is publicly made available by SEC and FINRA, and show that the significant decrease in our measure of the inventory cost component of the spread is consistent with the decreased inventory risk borne by the market makers, during the TSP.[9] Appendix B (ii) of SEC data contains the daily cumulative number of share buys and sells by all registered market-makers. Based on this daily trading activity of market-makers, we use two measures that reflect the daily inventory costs/risks borne by them. The first measure is

[9] The data are available at https://www.finra.org/rules-guidance/key-topics/tick-size-pilot-program/appendix-b-data-publication.

the absolute order imbalance, which is given by

$$Inv_{1it} = |\text{Number of Shares } Bought_{it} \text{ - Number of Shares } Sold_{it}|, \text{ for stock } i, \text{ on day } t. \quad (17)$$

Higher value of Inv_{1it} implies higher trade imbalance, and thus higher inventory costs/risk borne by the market-makers for stock i at the end of day t (Comerton-Forde et al. (2010)). Further, Chordia and Subrahmanyam (2004) argue that higher order imbalance may also reflect higher adverse selection risk when informed traders optimally choose to split their orders. However, they also assert that the order imbalances are significantly predicted by the lagged order imbalances, and this predictable component captures only inventory holding risk but not adverse selection risk. The intuition is that if today's high order imbalance predicts high order imbalance even for the next day, then it indicates high inventory holidng costs or risks borne by the market-makers. Recognizing this insight, Muravyev (2016) uses the order imbalance component that is predicted by the lagged order imbalance as measure of inventory holding risk. We use the same metric as another measure of inventory holding costs, as below,

$$\begin{aligned} Inv_{2it} &= |\hat{\alpha}_i + \hat{\beta}_{1i} OIB_{i,t-1} + \hat{\beta}_{2i} OIB_{i,t-2},| \quad \text{where,} \\ OIB_{it} &= \alpha_i + \beta_{1i} OIB_{i,t-1} + \beta_{2i} OIB_{i,t-2} + \epsilon_{it}, \quad \text{and} \\ OIB_{it} &= \text{Number of Shares } Bought_{it} \text{ - Number of Shares } Sold_{it}. \end{aligned} \quad (18)$$

Table 36 shows that the DD estimator of the above measures are significantly negative across all but the G_1-constrained stocks, thus indicating that the TSP decreases inventory costs borne by the market makers. Recall from Panel A of Table 35 that the DD estimates of the derived inventory component of the spread for the unconstrained stocks, using our methodology are negative as well, and therefore consistent with the registered decrease in inventory holding costs during the TSP.

In contrast to the results in Panel A of 35, Panel B computes the adverse selection and the inventory cost component using quoted spreads and without accounting for rounding as in Huang and Stoll (1997). Both the adverse selection and the inventory cost component increase for the constrained stocks and remain unchanged for the unconstrained stocks. Ignoring discretization seems to lead to biased inferences. At least in the case of inventory costs, aggregate market maker participation data strongly indicate a decrease in inventory costs borne by the market makers.

Overall, we find that the TSP decreases true spreads, and increases market maker profits across all the treated stocks. Both, the adverse selection and inventory component of the spreads decrease. In the following sections, we ask whether these increased profits come at the expense of increased trading costs for liquidity demanders and whether price discovery improves.

23. Effective Spreads

In this section we study the impact of TSP on investor transaction costs as measured by effective spreads, which are defined as twice the absolute value of the transaction price less a reference price,

$$Effective\ Spread_t = 2|P_t - P_t^R|, \tag{19}$$

where P_t is the transaction price at time t and P_t^R is the reference price. We use three different reference prices: (i) the mid-point of the bid-ask quote, $P_t^R = (A_t + B_t)/2$, (ii) the depth weighted mid-point as in Hagstromer (2020), $P_t^R = (B_t D_t^A + A_t D_t^B)/(D_t^A + D_t^B)$, and (iii) the fundamental price, $P_t^R = m_t$. Panel A (B) of Table 37 documents the impact of TSP on effective spreads using the three reference prices for the non-pilot and pilot sample periods in 2018 (2016).

Across both panels and for all three measures, the DD results show that the effective spreads have increased for the constrained stocks but have not changed significantly for the unconstrained stocks. However, note that only in the scenario where the reference price is the fundamental price do we see a monotonic increase (2.67, 2.88, and 3.16 cents per share on average during the sample period in 2018 and 2.57, 2.74, and 3.08 cents in 2016) in the effective spreads for the G_1, G_2, and G_3 stocks. This monotonic increase is consistent with the restrictions imposed on each of the three groups of treated stocks. When the reference price is the quote mid-point or the weighted quote mid-point, there is monotonic *decrease* in the effective spreads across the G_1, G_2, and G_3 stocks.

Hagstromer (2020) has argued that the effective spread using the quote mid-point overstates the true effective spread. On the contrary, we find that, during the non-pilot and the pilot periods, the point estimate of the effective spread with the fundamental price as the reference is generally higher than that using the quote mid-point as the reference price. In fact, even with the weighted quote mid-point as the reference, the point estimate of the effective spread, during the non-pilot period, is higher than the estimate of the effective spread with the quote mid-point as the reference price. This difference in results could be driven by an assumption in Hagstromer (2020) that informed liquidity demanders submit market orders or marketable limit orders. However, if the informed traders submit limit orders that are successfully executed then it is not necessarily the case that the effective spread using the quote mid-point would overstate the true effective spread.

In sum, the increased *mmp* due to TSP have not led to a unilateral decrease in transaction costs as envisioned by the Jobs ACT. In fact, due to the impact of rounding, transaction costs faced by liquidity demanders in the constrained stocks have increased.

24. Price Discovery

We now check whether large *mmp* and market-makers participation induced by the TSP enhances price discovery directly through new information, rather than indirectly through trading. We also check whether it increases the speed of price discovery.

24a. Proportion of Price Discovery through Trading and New Information

Our framework allows us to estimate the contribution of market, limit orders and new information to the price discovery. From equation (2), for each stock, we estimate the proportion of price discovery through market orders (R^2_M) as R-squared in the regression of fundamental price changes on the signed half spread,

$$m_t - m_{t-1} = \lambda(q_t s_t)/2 + u_t. \tag{20}$$

Similarly, we estimate the proportion of price discovery through limit orders (R^2_L) as R-squared in the regression of fundamental price changes on the changes in limit order book,

$$\begin{aligned} m_t - m_{t-1} &= \lambda_1(\Delta A_t)\,\mathrm{D}^A_t + \lambda_2(\Delta B_t)\,\mathrm{D}^B_t \\ &+\lambda_3\,(\Delta D^A_t)\,I_{\Delta A=0}\,Q_t + \lambda_4(\Delta D^B_t)\,I_{\Delta B=0}\,Q_t + v_t. \end{aligned} \tag{21}$$

Lastly, we estimate the proportion of price discovery through new information (R^2_I) as one minus R-squared in the regression of fundamental price changes on the signed half spread and changes in the limit order book,

$$\begin{aligned} m_t - m_{t-1} &= \lambda(q_t s_t)/2 + \lambda_1(\Delta A_t)\,\mathrm{D}^A_t + \lambda_2(\Delta B_t)\,\mathrm{D}^B_t \\ &+\lambda_3\,(\Delta D^A_t)\,I_{\Delta A=0}\,Q_t + \lambda_4(\Delta D^B_t)\,I_{\Delta B=0}\,Q_t + \epsilon_t. \end{aligned} \tag{22}$$

The components of price discovery from market and limit orders could be correlated and thus it could be the case that $R^2_M + R^2_L + R^2_I \neq 100\%$. To interpret the relative importance of the contributions price discovery through each channel and make uniform comparisons across various stocks and over different periods, we normalize the obtained R^2s so that $R^2_M + R^2_L + R^2_I = 100\%$.

Table 38 presents the proportion of price discovery through market orders, limit orders and new information in the same format as the earlier tables. In general, the DD results show that the proportion of price discovery through market orders decreases but the proportion of price discovery through limit orders increases even more such that the proportion of price discovery through new information decreases. This does not mean that the amount of new information in the economy had declined during the pilot period. What it means is that, relative to the control stocks, the TSP has led to a decrease in the *proportion* of price discovery through new information for the treated stocks. In fact, TSP increases the proportion of price discovery from new information for the control stocks. Table 38 also shows that proportion of price discovery from new information is about three times larger than from limit orders, which in turn is about four time larger than from market orders. For instance, in the case of the control stocks, the proportion of price discovery from new information is 72.2% (75.6%); from limit orders it is 22.7% (19.5%); and from market orders it is 5.6% (5.2%) during the non-pilot (pilot) sample period in 2018. Given that a lower proportion of price discovery is through new information, this indicates that prices are more responsive to previous trades and quotes rather than to new information, and thus are less efficient as suggested by Hendershott et al. (2011) and Chordia et al. (2018).

In the case of the unconstrained stocks it is not surprising to find that the TSP leads to increased price discovery through limit orders because, as we argued earlier, the non-HFT informed traders are relatively more likely to submit price improved limit orders that are successfully executed and this is likely to lead to more information being incorporated into price through limit orders. This will lead to a lower proportion of price discovery through market orders and more through limit orders.

Even in the case of the constrained stocks, with the artificially high spreads, the non-HFT informed traders will prefer to submit fewer market orders and more limit orders that are sometimes successfully executed. This is confirmed by the TSP driven decline in the proportion of price discovery from market orders for the constrained stocks that is larger in absolute terms than the decline for the unconstrained stocks. For instance, during the sample period in 2018, the DD estimate of the proportion of price discovery from market orders is -2.4% (-1.7%), -3.1% (-1.4%), and -3.0% (-1.0%) for the G_1, G_2, and G_3 constrained (unconstrained) stocks, respectively. In addition, the queue competition leads to faster updating of the HFT limit orders and this may also lead to increased proportion of price discovery through limit orders as more information about the order flow is incorporated into prices.

Overall, TSP leads to a lower proportion of price discovery through market orders and more through limit orders for both the constrained and the unconstrained stocks. However, in the case of the constrained stocks, the caveat is that the queue competition crowds out the non-informed traders thereby reducing their incentive to collect fundamental information as suggested by O'Hara (2015) and Li et al. (2020).

24b. Speed of Price Discovery

To evaluate whether the TSP increases the speed of price of discovery, we construct price delay measures of Chordia and Swaminathan (2000) and Hou and Moskowitz (2005) (that are also used in Chung et al. (2019) and Albuquerque et al. (2020)), but measured using the true fundamental prices obtained from our methodology instead of traditional mid-quote prices. In particular, we construct three delay measures D_1, D_2 and D_3 using the following regression specifications:

$$r_{it} = \alpha_{ic} + \beta_{ic} r_{mkt,t} + \eta_{cit}, \tag{23}$$

$$r_{it} = \alpha_i + \sum_{k=0}^{5} \beta_{ik} r_{mkt,t-k} + \eta_{it}, \tag{24}$$

where r_{it} denotes returns of stock i at time t computed with fundamental prices and $r_{mkt,t}$ is the return of the SPY index at time t. Consistent with earlier work, we aggregate returns over one minute time intervals. Equation (23) is the constrained regression of stock returns on the contemporaneous market returns, whereas equation (24) represents the unconstrained regression of stock returns on the contemporaneous and several lagged market returns. By construction, R^2 in the constrained regression, (R_c^2), is always less than or equal to the R^2 in the unconstrained regression, (R_u^2). If the stock i responds immediately to market returns, then β_{ic} significantly differs from zero but none of β_{ik}, $k > 0$ significantly differ from zero, and additionally $R_c^2 = R_u^2$. If, however, stock i responds with a lag then β_{ik} differ from zero and $R_c^2 < R_u^2$. Using the above insight, we have the following three metrics that quantify delays in price discovery:

$$\begin{aligned} D_1 &= 1 - \frac{R_c^2}{R_u^2} \\ D_2 &= \frac{\sum_{k=1}^{5} k|\beta_{ik}|}{\beta_{ic} + \sum_{k=1}^{5} k|\beta_{ik}|}, \\ D_3 &= \frac{\sum_{k=1}^{5} k|z_{ik}|}{z_{ic} + \sum_{k=1}^{5} k|z_{ik}|}, \end{aligned} \tag{25}$$

where z_{ik} is the standard z-statistic for the coefficient estimate β_{ik}. A larger D_i implies that a stock's return responds with a delay to market returns. For each stock on a given day, we conduct independent regressions of equations (23) and (24) and obtain a panel data of all the above delay measures that span across all stocks and days in the sample.

The DD results in Table 39 show that the TSP causes significant delays in price discovery for constrained stocks during the sample period in 2016 and 2018, regardless of the delay measure. Recall that trading costs as proxied by the effective spreads are higher for the constrained stocks due to the binding tick size. This possibly causes delays in the speed of incorporation of information into prices because it takes the impact of information to become larger than transaction costs before it is acted upon. On the other hand, except for the delay measures $D2$ and $D3$ over the sample period in 2016, TSP generally speeds up price discovery for the unconstrained stocks. Recall that there is essentially no change in the effective spreads in unconstrained stocks and thus, transaction costs do not hinder the incorporation of information into prices. Also, with informed traders placing relatively more price improving limit orders that are successfully executed, there is more and quicker price discovery through the limit orders.

Overall, the increase in market maker profits does not translate into an increase in the speed of price discovery but promotes market making by HFTs in the constrained stocks. The results are consistent with Li et al. (2020), where increased *mmp* promotes queue competition rather than price competition, thereby decreasing the proportion of price discovery through new information, and also delaying price discovery. In the case of unconstrained stocks, with no increase in transaction costs, informed traders compete to place limit orders, thereby increasing the speed of price discovery through limit orders.

In this section, we are measuring the speed of incorporation of information into prices at the one minute frequency (equation (24) is estimated over five minutes) as opposed to our earlier estimations at the transaction-level frequency. This probably is related more to fundamental information than to order flow information. The TSP driven decline in the speed of price discovery for the constrained stocks is consistent with queue competition crowding out the non-HFT informed traders with fundamental information.

25. Conclusion

Observed prices and quoted spreads do not correspond to fundamental prices and true spreads when stocks are traded at prices rounded to the grid determined by the minimum tick size. Estimating these unobserved true liquidity and price measures are extremely challenging, not only because traditional methods cannot accommodate non-gaussian rounding errors but also computationally infeasible with the recently available millisecond-level TAQ data. We develop a novel methodology using Variational Inference that scales to high-frequency data to estimate the unobserved fundamental prices and true liquidity, explicitly accounting for the rounding specification.

We apply our method to evaluate the recently conducted tick-size pilot program (TSP). We find that the TSP increases (decreases) price discovery through limit (market) orders. The TSP increases market-maker profits but deteriorates liquidity and speed of price discovery, especially for stocks whose spreads are constrained by the tick size. These results contrast existing empirical studies but are consistent with recent theoretical studies that account for rounding.

A. Appendix

This section derives the equations for updating the optimal density of a parameter or a hidden variable given other variables and parameters. First, the joint likelihood of the observed data $\{Y_1, Y_2, \ldots Y_T\}$ and the hidden variables $\{z_1, z_2, \ldots, z_T\}$, given the parameters Θ is :

$$
\begin{aligned}
& f\left(Y_{t=1}^T, z_{t=1}^T|\Theta\right) = f\left(Y_{t=1}^T|z_{t=1}^T, \Theta\right) f(z_{t=1}^T|\Theta) \\
& \propto \Pi_{t=1}^T \exp\left(-\frac{1}{2}\left(z_{t+1} - \mu_{t+1} - A_{t+1}z_t\right)^T \Sigma^{-1}\left(z_{t+1} - \mu_{t+1} - A_{t+1}z_t\right)\right) \mathcal{I}\left(I_{1t} \leq Bz_t \leq I_{2t}\right), \quad (26)
\end{aligned}
$$

where $z_t = \begin{bmatrix} x_t \\ \gamma_t \end{bmatrix}$, $\mu_t = \begin{bmatrix} \sum_{i=1}^4 l_i L_{it} \\ \alpha + \sum_{i=1}^4 d_i D_{it} \end{bmatrix}$, $A_t = \begin{bmatrix} 1 & \lambda q_t/2 \\ 0 & \beta \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1/2 \\ 1 & -1/2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{bmatrix}$, and $\mathcal{I}(.)$ is the indicator function that takes the value 1 if the condition in (.)holds, takes zero otherwise, I_{1t}, I_{2t} are the bounds described in equation (8).

Under the diffuse prior specification for the parameters with $P(\Theta) = P\left(\Sigma, \lambda, \alpha, \beta, \{l_i, d_i\}_{i=1}^4\right) \propto \Sigma^{-\frac{2+1}{2}}$, update equations for each parameter and hidden state variable are derived as below:

Proposition 2. *The optimal density of λ, $q^*(\lambda)$ given the optimal densities of all other parameters and hidden variables is given by:*

$$
q^*(\lambda) \sim \mathcal{N}\left(\frac{E(\sum_t A_t^\lambda B_t^\lambda)}{E(\sum A_t^{\lambda^2})}, \frac{E(\sigma_\epsilon^2)}{E(\sum_t A_t^{\lambda^2})}\right), \quad (27)
$$

where $A_t = \gamma_{t-1}$, $B_t = x_t - x_{t-1} - \sum_{i=1}^4 l_i L_{it}$, and the expectations are taken with respect to the optimal variational densities of parameters and state variables other than λ.

Proof.

$$q^*(\lambda) \propto \exp\left[-\sum_{t=1}^{T} \frac{E\left(x_t - x_{t-1} - \lambda\gamma_{t-1} - \sum_{i=1}^{4} l_i L_{it}\right)^2}{2E(\sigma_\epsilon^2)}\right]$$

$$\propto exp\left[-\sum_{t=1}^{T} \frac{\left(B_t^\lambda - \lambda A_t^\lambda\right)^2}{2E(\sigma_\epsilon^2)}\right] = \exp\left[-\frac{\left(\lambda - \frac{E(\sum_t A_t^\lambda B_t^\lambda)}{\sum_t A_t^{\lambda^2}}\right)^2}{2E(\sigma_\epsilon^2)/E(\sum_t A_t^{\lambda^2})}\right]$$

$$\implies q^*(\lambda) \sim \mathcal{N}\left(\frac{E(\sum_t A_t^\lambda B_t^\lambda)}{E(\sum A_t^{\lambda^2})}, \frac{E(\sigma_\epsilon^2)}{E(\sum_t A_t^{\lambda^2})}\right) \tag{28}$$

□

Proposition 3. *The optimal density of l_i, $q^*(l_i)$ given the optimal densities of all other parameters and hidden variables is given by:*

$$q^*(l_i) \sim \mathcal{N}\left(\frac{E\left(\sum_{t=1}^{T} A_t^{l_i} L_{it}\right)}{\sum_{t=1}^{T} L_{it}^2}, \frac{E(\sigma_\epsilon^2)}{\sum_{t=1}^{T} L_{it}^2}\right), \tag{29}$$

where $A_t^{l_i} = x_t - x_{t-1} - \sum_{j \neq i} l_j L_{jt}$, and the expectation is taken with respect to optimal variational densities of parameters and state variables other than l_i.

Proof.

$$q^*(l_i) \propto \exp\left[-\sum_{t=1}^{T} E\left(l_i L_{it} - A_t^{l_i}\right)^2 / 2E(\sigma_\epsilon^2)\right] \propto \exp\left[-\frac{\left(l_i - \frac{E(\sum_{t=1}^{T} A_t^{l_i} L_{it})}{\sum_{t=1}^{T} L_{it}}\right)^2}{2E(\sigma_\epsilon^2)/(\sum_{t=1}^{T} L_{it})^2}\right]$$

$$\implies q^*(\beta) \sim \mathcal{N}\left(\frac{E(\sum_{t=1}^{T} \gamma_{t-1} A_t^\gamma)}{E\left(\sum_{t=1}^{T} \gamma_{t-1}^2\right)}, \frac{E(\sigma_\eta^2)}{E(\sum_{t=1}^{T} \gamma_{t-1}^2)}\right) \tag{30}$$

□

Proposition 4. *The optimal density of β, $q^*(\beta)$ given the optimal densities of all other parameters and hidden variables is given by:*

$$q^*(\beta) \sim \mathcal{N}\left(\frac{E(\sum_{t=1}^{T}\gamma_{t-1}A_t^{\gamma})}{E\left(\sum_{t=1}^{T}\gamma_{t-1}^2\right)}, \frac{E(\sigma_\eta^2)}{E(\sum_{t=1}^{T}\gamma_{t-1}^2)}\right), \tag{31}$$

where $A_t^{\gamma} = \gamma_t - \alpha - (\sum_{i=1}^{4} d_i D_{it})$, and the expectations are taken with respect to the variational densities of all the parameters and variables, but β.

Proposition 5. *The optimal density of each d_i, $q^*(d_i)$ given the optimal densities of all other parameters and hidden variables is given by:*

$$q^*(d_i) \sim \mathcal{N}\left(\frac{\sum E(D_{it}A_t^{D_i})}{\sum_{t=1}^{T} D_{it}^2}, \frac{E(\sigma_\epsilon^2)}{\sum_{t=1}^{T} D_{it}^2}\right), \tag{32}$$

where $A_t^{D_i} = \gamma_t - \left(\alpha + \beta\gamma_{t-1} + \sum_{k\neq i} D_{kt}d_k\right)$

Proposition 6. *The optimal density of each state variable z_t, $q^*(z_t)$ given the optimal densities of all other parameters and hidden variables is given by:*

$$z_t \sim \mathcal{N}\left(\Sigma^{*^{-1}}\left(\Sigma^{-1}\mu_{1t} + A_{t+1}^T\Sigma^{-1}\mu_{2t}\right), \Sigma^{*^{-1}}\right), I_{1t} \leq Bz_t \leq I_{2t}, \tag{33}$$

where $\mu_{1t} = A_tE(Z_{t-1}) + E(\mu_t)$ and $\mu_{2t} = E(z_{t+1})$; $\Sigma^* = \Sigma^{-1} + A_{t+1}^T\Sigma^{-1}A_{t+1}$; B, I_{1t}, I_{2t} are given in equation (26).

Proof.

$$\begin{aligned} q^*(z_t) \propto \exp&\left(-\frac{1}{2}E\left[\left(z_{t+1} - \mu_{t+1} - A_{t+1}z_t\right)^T\Sigma^{-1}\left(z_{t+1} - \mu_{t+1} - A_{t+1}z_t\right)\right]\right) \times \\ &\exp\left(-\frac{1}{2}E\left[\left(z_t - \mu_t - A_tz_{t-1}\right)^T\Sigma^{-1}\left(z_t - \mu_t - A_tz_{t-1}\right)\right]\right) \times \mathcal{I}(I_{1t} \leq Bz_t \leq I_{2t}) \end{aligned}$$

Letting $\mu_{1t} = A_t E(Z_{t-1}) + E(\mu_t)$ and $\mu_{2t} = E(z_{t+1})$, and $\Sigma^* = \Sigma^{-1} + A_{t+1}^T \Sigma^{-1} A_{t+1}$, we have

$$
\begin{aligned}
& q^*(z_t) \propto \mathcal{I}(I_{1t} \leq A z_t \leq I_{2t}) \times \\
& exp\left(-\frac{1}{2}\left(z_t - {\Sigma^*}^{-1}\left(\Sigma^{-1}\mu_{1t} + A_{t+1}^T\Sigma^{-1}\mu_t\right)\right)^T \Sigma^* \left(z_t - {\Sigma^*}^{-1}\left(\Sigma^{-1}\mu_{1t} + A_{t+1}^T\Sigma^{-1}\mu_t\right)\right)\right)
\end{aligned}
$$

Therefore,

$$
z_t \sim \mathcal{N}\left({\Sigma^*}^{-1}\left(\Sigma^{-1}\mu_{1t} + A_{t+1}^T\Sigma^{-1}\mu_t\right), {\Sigma^*}^{-1}\right), I_{1t} \leq B z_t \leq I_{2t} \tag{34}
$$

□

Thus, the optimal density of z_t follows a linearly constrained (truncated) version of a normally distributed random variable, given the moments of all other parameters and the hidden variables, z_{t-1} and z_{t+1}. The first two moments of a linearly constrained normal random variable can be computed using the procedure outlined by Kan and Robotti (2017).

References

Albuquerque, Rui A., Shiyun Song, and Chen Yao, 2020, The Price Effects of Liquidity Shocks: A Study of SEC's Tick-Size Experiment, *Journal of Financial Economics (JFE), Forthcoming* .

Amihud, Yakov, and Haim Mendelson, 1986, Asset pricing and the bid-ask spread, *Journal of Financial Economics* 17, 223–249.

Backus, David, Mikhail Chernov, and Stanley Zin, 2014, Sources of Entropy in Representative Agent Models, *The Journal of Finance* 69, 51–99.

Ball, Clifford A, and Tarun Chordia, 2001, True Spreads and Equilibrium Prices, *The Journal of Finance* 56, 1801–1835.

Bessembinder, Hendrik, 2003, Quote-based competition and trade execution costs in NYSE-listed stocks, *Journal of Financial Economics* 70, 385–422.

Bharath, Sreedhar T., Paolo Pasquariello, and Guojun Wu, 2009, Does asymmetric information drive capital structure decisions, *Review of Financial Studies* 22, 3211–3243.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe, 2017, Variational Inference: A Review for Statisticians, *Journal of the American Statistical Association* 112, 859–877.

Brennan, Michael J, Tarun Chordia, and Avanidhar Subrahmanyam, 1998, Alternative factor specifications, security characteristics, and the cross-section of expected stock returns, *Journal of Financial Economics* 49, 354–373.

Brennan, Michael J., Tarun Chordia, Avanidhar Subrahmanyam, and Qing Tong, 2012, Sell-order liquidity and the cross-section of expected stock returns, *Journal of Financial Economics* 105, 523–541.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2019, Price Discovery without Trading: Evidence from Limit Orders, *The Journal of Finance* 74, 1621–1658.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2019, Deep Learning in Asset Pricing, *SSRN Electronic Journal* .

Chinco, Alex, Adam D. Clark-Joseph, and Mao Ye, 2019, Sparse Signals in the Cross-Section of Returns, *Journal of Finance* 74, 449–492.

Chordia, Tarun, T. Clifton Green, and Badrinath Kottimukkalur, 2018, Rent seeking by low-latency traders: Evidence from trading on macroeconomic announcements, *Review of Financial Studies* 31, 4650–4687.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2001, Market Liquidity and Trading Activity, *The Journal of Finance* 56, 501–530.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2008, Liquidity and market efficiency, *Journal of Financial Economics* 87, 249–268.

Chordia, Tarun, and Avanidhar Subrahmanyam, 1995, Market Making, the Tick Size, and Payment-for-Order Flow: Theory and Evidence, Technical Report 4.

Chordia, Tarun, and Avanidhar Subrahmanyam, 2004, Order imbalance and individual stock returns: Theory and evidence, *Journal of Financial Economics* 72, 485–518.

Chordia, Tarun, and Bhaskaran Swaminathan, 2000, Trading Volume and Cross-Autocorrelations in Stock Returns, Technical Report 2.

Chung, Kee H., Albert Lee, and Dominik Rösch, 2019, Tick Size, Liquidity for Small and Large Orders, and Price Informativeness: Evidence from the Tick Size Pilot Program, *Journal of Financial Economics (JFE), Forthcoming* .

Comerton-Forde, Carole, Vincent Grégoire, and Zhuo Zhong, 2019, Inverted fee structures, tick size, and market quality, *Journal of Financial Economics* .

Comerton-Forde, Carole, Terrence Hendershott, Charles M. Jones, Pamela C. Moulton, and Mark S. Seasholes, 2010, Time variation in liquidity: The role of market-maker inventories and revenues, *Journal of Finance* 65, 295–331.

Demsetz, Harold, 1968, The Cost of Transacting, *The Quarterly Journal of Economics* 82, 33.

Easley, David, and Maureen O'Hara, 1992, Time and the Process of Security Price Adjustment, *The Journal of Finance* 47, 577–605.

FINRA Report, 2018, Assessment of the plan to implement a Tick Size Pilot program, *Originally Submitted to the NMS Plan Participants* .

Geman, Stuart, and Donald Geman, 1993, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, *Journal of Applied Statistics* 20, 25–62.

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.

Goettler, Ronald L., Christine A. Parlour, and Uday Rajan, 2009, Informed traders and limit order markets, *Journal of Financial Economics* 93, 67–87.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2018, Empirical Asset Pricing via Machine Learning, Technical report, National Bureau of Economic Research, Cambridge, MA.

Hagstromer, Bjorn, 2020, Bias in the Effective Bid-Ask Spread, *Journal of Financial Economics (JFE), Forthcoming* .

Hasbrouck, Joel, 1999a, Security bid/ask dynamics with discreteness and clustering: Simple strategies for modeling and estimation, *Journal of Financial Markets* 2, 1–28.

Hasbrouck, Joel, 1999b, The dynamics of discrete bid and ask quotes, *Journal of Finance* 54, 2109–2142.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1–33.

Hoffman, Matthew D, David M Blei, Chong Wang, John Paisley, and Jpaisley@berkeley Edu, 2013, Stochastic Variational Inference, *Journal of Machine Learning Research* 14, 1303–1347.

Holden, Craig W, and Stacey Jacobsen, 2014, Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions, *THE JOURNAL OF FINANCE* • LXIX.

Hou, Kewei, and Tobias J. Moskowitz, 2005, Market frictions, price delay, and the cross-section of expected returns.

Huang, Roger D., and Hans R. Stoll, 1997, The Components of the Bid-Ask Spread: A General Approach, *Review of Financial Studies* 10, 995–1034.

Kalman, R E, 1960, A New Approach to Linear Filtering and Prediction Problems, *Transactions of the ASME-Journal of Basic Engineering* 82, 35–45.

Kan, Raymond, and Cesare Robotti, 2017, On Moments of Folded and Truncated Multivariate Normal Distributions, *Journal of Computational and Graphical Statistics* 26, 930–934.

Kyle, Albert S., 1985, Continuous Auctions and Insider Trading, *Econometrica* 53, 1315.

Li, Sida, Xin Wang, and Mao Ye, 2020, Who Provides Liquidity, and When?, *Journal of Financial Economics (JFE), Forthcoming* .

Muravyev, Dmitriy, 2016, Order Flow and Expected Option Returns, *Journal of Finance* 71, 673–708.

O'Hara, Maureen, 2015, High frequency market microstructure, *Journal of Financial Economics* 116, 257–270.

Riccó, Roberto, Barbara Rindi, and Duane J Seppi, 2018, Information, Liquidity, and Dynamic Limit Order Markets *, Technical report.

Rindi, Barbara, and Ingrid M. Werner, 2017, U.S. Tick Size Pilot, *SSRN Electronic Journal* .

Roll, Richard, 1984, A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market, *The Journal of Finance* 39, 1127–1139.

Stoikov, Sasha, 2018, The micro-price: a high-frequency estimator of future prices, *Quantitative Finance* 18, 1959–1966.

Stoll, Hans R, 1978, The Supply ofDealer Services in Securities Markets, *The Journal of Finance* XXXl, 1133–1151.

Wang, Yixin, and David M. Blei, 2018, Frequentist Consistency of Variational Bayes, *Journal of the American Statistical Association* 1–15.

Werner, Ingrid, Yuanji Wen, Barbara Rindi, and Sabrina Buti, 2019, Tick Size, Trading Strategies and Market Quality.

Yao, Chen, and Mao Ye, 2018, Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls, *The Review of Financial Studies* 31, 2157–2183.

Figure 12. ELBO (y-axis) vs Number of Epochs (x-axis)

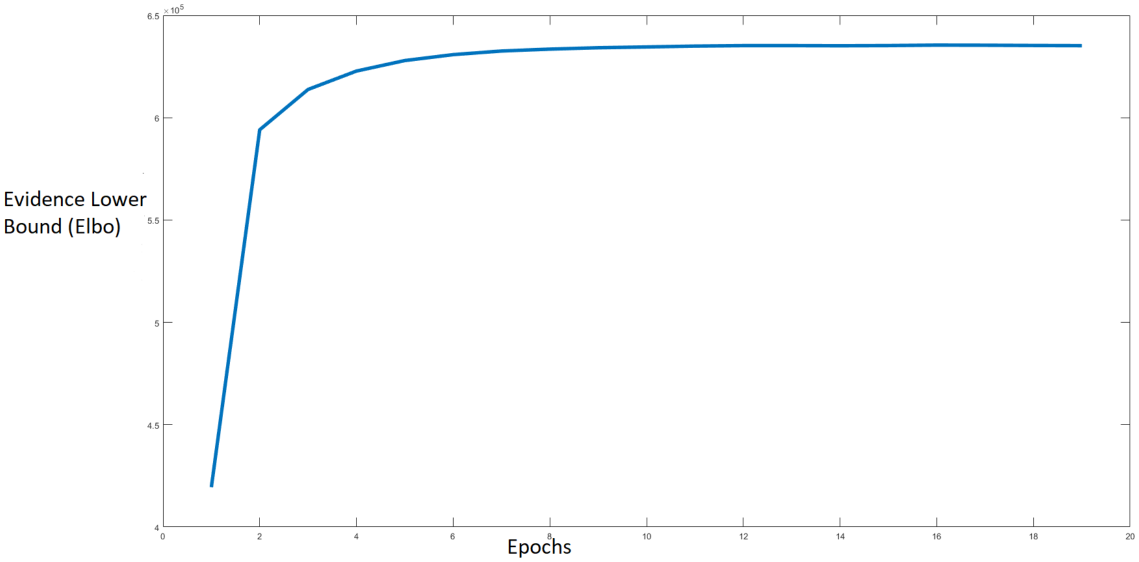


Table 31
Performance of the Methodology: Monte Carlo Evidence

This table reports the performance of our methodology in estimating the true simulated parameters. We conduct monte-carlo simulations calibrated to three randomly selected stocks, FCB, SAM and AMC. The first column represents the true simulated values, the second reports the estimated values using our methodology and the final column depicts the naively estimated parameters with the observed transaction prices and quotes, without adjusting for rounding. The parameters include means and standard deviations of true spreads (tspr), fundamental prices (price), effective spreads (espr) and market-maker profits (mmp) per share traded. All of the above values are reported in dollars. It also includes squared sum of error in estimating fundamental prices and true spreads.

Stock	Variable	True Simulated Values	Estimated Values	Naïve Values
FCB	mean tspr	0.0441	0.0445	0.0941
	std tspr	0.0275	0.0215	0.0338
	mean price	58.3545	58.3538	58.3489
	std price	14.3335	14.3349	14.3326
	mean espr	0.0861	0.0859	0.0941
	std espr	0.0367	0.0365	0.0340
	mean mmp	0.0501	0.0497	0.0941
	std mmp	0.0199	0.0221	0.0338
	Price Impact (λ)	0.1923	0.1927	0.0917
	Estimation Error - tspr		15.8693	170.6501
	Estimation Error - price		9.3137	45.9682
SAM	mean tspr	0.1817	0.1821	0.2318
	std tspr	0.1221	0.1146	0.1237
	mean price	222.2710	222.2699	222.2585
	std price	38.1229	38.1425	38.1226
	mean espr	0.1931	0.1936	0.2318
	std espr	0.1006	0.1021	0.1237
	mean mmp	0.0501	0.0497	0.2318
	std mmp	0.0197	0.0269	0.1237
	Price Impact (λ)	0.1858	0.1888	0.1530
	Estimation Error - tspr		27.5806	126.3707
	Estimation Error - price		11.6936	536.0978
AMC	mean tspr	0.0459	0.0460	0.0960
	std tspr	0.0247	0.0278	0.0318
	mean price	37.1058	37.1046	37.1022
	std price	2.5691	2.5748	2.5687
	mean espr	0.0925	0.0926	0.0960
	std espr	0.0368	0.0368	0.0318
	mean mmp	0.0501	0.0500	0.0960
	std mmp	0.0201	0.0188	0.0318
	Price Impact (λ)	0.0709	0.0695	0.0353
	Estimation Error - tspr		9.6458	157.3844
	Estimation Error - price		8.1756	42.2229

Table 32
Quoted Spreads, True Spreads and Market-Maker Profits

This table reports the average estimated values of quoted spreads, true spreads and market-maker profits per share traded across all the treated groups G_1, G_2, G_3, and the Control group. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column "Non-Pilot" ("Pilot") presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column "Diff" is the difference estimate of variables of each group, prior and post the pilot program. "DD" is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Panel A: Non-Pilot Period = November 1-30, 2018 ; Pilot Period = September 1-30, 2018

	Quoted Spread				True Spread				Market-Maker Profits			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.0181	0.0557	0.0376** (0.0014)	0.0425** (0.00397)	0.0057	0.0054	−0.0003 (0.00103)	0.0034 (0.00335)	0.0123	0.0502	0.0379** (0.00038)	0.0391** (0.00083)
G1 Unconstrained	0.1463	0.1241	−0.0221* (0.0113)	−0.0174 (0.01189)	0.1137	0.0633	−0.0504** (0.01032)	−0.0467** (0.0108)	0.0325	0.0608	0.0282** (0.00143)	0.0294** (0.00161)
G1	0.0517	0.0753	0.0236** (0.0048)	0.0284** (0.00606)	0.0336	0.0217	−0.012** (0.00402)	−0.0083* (0.00513)	0.0181	0.0536	0.0355** (0.00088)	0.0367** (0.00115)
G2 Constrained	0.0181	0.0547	0.0365** (0.0011)	0.0414** (0.00388)	0.0057	0.0035	−0.0022** (0.0008)	0.0015 (0.00329)	0.0124	0.0512	0.0388** (0.00034)	0.0399** (0.00082)
G2 Unconstrained	0.1380	0.1221	−0.016 (0.0108)	−0.0119 (0.0115)	0.1054	0.0565	−0.049** (0.00951)	−0.0453** (0.01003)	0.0326	0.0656	0.033** (0.00162)	0.0342** (0.00178)
G2	0.0701	0.0844	0.0142** (0.0065)	0.019** (0.00751)	0.0488	0.0266	−0.0221** (0.00541)	−0.0184** (0.00628)	0.0214	0.0577	0.0363** (0.0012)	0.0375** (0.00141)
G3 Constrained	0.0187	0.0551	0.0364** (0.001)	0.0412** (0.00385)	0.0061	0.0019	−0.0042** (0.00065)	−0.0005 (0.00325)	0.0126	0.0531	0.0406** (0.00039)	0.0418** (0.00084)
G3 Unconstrained	0.1620	0.1323	−0.0297 (0.0204)	−0.0248 (0.02077)	0.1291	0.0622	−0.0669** (0.01926)	−0.0632** (0.01952)	0.0329	0.0702	0.0372** (0.00415)	0.0384** (0.00422)
G3	0.0625	0.0785	0.016** (0.0064)	0.0209** (0.00744)	0.0432	0.0194	−0.0238** (0.00583)	−0.0201** (0.00664)	0.0194	0.0592	0.0398** (0.00176)	0.041** (0.00191)
Control	0.0610	0.0561	−0.0049 (0.0037)		0.0411	0.0374	−0.0037 (0.00319)		0.0199	0.0187	−0.0012 (0.00074)	

Panel B: Non-Pilot Period = September 1-30, 2016 ; Pilot Period = November 1-30, 2016

	Quoted Spread				True Spread				Market-Maker Profits			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.0186	0.0583	0.0398** (0.001)	0.0315** (0.00281)	0.0064	0.0071	0.0007 (0.00076)	−0.0057** (0.00238)	0.0122	0.0513	0.0391** (0.0003)	0.0373** (0.0007)
G1 Unconstrained	0.1503	0.1487	−0.0016 (0.0125)	−0.0099 (0.01278)	0.1184	0.0836	−0.0349** (0.01162)	−0.0413** (0.01183)	0.0319	0.0651	0.0332** (0.00344)	0.0314** (0.0035)
G1	0.0392	0.0741	0.0349** (0.0034)	0.0266** (0.00432)	0.0236	0.0201	−0.0035 (0.003)	−0.01** (0.00375)	0.0156	0.0540	0.0384** (0.00108)	0.0366** (0.00125)
G2 Constrained	0.0187	0.0588	0.0402** (0.0012)	0.0319** (0.00289)	0.0064	0.0064	−0.0001 (0.00085)	−0.0065** (0.00241)	0.0122	0.0525	0.0403** (0.0004)	0.0385** (0.00075)
G2 Unconstrained	0.1334	0.1448	0.0114 (0.0117)	0.0032 (0.01197)	0.1038	0.0751	−0.0287** (0.0107)	−0.0352** (0.01093)	0.0296	0.0697	0.0401** (0.00216)	0.0383** (0.00225)
G2	0.0411	0.0765	0.0354** (0.0037)	0.0272** (0.0045)	0.0251	0.0200	−0.0051* (0.00302)	−0.0116** (0.00377)	0.0160	0.0565	0.0405** (0.00089)	0.0388** (0.00109)
G3 Constrained	0.0193	0.0612	0.0419** (0.0017)	0.0336** (0.00311)	0.0069	0.0052	−0.0017* (0.00102)	−0.0081** (0.00247)	0.0124	0.0560	0.0436** (0.00074)	0.0418** (0.00097)
G3 Unconstrained	0.1343	0.1717	0.0374 (0.0288)	0.0292 (0.02892)	0.1045	0.0964	−0.0081 (0.02837)	−0.0146 (0.02845)	0.0298	0.0753	0.0455** (0.00602)	0.0437** (0.00605)
G3	0.0385	0.0734	0.0349** (0.0032)	0.0266** (0.00416)	0.0230	0.0150	−0.008** (0.00302)	−0.0145** (0.00376)	0.0155	0.0584	0.0429** (0.00162)	0.0411** (0.00174)
Control	0.0435	0.0518	0.0082** (0.0026)		0.0272	0.0336	0.0065** (0.00225)		0.0164	0.0182	0.0018** (0.00063)	

Table 33

Aggregated Dollar Value of Quoted Spreads, True Spreads and Market-Maker Profits

This table reports the average estimated dollar value (in 000's) of quoted spreads, true spreads and market-maker profits per stock, aggregated across all the trades for all the treated groups G_1, G_2, G_3, and the Control group. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column "Non-Pilot" ("Pilot") presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column "Diff" is the difference estimate of variables of each group, prior and post the pilot program. "DD" is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Panel A: Non-Pilot Period = November 1-30, 2018 ; Pilot Period = September 1-30, 2018

	Quoted Spread				True Spread				Market-Maker Profits			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	226.44	603.45	377.02** (49.37)	447.53** (50.9)	68.79	57.26	−11.52 (11.46)	35.47** (15.7)	157.65	546.19	388.54** (51.63)	412.06** (51.68)
G1 Unconstrained	273.00	189.72	−83.28** (15.65)	−12.76 (19.95)	207.65	92.20	−115.45** (13.92)	−68.45** (17.58)	65.35	97.52	32.17** (3.39)	55.69** (4.1)
G1	244.74	327.09	82.35** (20.17)	152.87** (23.66)	144.95	73.21	−71.74** (14.2)	−24.74 (17.8)	99.79	253.88	154.09** (17.72)	177.61** (17.87)
G2 Constrained	182.09	430.21	248.12** (23.37)	318.64** (26.44)	57.14	25.40	−31.73** (6.18)	15.27 (12.38)	124.95	404.81	279.86** (23.07)	303.38** (23.18)
G2 Unconstrained	364.09	244.79	−119.3** (19.51)	−48.78** (23.1)	270.76	108.27	−162.49** (16.97)	−115.49** (20.08)	93.33	136.52	43.19** (4.61)	66.71** (5.15)
G2	272.79	285.08	12.29 (19.65)	82.81** (23.22)	177.63	72.12	−105.51** (16.41)	−58.51** (19.6)	95.16	212.96	117.8** (8.2)	141.32** (8.52)
G3 Constrained	165.34	396.24	230.9** (17.35)	301.42** (21.31)	54.83	12.87	−41.96** (4.89)	5.04 (11.79)	110.51	383.37	272.86** (16.61)	296.38** (16.77)
G3 Unconstrained	311.22	187.41	−123.82 (28.41)	−53.3* (30.99)	242.04	78.81	−163.23** (27.51)	−116.23** (29.53)	69.19	108.60	39.41** (2.89)	62.93** (3.7)
G3	240.31	266.17	25.87 (17.6)	96.38** (21.52)	154.35	46.70	−107.65** (15.32)	−60.65** (18.7)	85.96	219.47	133.51** (6.67)	157.03** (7.06)
Control	284.75	214.23	−70.52** (12.37)		177.50	130.50	−47** (10.73)		107.26	83.74	−23.52** (2.31)	

Panel B: Non-Pilot Period = September 1-30, 2016 ; Pilot Period = November 1-30, 2016

	Quoted Spread				True Spread				Market-Maker Profits			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	147.56	485.03	337.47** (16.21)	277.84** (18.38)	54.52	58.06	3.54 (5.91)	−36.75** (9.3)	93.04	426.97	333.93** (15.18)	314.59** (15.28)
G1 Unconstrained	172.37	195.17	22.8 (16.49)	−36.83** (18.62)	133.53	105.77	−27.76** (13.12)	−68.05** (14.95)	38.84	89.40	50.56** (5.84)	31.22** (6.1)
G1	144.74	317.32	172.58** (13.51)	112.95** (16.04)	83.24	76.83	−6.41 (9.88)	−46.7** (12.21)	61.50	240.49	178.99** (7.09)	159.64** (7.31)
G2 Constrained	147.12	492.04	344.92** (20.41)	285.29** (22.17)	51.09	44.88	−6.21 (5.94)	−46.5** (9.31)	96.03	447.16	351.13** (20.39)	893.49** (20.47)
G2 Unconstrained	165.41	188.49	23.08* (13.08)	−36.56** (15.68)	126.85	95.63	−31.22** (11.07)	−71.51** (13.19)	38.56	92.86	54.3** (4.24)	596.66** (4.59)
G2	147.54	319.89	172.36** (12.15)	112.72** (14.92)	85.45	70.43	−15.03 (9.14)	−55.32** (11.62)	62.08	249.47	187.38** (7.51)	729.74** (7.72)
G3 Constrained	167.41	474.89	307.48** (19.18)	247.84** (21.04)	63.55	36.82	−26.73** (7.44)	−67.02** (10.34)	103.86	438.07	334.2** (18.93)	876.57** (19.01)
G3 Unconstrained	217.97	173.59	−44.39 (27.32)	−104.02** (28.66)	164.94	90.18	−74.75** (23.43)	−115.04** (24.5)	53.04	83.40	30.37** (7.31)	572.73** (7.52)
G3	167.38	309.11	141.73** (13.9)	82.1** (16.38)	95.68	50.53	−45.15** (10.51)	−85.44** (12.73)	71.70	258.59	186.89** (7.24)	729.25** (7.46)
Control	155.68	215.31	59.63** (8.65)		93.27	133.55	40.29** (7.18)		62.41	81.76	19.34** (1.77)	

Table 34
Realized Market-Maker Profits

This table reports the average estimated realized market-maker profits per share traded, and the average dollar value (in 000's) of the realized profits per stock, aggregated across all the trades for all the treated groups G_1, G_2, G_3, and the Control group. The realized market-maker profit per trade is equal to the actual transaction price of the trade minus the true ask for the customer buy (market maker sell) transactions and the true bid minus the transaction price for the customer sell (market maker buy) transactions. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column "Non-Pilot" ("Pilot") presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column "Diff" is the difference estimate of variables of each group, prior and post the pilot program. "DD" is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Panel A: Non-Pilot Period = November 1-30, 2018 ; Pilot Period = September 1-30, 2018

	Realized Profits				Aggregate Dollar Realized Profits			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.0051	0.0183	0.0132^{**} (0.00045)	0.0128^{**} (0.00108)	61.36	184.73	123.37^{**} (17.44058)	127.62^{**} (17.44)
G1 Unconstrained	0.0051	0.0191	0.014^{**} (0.00112)	0.0136^{**} (0.00149)	9.64	27.00	17.36^{**} (0.83074)	21.61^{**} (0.83)
G1	0.0051	0.0189	0.0138^{**} (0.00051)	0.0134^{**} (0.00111)	28.66	82.05	53.38^{**} (6.22149)	57.63^{**} (6.22)
G2 Constrained	0.0051	0.0203	0.0152^{**} (0.00069)	0.0147^{**} (0.0012)	48.18	149.99	101.82^{**} (8.52915)	106.06^{**} (8.53)
G2 Unconstrained	0.0051	0.0235	0.0184^{**} (0.00055)	0.018^{**} (0.00112)	13.46	43.38	29.92^{**} (1.26057)	34.16^{**} (1.26)
G2	0.0052	0.0220	0.0168^{**} (0.00037)	0.0164^{**} (0.00105)	23.27	75.48	52.2^{**} (3.1418)	56.45^{**} (3.14)
G3 Constrained	0.0050	0.0224	0.0174^{**} (0.00032)	0.0169^{**} (0.00103)	41.44	155.63	114.19^{**} (7.69157)	118.44^{**} (7.69)
G3 Unconstrained	0.0051	0.0286	0.0235^{**} (0.00648)	0.0231^{**} (0.00656)	10.09	39.77	29.68^{**} (2.30997)	33.93^{**} (2.31)
G3	0.0050	0.0249	0.0199^{**} (0.00259)	0.0195^{**} (0.00277)	22.53	87.05	64.52^{**} (2.95481)	68.77^{**} (2.95)
Control	0.0050	0.0054	0.0004^{**} (0.00054)		28.12	23.87	-4.25^{**} (0.86475)	

Panel B: Non-Pilot Period = September 1-30, 2016 ; Pilot Period = November 1-30, 2016

	Realized Profits				Aggregate Dollar Realized Profits			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.0054	0.0198	0.0144^{**} (0.00031)	0.0147^{**} (0.00103)	36.69	154.63	117.9407^{**} (6.40375)	113.98^{**} (6.4)
G1 Unconstrained	0.0052	0.0171	0.0118^{**} (0.00261)	0.0121^{**} (0.00279)	5.59	22.02	16.4279^{**} (1.56207)	12.47^{**} (1.56)
G1	0.0054	0.0194	0.014^{**} (0.00085)	0.0143^{**} (0.0013)	19.51	81.26	61.7585^{**} (2.83585)	57.8^{**} (2.84)
G2 Constrained	0.0054	0.0211	0.0157^{**} (0.00841)	0.016^{**} (0.00846)	37.35	174.26	136.9043^{**} (8.43943)	132.95^{**} (8.44)
G2 Unconstrained	0.0054	0.0224	0.017^{**} (0.00231)	0.0173^{**} (0.00251)	6.14	28.43	22.2898^{**} (1.1284)	18.33^{**} (1.13)
G2	0.0054	0.0214	0.016^{**} (0.00288)	0.0163^{**} (0.00304)	19.23	91.98	72.7542^{**} (3.15926)	68.8^{**} (3.16)
G3 Constrained	0.0056	0.0239	0.0183^{**} (0.00177)	0.0186^{**} (0.00202)	40.17	177.57	137.4054^{**} (9.21486)	133.45^{**} (9.21)
G3 Unconstrained	0.0051	0.0258	0.0208^{**} (0.00975)	0.0211^{**} (0.0098)	8.26	27.76	19.5053^{**} (2.66529)	15.55^{**} (2.67)
G3	0.0054	0.0242	0.0187^{**} (0.00287)	0.019^{**} (0.00303)	22.62	102.66	80.0478^{**} (3.28981)	76.09^{**} (3.29)
Control	0.0054	0.0051	-0.0003^{**} (0.00098)		18.87	22.82	3.9557^{**} (0.57055)	

Table 35
Components of Bid-Ask Spread

This table presents the average estimated values of the adverse selection and inventory cost component per share traded of the true (Panel A) and quoted spread (Panel B) using our methodology and that using the Huang and Stoll (1997) methodology, respectively, across all the treated groups G_1, G_2, G_3, and the Control group. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column "Non-Pilot" ("Pilot") presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column "Diff" is the difference estimate of variables of each group, prior and post the pilot program. "DD" is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Non-Pilot = Sep 1-30, 2016 & Nov 1-30, 2018 ; Pilot = Nov 1-30, 2016 & Sep 1-30, 2018

Panel A	Adverse Selection using True Spreads				Inventory Costs using True Spreads			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.0010	0.0005	−0.0004** (0.0001)	−0.0002 (0.0003)	0.0050	0.0058	0.0008 (0.0006)	0.0004 (0.0023)
G1 Unconstrained	0.0119	0.0067	−0.0052** (0.0007)	−0.0049** (0.0008)	0.1031	0.0636	−0.0394** (0.0089)	−0.0398** (0.0091)
G1	0.0035	0.0020	−0.0015** (0.0003)	−0.0013** (0.0004)	0.0258	0.0189	−0.007** (0.0027)	−0.0073** (0.0035)
G2 Constrained	0.0011	0.0007	−0.0004** (0.0001)	−0.0001 (0.0003)	0.0050	0.0047	−0.0004 (0.0006)	−0.0007 (0.0023)
G2 Unconstrained	0.0113	0.0077	−0.0036** (0.0008)	−0.0033** (0.0008)	0.0937	0.0534	−0.0403** (0.0076)	−0.0407** (0.0079)
G2	0.0046	0.0030	−0.0016** (0.0005)	−0.0013** (0.0005)	0.0330	0.0200	−0.013** (0.0034)	−0.0133** (0.0041)
G3 Constrained	0.0012	0.0006	−0.0005** (0.0001)	−0.0003 (0.0003)	0.0054	0.0026	−0.0028** (0.0005)	−0.0032 (0.0023)
G3 Unconstrained	0.0124	0.0085	−0.0039** (0.0014)	−0.0037** (0.0014)	0.1089	0.0573	−0.0516** (0.0147)	−0.052** (0.0149)
G3	0.0040	0.0026	−0.0014** (0.0004)	−0.0011** (0.0005)	0.0293	0.0144	−0.0149** (0.0035)	−0.0153** (0.0041)
Control	0.0044	0.0041	−0.0003 (0.0002)		0.0310	0.0313	0.0004 (0.0022)	

Panel B	Adverse Selection using Quoted Spreads				Inventory Costs using Quoted Spreads			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.0054	0.0125	0.0071** (0.0005)	0.0079** (0.001)	0.0129	0.0447	0.0318** (0.0006)	0.0309** (0.0021)
G1 Unconstrained	0.0459	0.0415	−0.0044 (0.0038)	−0.0036 (0.0039)	0.1014	0.0912	−0.0102 (0.0068)	−0.0111 (0.0071)
G1	0.0140	0.0194	0.0054** (0.0013)	0.0062** (0.0015)	0.0324	0.0553	0.0229** (0.0024)	0.022** (0.0032)
G2 Constrained	0.0055	0.0130	0.0074** (0.0004)	0.0082** (0.0009)	0.0129	0.0444	0.0315** (0.0007)	0.0307** (0.0021)
G2 Unconstrained	0.0418	0.0392	−0.0027 (0.0034)	−0.0019 (0.0035)	0.0951	0.0887	−0.0064 (0.0061)	−0.0073 (0.0064)
G2	0.0171	0.0215	0.0044** (0.0015)	0.0052** (0.0018)	0.0393	0.0585	0.0193** (0.0032)	0.0184** (0.0038)
G3 Constrained	0.0055	0.0111	0.0056** (0.0005)	0.0064** (0.001)	0.0135	0.0463	0.0328** (0.0006)	0.0319** (0.0021)
G3 Unconstrained	0.0483	0.0415	−0.0068 (0.0065)	−0.006 (0.0065)	0.1050	0.0950	−0.0099 (0.011)	−0.0108 (0.0111)
G3	0.0154	0.0185	0.0031** (0.0016)	0.0039** (0.0018)	0.0353	0.0572	0.0219** (0.0029)	0.021** (0.0036)
Control	0.0157	0.0150	−0.0008 (0.0009)	0.0381	0.0389	0.0009 (0.002)		

Table 36
Inventory Risks of Aggregate Market-Makers

This table reports the averages of two measures of inventory costs estimated directly from the SEC market-makers' participation data, across all the treated groups G_1, G_2, G_3, and the Control group. The inventory costs are estimated using order imbalances, measured as the difference in the number of shares bought less the number of shares sold. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column "Non-Pilot" ("Pilot") presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column "Diff" is the difference estimate of variables of each group, prior and post the pilot program. "DD" is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Inventory Costs of Aggregate Market-Makers

Non-Pilot = Sep 1-30, 2016 & Nov 1-30, 2018 ; Pilot = Nov 1-30, 2016 & Sep 1-30, 2018

	Order Imbalance (Inv_{1it})				Expected Order Imbalance (Inv_{2it})			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	13375	17597	4221^{**} (493)	1897^{**} (378)	6587	8402	1815^{**} (200)	713^{**} (148)
G1 Unconstrained	2595	2732	137 (89)	-2187^{**} (391)	1614	1337	-278^{**} (98)	-1380^{**} (156)
G1	7321	9333	2012^{**} (203)	-312 (234)	3765	4536	772^{**} (84)	-1180^{**} (82)
G2 Constrained	11711	13796	2085^{**} (322)	-239 (382)	6266	6683	417^{**} (155)	-685^{**} (151)
G2 Unconstrained	2362	2149	-212^{**} (54)	-2536^{**} (389)	1202	1054	-149^{**} (29)	-1251^{**} (153)
G2	6640	8109	1468^{**} (146)	-855^{**} (224)	3501	3971	471^{**} (62)	-631^{**} (88)
G3 Constrained	11847	11257	-590^{**} (290)	-2914^{**} (359)	5806	5656	-150 (119)	-1252^{**} (140)
G3 Unconstrained	2615	2503	-112 (120)	-2436^{**} (401)	1503	1832	329 (329)	-773^{**} (184)
G3	6701	6432	-269^{**} (132)	-2593^{**} (222)	3417	3597	180 (126)	-923^{**} (108)
Control	7179	9502	2324^{**} (118)		3653	4755	1102^{**} (46)	

Table 37

Effective Spreads with Fundamental Prices, mid-Quotes, and Weighted Mid-quotes

This table reports the average estimated values of the effective spreads (per quote update) using, as reference prices, mid-quotes, weighted mid-quotes (Hagstromer (2020)), and fundamental prices extracted using our methodology, across all the treated groups G_1, G_2, G_3, and the Control group. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column “Non-Pilot” (“Pilot”) presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column “Diff” is the difference estimate of variables of each group, prior and post the pilot program. “DD” is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Panel A: Non-Pilot Period = November 1-30, 2018 ; Pilot Period = September 1-30, 2018

	Effective Spreads using Midquotes				Effective Spreads using Weighted Mid-quotes				Effective Spreads using Fundamental Prices			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.012	0.038	0.026** (0.0013)	0.0268** (0.00305)	0.012	0.034	0.0224** (0.0015)	0.0229** (0.00316)	0.014	0.041	0.027** (0.00145)	0.0267** (0.00307)
G1 Unconstrained	0.107	0.089	−0.018 (0.0115)	−0.0173 (0.01187)	0.110	0.093	−0.0177 (0.01211)	−0.0172 (0.01243)	0.116	0.111	−0.0048 (0.0178)	−0.0051 (0.018)
G1	0.035	0.051	0.0157** (0.0035)	0.0164** (0.00445)	0.036	0.049	0.0139** (0.00366)	0.0144** (0.0046)	0.038	0.056	0.0183** (0.00359)	0.0179** (0.00449)
G2 Constrained	0.012	0.038	0.0255** (0.0008)	0.0262** (0.00287)	0.012	0.034	0.0216** (0.00087)	0.0221** (0.00292)	0.014	0.044	0.0292** (0.00121)	0.0288** (0.00296)
G2 Unconstrained	0.092	0.082	−0.01 (0.0085)	−0.0092 (0.00893)	0.093	0.085	−0.0082 (0.0084)	−0.0076 (0.00885)	0.095	0.093	−0.002 (0.00811)	−0.0023 (0.00855)
G2	0.048	0.057	0.0092** (0.0045)	0.01** (0.0053)	0.048	0.056	0.008** (0.00466)	0.0086** (0.00543)	0.051	0.065	0.0147** (0.00451)	0.0143** (0.00526)
G3 Constrained	0.013	0.037	0.0238** (0.0007)	0.0245** (0.00283)	0.013	0.032	0.0198** (0.00077)	0.0203** (0.00289)	0.014	0.046	0.032** (0.00087)	0.0316** (0.00284)
G3 Unconstrained	0.126	0.097	−0.0284 (0.0202)	−0.0277 (0.02043)	0.130	0.101	−0.0285 (0.02106)	−0.0279 (0.02124)	0.136	0.122	−0.0137 (0.02222)	−0.0141 (0.02238)
G3	0.043	0.052	0.0089** (0.005)	0.0097* (0.0057)	0.044	0.051	0.0071 (0.00503)	0.0077 (0.00575)	0.045	0.065	0.0203** (0.00462)	0.02** (0.00536)
Control	0.041	0.040	−0.0007 (0.0028)		0.041	0.041	−0.0005 (0.00279)		0.043	0.043	0.0003 (0.0027)	

Panel B: Non-Pilot Period = September 1-30, 2016 ; Pilot Period = November 1-30, 2016

	Effective Spreads with Midquotes				Effective Spreads with Weighted Mid-quotes				Effective Spreads with Fundamental Prices			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.013	0.043	0.0299** (0.0006)	0.0261** (0.00198)	0.013	0.038	0.0248** (0.00092)	0.021** (0.00215)	0.016	0.046	0.0299** (0.00079)	0.0257** (0.00208)
G1 Unconstrained	0.106	0.102	−0.0046 (0.01)	−0.0083 (0.0102)	0.109	0.103	−0.0057 (0.0099)	−0.0095 (0.01009)	0.107	0.104	−0.0036 (0.00932)	−0.0078 (0.00952)
G1	0.028	0.052	0.0239** (0.0024)	0.0201** (0.00307)	0.029	0.049	0.0201** (0.00269)	0.0163** (0.00332)	0.031	0.055	0.0245** (0.00254)	0.0203** (0.00319)
G2 Constrained	0.013	0.043	0.0297** (0.0007)	0.0259** (0.002)	0.013	0.038	0.0248** (0.001)	0.021** (0.00219)	0.016	0.048	0.0316** (0.00089)	0.0274** (0.00212)
G2 Unconstrained	0.094	0.098	0.0039 (0.0096)	0.0001 (0.00982)	0.097	0.101	0.0035 (0.00916)	−0.0003 (0.00936)	0.096	0.105	0.0099 (0.00901)	0.0057 (0.00921)
G2	0.030	0.053	0.0238** (0.0026)	0.02** (0.00318)	0.031	0.051	0.0205** (0.00284)	0.0168** (0.00344)	0.032	0.059	0.027** (0.0027)	0.0228** (0.00331)
G3 Constrained	0.014	0.041	0.0272** (0.0009)	0.0234** (0.00207)	0.014	0.037	0.0229** (0.00136)	0.0191** (0.00237)	0.017	0.052	0.035** (0.0017)	0.0308** (0.00256)
G3 Unconstrained	0.093	0.120	0.0271 (0.0272)	0.0233 (0.02726)	0.097	0.122	0.0249 (0.02667)	0.0211 (0.02674)	0.093	0.128	0.0344 (0.02339)	0.0302 (0.02347)
G3	0.028	0.049	0.0216** (0.0022)	0.0178** (0.00292)	0.029	0.047	0.0184** (0.00249)	0.0146** (0.00316)	0.030	0.060	0.0299** (0.0023)	0.0257** (0.00299)
Control	0.031	0.035	0.0038** (0.0019)		0.032	0.036	0.0038** (0.00195)		0.034	0.038	0.0042** (0.00192)	

Table 38

Proportion of Price Discovery through Market Orders, Limit Orders and New Information

This table reports the average estimated proportions of price discovery through market orders, limit orders and new information, across all the treated groups G_1, G_2, G_3, and the Control group. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column "Non-Pilot" ("Pilot") presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column "Diff" is the difference estimate of variables of each group, prior and post the pilot program. "DD" is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Panel A: Non-Pilot Period = November 1-30, 2018 ; Pilot Period = September 1-30, 2018

	Price Discovery through Market Orders				Price Discovery through Limit Orders				Price Discovery through New Information			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.044	0.017	−0.027** (0.004)	−0.024** (0.004)	0.245	0.227	−0.015 (0.013)	0.016 (0.017)	0.715	0.756	0.042** (0.016)	0.008 (0.017)
G1 Unconstrained	0.059	0.038	−0.02** (0.003)	−0.017** (0.003)	0.248	0.274	0.026** (0.011)	0.057** (0.015)	0.699	0.693	−0.006 (0.014)	−0.04** (0.014)
G1	0.054	0.025	−0.029** (0.002)	−0.025** (0.002)	0.237	0.255	0.02** (0.009)	0.05** (0.01)	0.714	0.723	0.009 (0.009)	−0.025** (0.009)
G2 Constrained	0.046	0.011	−0.035** (0.004)	−0.031** (0.004)	0.231	0.286	0.057** (0.016)	0.088** (0.017)	0.727	0.705	−0.022** (0.009)	−0.056** (0.017)
G2 Unconstrained	0.06	0.042	−0.017** (0.002)	−0.014** (0.003)	0.251	0.252	0.002** (0.012)	0.033** (0.013)	0.695	0.711	0.015 (0.012)	−0.019 (0.013)
G2	0.055	0.028	−0.027** (0.0018)	−0.023** (0.0021)	0.231	0.25	0.02** (0.008)	0.0502** (0.01)	0.718	0.725	0.007 (0.022)	−0.0269** (0.0095)
G3 Constrained	0.042	0.008	−0.034** (0.0027)	−0.03** (0.0029)	0.219	0.196	−0.021 (0.0141)	0.01 (0.015)	0.742	0.797	0.0544** (0.014)	0.0204 (0.014)
G3 Unconstrained	0.062	0.049	−0.013** (0.0044)	−0.01** (0.004)	0.242	0.244	0.002 (0.0147)	0.033** (0.015)	0.702	0.713	0.011 (0.015)	−0.023** (0.0155)
G3	0.056	0.029	−0.026** (0.002)	−0.023** (0.002)	0.229	0.225	−0.003 (0.005)	0.028** (0.009)	0.72	0.749	0.029** (0.008)	−0.0046 (0.006)
Control	0.056	0.052	−0.003** (0.001)		0.227	0.195	−0.03** (0.005)		0.722	0.756	0.034** (0.005)	

Panel B: Non-Pilot Period = September 1-30, 2016 ; Pilot Period = November 1-30, 2016

	Price Discovery through Market Orders				Price Discovery through Limit Orders				Price Discovery through New Information			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.049	0.022	−0.027** (0.004)	−0.023** (0.004)	0.199	0.278	0.0783** (0.012)	0.0891** (0.0127)	0.751	0.7	−0.0513** (0.01)	−0.066** (0.012)
G1 Unconstrained	0.066	0.044	−0.0218** (0.003)	−0.0178** (0.004)	0.271	0.273	0.0026 (0.038)	0.0134 (0.015)	0.663	0.683	0.019 (0.017)	0.005 (0.03)
G1	0.058	0.033	−0.024** (0.002)	−0.02** (0.003)	0.224	0.268	0.044** (0.008)	0.055** (0.009)	0.718	0.698	−0.02** (0.006)	−0.035** (0.008)
G2 Constrained	0.048	0.024	−0.024** (0.003)	−0.02** (0.003)	0.223	0.275	0.052** (0.013)	0.063** (0.014)	0.729	0.701	−0.028** (0.011)	−0.042** (0.012)
G2 Unconstrained	0.061	0.049	−0.012** (0.004)	−0.008** (0.004)	0.259	0.289	0.03** (0.017)	0.0408** (0.018)	0.68	0.662	−0.0184 (0.014)	−0.0332** (0.017)
G2	0.055	0.037	−0.019** (0.002)	−0.014** (0.002)	0.229	0.263	0.034** (0.008)	0.04** (0.009)	0.715	0.7	−0.015** (0.006)	−0.03** (0.008)
G3 Constrained	0.051	0.021	−0.03** (0.004)	−0.026** (0.0046)	0.242	0.236	−0.006 (0.01)	0.0051 (0.0163)	0.707	0.743	0.036* (0.021)	0.021 (0.026)
G3 Unconstrained	0.068	0.067	−0.01 (0.015)	0.003** (0.0103)	0.277	0.315	0.038 (0.029)	0.049* (0.029)	0.655	0.618	−0.037* (0.028)	−0.052* (0.03)
G3	0.059	0.038	−0.021** (0.002)	−0.017** (0.0023)	0.25	0.243	−0.007 (0.006)	0.003 (0.0092)	0.691	0.72	0.028** (0.01)	0.014 (0.013)
Control	0.058	0.054	−0.004** (0.001)		0.241	0.23	−0.011** (0.005)		0.702	0.716	0.015** (0.005)	

Table 39

Speed of Price Discovery with Fundamental Prices

This table reports the average estimated delay measures $\{D_1, D_2, D_3\}$ (Chordia and Swaminathan (2000) and Hou and Moskowitz (2005)) using fundamental prices (Panel A) extracted from our methodology and quote mid-points (Panel B), across all the treated groups G_1, G_2, G_3, and the Control group. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The column "Non-Pilot" ("Pilot") presents the estimated mean values of variables of each group during the non-pilot (pilot) regime. The column "Diff" is the difference estimate of variables of each group, prior and post the pilot program. "DD" is the difference-in-differences estimate of variables in treated groups with respect to the variables in control group. Standard errors are in parenthesis and ** denotes significance at the 5% level.

Panel A: Delay Measures with Fundamental Prices ; Non-Pilot Period = November 1-30, 2018 ; Pilot Period = September 1-30, 2018

	Delay Measure D1				Delay Measure D2				Delay Measure D3			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.447	0.729	0.2822** (0.02779)	0.1741** (0.02959)	0.778	0.896	0.1179** (0.01224)	0.0699** (0.01294)	0.774	0.894	0.1203** (0.01227)	0.0714** (0.01297)
G1 Unconstrained	0.668	0.726	0.0575** (0.01947)	−0.0507** (0.02197)	0.872	0.890	0.0183** (0.00905)	−0.0297** (0.00997)	0.868	0.887	0.0185** (0.00915)	−0.0305** (0.01008)
G1	0.585	0.725	0.1401** (0.01553)	0.0319* (0.01857)	0.835	0.895	0.0599** (0.00656)	0.012 (0.00778)	0.831	0.892	0.0611** (0.00659)	0.0121 (0.00782)
G2 Constrained	0.463	0.739	0.2757** (0.03085)	0.1675** (0.03248)	0.780	0.898	0.1187** (0.01292)	0.0707** (0.01358)	0.776	0.897	0.1215** (0.01304)	0.0726** (0.01371)
G2 Unconstrained	0.644	0.710	0.0651** (0.02262)	−0.0431* (0.02481)	0.860	0.889	0.0286** (0.0081)	−0.0194** (0.00912)	0.857	0.886	0.0297** (0.00818)	−0.0192** (0.0092)
G2	0.596	0.715	0.1183** (0.01955)	0.0101 (0.02204)	0.838	0.895	0.0564** (0.00659)	0.0084 (0.0078)	0.835	0.893	0.0577** (0.00665)	0.0088 (0.00788)
G3 Constrained	0.512	0.749	0.2372** (0.02903)	0.129** (0.03076)	0.801	0.903	0.1024** (0.01253)	0.0544** (0.0132)	0.797	0.902	0.105** (0.01258)	0.0561** (0.01327)
G3 Unconstrained	0.757	0.821	0.064** (0.02002)	−0.0441** (0.02246)	0.873	0.898	0.0253** (0.00919)	−0.0227** (0.0101)	0.869	0.895	0.026** (0.0093)	−0.023** (0.01021)
G3	0.601	0.740	0.1391** (0.01567)	0.0309* (0.01869)	0.840	0.900	0.0599** (0.00657)	0.0119 (0.00778)	0.837	0.898	0.0613** (0.00662)	0.0124 (0.00785)
Control	0.594	0.702	0.1082** (0.01017)		0.838	0.886	0.048** (0.00418)		0.834	0.883	0.0489** (0.00421)	

Panel B: Delay Measures with Fundamental Prices ; Non-Pilot Period = September 1-30, 2016 ; Pilot Period = November 1-30, 2016

	Delay Measure D1				Delay Measure D2				Delay Measure D3			
Group	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD	Non-Pilot	Pilot	Diff	DD
G1 Constrained	0.542	0.682	0.1404** (0.02363)	0.112** (0.02523)	0.821	0.878	0.0571** (0.00935)	0.0475** (0.00979)	0.817	0.877	0.0602** (0.00942)	0.0482** (0.00987)
G1 Unconstrained	0.775	0.751	−0.024** (0.01841)	−0.0524** (0.02042)	0.931	0.922	−0.0098** (0.00607)	−0.0194** (0.00673)	0.927	0.919	−0.0085 (0.00619)	−0.0206** (0.00685)
G1	0.654	0.707	0.0536** (0.01369)	0.0252** (0.01629)	0.865	0.887	0.022** (0.00554)	0.0124** (0.00626)	0.861	0.885	0.0245** (0.00559)	0.0125* (0.00632)
G2 Constrained	0.523	0.684	0.1619** (0.02267)	0.1335** (0.02433)	0.813	0.878	0.0652** (0.00902)	0.0556** (0.00948)	0.809	0.877	0.0686** (0.00913)	0.0565** (0.0096)
G2 Unconstrained	0.774	0.764	−0.0103 (0.01993)	−0.0386 (0.0218)	0.855	0.927	0.0721** (0.00637)	0.0625** (0.00701)	0.850	0.925	0.0749** (0.00644)	0.0628** (0.00709)
G2	0.645	0.712	0.0673** (0.01307)	0.0389** (0.01577)	0.862	0.890	0.028** (0.00524)	0.0184** (0.006)	0.857	0.888	0.0307** (0.00529)	0.0186** (0.00606)
G3 Constrained	0.541	0.675	0.1337** (0.02906)	0.1053** (0.03037)	0.821	0.876	0.0546** (0.01112)	0.045** (0.0115)	0.817	0.874	0.0576** (0.01124)	0.0456** (0.01162)
G3 Unconstrained	0.771	0.766	−0.0049 (0.02806)	−0.0333 (0.02941)	0.929	0.928	−0.0014 (0.00839)	−0.011** (0.00889)	0.926	0.925	−0.0005 (0.00853)	−0.0126 (0.00902)
G3	0.648	0.725	0.0769** (0.01323)	0.0485** (0.01591)	0.864	0.895	0.0309** (0.00528)	0.0213** (0.00604)	0.860	0.893	0.0335** (0.00533)	0.0214** (0.00609)
Control	0.669	0.697	0.0284** (0.00883)		0.894	0.903	0.0096** (0.00292)		0.890	0.902	0.0121** (0.00295)	

B. Internet Appendix

Table 40
Market-maker Profits and Depths

This table reports the average estimated total bid-ask depth at the best bid and best ask quotes conditional on market maker profits. For every stock on each day, we split the transactions into two depending on their profitability. "High_{mmp}" ("Low_{mmp}") represents the subsample of transactions that have above (below) the median-market maker profits. "Average depth" column in Panel A represents the average of the total depth (sum of the best bid and best ask depths) on the "High_{mmp}" and "Low_{mmp}" subsamples across all stocks in various groups, such as G1, G2, and G3. "High_{qspr}" ("Low_{qspr}") represents the subsample of transactions that have above (below) the quoted spreads for each stock, each trading day. "Average depth" column in Panel B is the average of the total depth on the "High_{qspr}" and "Low_{qspr}" subsamples. Constrained (unconstrained) stocks are those whose quoted bid-ask spreads were lower (higher) than 5 cents prior to the TSP. The sample period is from September 1, 2018 to September 30, 2018.

	Panel A		Panel B	
Group	High/Low *mmp*	Average Depth	High/Low Quoted Spread	Average Depth
G1 Constrained	High_{mmp}	7638.53	High_{qspr}	6580.05
	Low_{mmp}	6232.22	Low_{qspr}	7100.35
G1 Unconstrained	High_{mmp}	815.40	High_{qspr}	736.97
	Low_{mmp}	837.44	Low_{qspr}	875.59
G1	High_{mmp}	3287.70	High_{qspr}	2860.56
	Low_{mmp}	2876.58	Low_{qspr}	3199.60
G2 Constrained	High_{mmp}	5697.19	High_{qspr}	5344.65
	Low_{mmp}	5401.25	Low_{qspr}	5659.71
G2 Unconstrained	High_{mmp}	799.08	High_{qspr}	740.87
	Low_{mmp}	915.07	Low_{qspr}	912.98
G2	High_{mmp}	3153.31	High_{qspr}	3026.46
	Low_{mmp}	4035.52	Low_{qspr}	4113.92
G3 Constrained	High_{mmp}	8819.17	High_{qspr}	8755.27
	Low_{mmp}	7653.87	Low_{qspr}	7656.41
G3 Unconstrained	High_{mmp}	1205.78	High_{qspr}	1226.55
	Low_{mmp}	1253.87	Low_{qspr}	1214.73
G3	High_{mmp}	3736.35	High_{qspr}	3724.87
	Low_{mmp}	3426.48	Low_{qspr}	3417.88