Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Arabind Swain Date

Detectability, Interpretability, and the Limits of Machine Learning in High-Dimensional Physical Systems

By

Arabind Swain
Doctor of Philosophy

Physics

Ilya Nemenman, Ph.D.
Advisor

Andrea J. Liu, Ph.D.
Committee Member

Daniel M. Sussman, Ph.D.
Committee Member

Eric R. Weeks, Ph.D.
Committee Member

Daniel B. Weissman, Ph.D.
Committee Member

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Detectability, Interpretability, and the Limits of Machine Learning in High-Dimensional Physical Systems

By

Arabind Swain
Int. M. Sc., NISER, HBNI , India, 2018

Advisor: Ilya Nemenman, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2025

Abstract

Detectability, Interpretability, and the Limits of Machine Learning in High-Dimensional Physical Systems
By Arabind Swain

In recent years, large-dimensional datasets have become increasingly common in physics, arising from simulations and experiments that capture complex systems across space and time. These datasets offer new opportunities for discovery but also pose significant challenges in separating meaningful physical structure from irrelevant correlations and statistical noise. This dissertation investigates the use of machine learning (ML) methods to uncover underlying physical structure in high-dimensional systems, focusing on two central challenges: interpreting ML predictions in complex glassy systems, and developing a theoretical foundation for understanding statistical significance of correlations between large datasets when conducting individual marginal covariance, joint covariance and cross-covariance analysis. In the context of glassy dynamics, where traditional approaches struggle due to the absence of clear structural order, ML classifiers such as Support Vector Machines (SVMs) have been shown to accurately predict local rearrangements of particles. However, using simple toy models and simulations, this work demonstrates that commonly used indicators—such as high classification accuracy, apparent Arrhenius scaling, or distance to hyperplanes—are not sufficient to guarantee that the ML model has captured meaningful physical quantities which is the size of the energy barriers in this case. This raises important questions about the inverse problem: under what conditions can interpretable physics be extracted from statistical learning models? To address broader issues of signal detection in high-dimensional data, the dissertation extends the well-known Marchenko-Pastur (MP) distribution from covariance to cross-covariance matrices. An exact analytical expression is derived for the distribution of singular values arising purely from noise-noise correlations, providing a null model for detecting shared structure between two large datasets. Furthermore, the work establishes a BBP-type (Baik–Ben Arous–Péché) detectability phase transition for cross-covariance and joint-covariance matrices, identifying critical thresholds for when rank-1 signals become distinguishable from noise, and showing that joint and cross-covariance methods can detect weaker signals—or do so with fewer samples—than individual marginal covariance based analysis. Altogether, this dissertation provides both conceptual insight and analytical tools for understanding when ML models truly learn the physics of the system, and how noise, dimensionality, and sample size fundamentally constrain that process.

Detectability, Interpretability, and the Limits of Machine Learning in High-Dimensional Physical Systems

By

Arabind Swain
Int. M. Sc., NISER, HBNI , India, 2018

Advisor: Ilya Nemenman, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2025

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Ilya Nemenman, for everything he has done for me throughout this journey. His faith in me, patience, and guidance have meant a great deal. He gave me the freedom to explore and the support to grow, while consistently teaching me how to ask simple questions that lead to deep scientific insight. I am truly grateful for his mentorship.

I would also like to thank my committee members—Drs. Andrea Liu, Daniel Sussman, Daniel Weissman, and Eric Weeks—for their valuable feedback, thoughtful discussions, and support throughout my research. Each of you has contributed in different and meaningful ways, and I appreciate the time and insight you've shared with me.

A special thank you goes to Sean Ridout. Working with you has been one of the most enjoyable parts of my PhD. Whether collaborating on science or talking about food and recipes, your ideas and support have shaped every chapter of this thesis. Thank you for always being there and for being the heart of our group with all the events you organize.

K. Michael Martini, I'm grateful for the time we spent together decoding transformers and discussing everything from deep learning to business monopolies. Thank you for your contributions and companionship. Michael Pasek, thank you for your practical advice on both science and life. Satya, I truly appreciate your help navigating the infinite loops of academic logistics. Ketuna, Zehui, Ahmed, and Eslam, thank you for the many helpful discussions and support along the way.

I'd also like to thank the broader TMLS community and all my former labmates. You have each contributed to my learning in many ways and made my time here more enriching and enjoyable.

To the friends who made Emory feel like home—Luka, Jamie, Kavinda, Linnea, and others—thank you for your companionship over the years.

Thank you to Barbara for always being there, helping me with everything from administrative puzzles to day-to-day challenges, and doing so with grace and generosity.

To my housemates and friends—Tuzo, Leoncia, Taesik, Bruce, Pedro, and Daniel—thank you for your company, the Friday dinners, the spontaneous debates, and the sense of community you created. Those moments will always stay with me.

Thanks to all my friends from undergrad at HBNI: Siddharth, Aroop, Sangeet, for all the amazing discussions on a variety of topics. You people did help me maintain my sanity especially during COVID.

I'm deeply grateful to those who funded this work—the Simons Foundation and NIH (via Ilya!)—for making it possible to pursue this research. Your support gave me the rare and valuable opportunity to explore, learn, and follow my curiosity.

Finally, I am deeply grateful to my family. To my extended family—thank you for your unwavering support and encouragement over the years. To my sister, thank you for your joy and optimism, which have always lifted my spirits. And to my parents and grandparents—thank you for your constant love, sacrifice, and belief in me.

Contents

List of Figures

Chapter 1

Introduction

The 2024 Nobel Prize awards were a recognition of the increasingly intertwined nature of the hard sciences and artificial intelligence (AI). The Physics Prize acknowledged the contribution of the laureates, drawing on their previous experience with physics concepts from magnetic materials and Boltzmann distributions to understand and improve AI [1]. The Chemistry Prize was partly awarded for AlphaFold, which is an AI model capable of predicting complex 3D structures of proteins with high accuracy from just amino acid sequences, solving a 50 year challenge [2]. The awards are a good representation of two distinct types of topics that are of interest to physicists. On the one hand, they involve leveraging our understanding of statistical physics to try to explain the inner workings of AI and its dependence on the statistics of data, the nature of algorithms used, etc. On the other hand, they involve trying to take advantage of advances in AI, to find new ways to analyze increasingly large and complex data sets, make generative models to simulate complex systems that are too expensive to simulate using traditional methods, and sometimes guide the discovery of new physics [3, 4]. In this dissertation, the two distinct topics discussed are along similar lines: 1) Reliability of AI-based predictors of dynamics in glassy systems; 2) the conditions for detecting shared correlations between two different

large-dimensional data sets, where much of the correlation between them is purely because of statistical sampling noise.

Machine Learning (ML) is a subset of AI, where we try to enable computers to *learn* from data rather than explicitly programming an algorithm to complete a task. ML algorithms have shown great advances in the area of image recognition and classification in recent years. Using these advancements in the context of physics has led to some promising results. For example, classification algorithms, applied to the problem of phase transitions [5, 6, 7, 8, 9, 10, 11, 12] and prediction of particle rearrangement in structural glasses [13, 14, 15, 16, 17, 18, 19, 20, 21] have shown promising success in identifying the different states of matter from snapshots of the systems. When trying to learn the phases of a 2D Ising model using a classification algorithm (Support Vector Machine in this case), it was observed that the SVMs kernels that satisfied the $\mathcal{Z}_2$ symmetry of the Ising model were able to learn the phases with a limited amount of data and were correctly able to predict the transition temperature [11] because the Ising model phase transition breaks the $\mathcal{Z}_2$ symmetry. The Ising model shows an ordered to disordered phase transition as the temperature goes above a critical temperature. In the disordered state, the system has zero magnetization as the spins are randomly oriented. Thus flipping all the spins does not change the magnetization and the system maintains its full $\mathcal{Z}_2$ symmetry. However, in the ordered state the system acquires a spontaneous magnetization and spins prefer to align mostly in one direction. Thus, the system chooses either spin-up or spin-down global magnetization state. If we flip all the spins it takes us from the global spin up to global spin down magnetization state, hence breaking the $\mathcal{Z}_2$ symmetry. The spontaneous breaking of $\mathcal{Z}_2$ symmetry for the ordered low-temperature state is what describes the Ising phase transition. Thus, infusing our algorithms with structures that respect the physical properties of the underlying problem helps us to achieve high prediction accuracy with limited data, as it introduces an inductive bias in our

model. Adding right constraints for the problem can be thought of as choosing the correct network architecture or the correct kernel that helps us solving the problem. As another example, consider that objects in images typically remain recognizable even if their position is shifted. Thus Convolutional Neural Networks (CNNs) [22], which are inherently designed to capture the translational symmetry [23], perform well on image recognition tasks, .

There is another approach of looking at phase transitions using ML. Instead of trying to predict the phases accurately using an interpretable SVM satisfying the correct set of symmetries of the problem using a very small amount of data, one can also train large networks without any inherent inductive bias with a very large amount of data. Training a fully connected feedforward neural network with 1 hidden layer on a large amount of data from the 2D Ising model allowed the network to learn the phases along with the magnetization as well [5]. The ML algorithm was trained to classify the phases but it ended up learning the order parameter—the magnetization—for the phase transition without being explicitly trained to learn it.

In summary, ML can

- Learn the correct order parameter of a phase transition in some cases (though we do not understand the conditions when this is possible), allowing us to develop intuition about what drives phase transitions in systems where we lack the intuition to identify the symmetries and conserved quantities.

- Learn the phases of the system with a limited amount of data, if the correct symmetries and conserved quantities are introduced into the structure of the learning machine.

Unlike in the case of images, where we know that we need to capture translational symmetry, for most real-life systems of interest, we lack a deeper intuition about the system. The hope is that we can use ML-based methods to get this intuition. A

ML model which has captured the statistical correlations of the problem may have captured the correct physics of the problem as well. There have been instances where a ML model with high prediction accuracy for phase transition did end up learning the correct order parameters [5, 12], showing the model had learned the correct physics of the problem. But evaluating if the correct physics has been learned for a model about which we lack intuition by looking at just the prediction accuracy is difficult. Thus, studies that shed more light on the inverse problem in ML, where we are using the prediction accuracy or some other quantity extracted from a ML model to get deeper insights about the symmetries, conserved quantities, quality of our data, quality of embeddings, etc. are important. They make the application of ML-based methods more reliable and accessible to real-world problems where we are trying to use these methods to develop deeper insights or intuition for the problem.

One of the areas where ML-based methods have helped our understanding of the underlying phenomenon is in glassy dynamics. Glassy liquids have heterogeneous rearrangement dynamics: in some regions particles rearrange quickly, while others are slow. The degree of heterogeneity, *i.e.*, the range of dynamical correlations, grows as the temperature is lowered [24]. Despite this, the structural order in a glass is hard to detect, making the origin of these correlations difficult to understand [25, 26]. There are changes in pair correlation function $g(r)$, but they are very gradual with change in temperature. If relaxation involves crossing energy barriers of a fixed, temperature-independent size ΔE, then the relaxation time would obey the Arrhenius law

$$\tau \sim e^{\frac{\Delta E}{T}} \tag{1.1}$$

The change in relaxation time with temperature can be simply thought of as the degree of difficulty in crossing an estimated energy barrier ΔE as the temperature decreases. But in many real life cases, this Arrhenius relationship is often not followed

and the relevant energy barrier may even decrease as a function of temperature [27]. ML based methods connect the changes in dynamical properties to structure [18]. A Support Vector Machine (SVM) was trained to distinguish rearranging particles from non-rearranging particles. SVM tries to find the best hyperplane that separates different classes data. The distance from the separating hyperplane which was named softness, was found to be linearly related to the energy barrier ΔE [14]. This is surprising as the SVM was only trained to to differentiate rearranging and non-rearranging particles with a high accuracy. This raises a fundamental question: does high predictive performance in classification imply that the model has learned physically meaningful quantities? Specifically, can the 'energy barriers' inferred from the softness be interpreted as the true thermodynamic barriers governing particle rearrangement, or are they simply effective parameters useful for classification? There have been attempts to use hybrid ML-theory models to explain heterogeneity in supercooled liquids [28]. Thus, understanding the reliability of ML prediction of energy in glasses is of critical importance to be able to build models based on ML predictions.

This challenge is not unique to glassy systems. More broadly, it reflects a growing challenge in modern physics between the pursuit of causal, law-like descriptions for complex large dimensional systems and the increasing reliance on statistical models. Physics often tries to discover laws that govern the natural world by establishing causal relationships. Sometimes we want to establish correlations between two large-dimensional datasets rather than trying to find the relationship between a set of input features to a discrete number of categories, as happens in classification tasks related to phase transitions. When data sets have obvious symmetries and causal relationships, we can write down explicit laws. For example, Newton's laws explain motion based on forces and the resultant acceleration. Similarly, Maxwell's equations relate electric and magnetic fields to the motion of charges and to currents. However, for many domains in more modern physics domains, like modeling the brain, identifying explicit

causes of phenomena or carrying out controlled experiments to establish the cause-effect relationships is exceedingly challenging. The brain has multiple interacting time scales [29] and lacks translational or rotational invariance [30]. Thus, we are increasingly relying on ML approaches to try and discover statistical correlations in such systems.

When trying to find significant signals in a single dataset, one of the simplest ML algorithms is called Principal Component Analysis (PCA) [31]. PCA involves calculating directions (which are called principal components) which capture the maximum variance in our data. The first principal component (which is the eigenvector corresponding to the largest eigenvalue) captures the maximum variance, the second principal component the second most variance and so on. Each of the principal components are orthogonal to each other.

To determine which principal components are truly significant, it's crucial to understand what kind of eigenvalues arise purely due to statistical noise. Thus we consider firstly an N-dimensional dataset which is pure noise, consisting of T points. The data can be written as a $T \times N$ matrix where each of the elements of the matrix is drawn from a i.i.d. standard Gaussian in N-dimensional space. This gives us a N-dimensional hypersphere with T data points randomly arranged in a thin shell near the surface of the hypersphere (here $N \gg 1$, hence the norm of every point is near $\sqrt{N}$). The distribution is isotropic and has equal variance in all directions. More importantly, we can calculate exactly the probability distribution of eigenvalues for the covariance matrix corresponding to the N-dimensional hypersphere because of pure noise.

The probability distribution of eigenvalues of the covariance matrix for such pure noise is given exactly by the Marcheko-Pastur (MP) distribution [32]. The MP distribution also allows us to understand how sampling and signal detection work. In an N-dimensional space, in the presence of sampling noise, a signal can be detected only

if the eigenvalue corresponding to the signal lies outside of the bulk of eigenvalues due to noise given by MP distribution. Not having enough data points makes the detection of a signal challenging as the noise bulk (the correlations purely because of the sampling noise given by MP distribution) is much larger. With increasing sample size the noise bulk reduces making a previously undetectable signal, detectable. Thus, many problems in signal detection reduce to being able to collect enough independent samples to be able to extract the weak signal from noise. More importantly, it requires an understanding of what part of the covariance eigenvalue spectrum is due to noise and which part corresponds to the signal.

Suppose there is a strong signal in one direction. The presence of this signal deforms the hypersphere along that direction. PCA, by construction, rotates the data so that the direction of maximum variance aligns with this signal. This is achieved by computing the eigenvectors of the covariance matrix and selecting the one associated with the largest eigenvalue. The result is typically one outlier eigenvalue (due to signal), while the rest are part of the noise bulk satisfying the MP distribution.

For real world data which may have many signals, the number of such outliers indicates how many correlations in the data are not purely due to noise. Initially, it was observed that in certain systems like in speech recognition and hand written digit recognition [33], most of the eigenvalues are in the noise bulk, and the few that are not, are well separated from the bulk. A mathematical model to explain this structure was first formalized in 2001 and is called the 'spiked population model' referring to the outlier eigenvalues [34]. Later for the spiked population model given the strength of a spike relative to the noise bulk, it was shown in 2005 that one had a spike non-detectability to detectability phase transition [35]. One adds a small perturbation 'a spike' to the covariance matrix produced by noise. Depending on the strength of the spike one sees an outlier eigenvalue. When we have the outlier we call the spike detectable. Furthermore, if we calculate the overlap between the unit vector

for the spike and normalized eigenvector corresponding to the outlier eigenvalue, the overlap is found to be a 2^{nd}-order phase transition as a function of the spike strength. This phase transition is called a Baik–Ben Arous–Péché (BBP) phase transition and is valid for $N \to \infty$ with a fixed ratio of $\frac{N}{T}$. For finite-sized matrices, the phase transition becomes a crossover. More importantly, the exactly solvable results gives us a good understanding of when one can detect a signal in PCA. We can also exactly evaluate, given the size of the outlier eigenvalue as compared to the noise bulk, how good our reconstruction of the signal associated with the outlier eigenvalue is. The understanding also allows us to develop methods for PCA where we can reduce the effect of sampling noise on the signal [36, 37].

Now suppose we want to find correlations between two large dimensional data sets ($\mathbf{X}$ and $\mathbf{Y}$) and we assume they have a shared signal. It means increase in the variance in the signal direction in $\mathbf{X}$ leads to a proportional increase in the variance in signal direction in $\mathbf{Y}$. Then given T time points, we have N_X and N_Y dimensional hyper-spheres for $\mathbf{X}$ and $\mathbf{Y}$ respectively. Given both have a shared signal, both the hyperspheres will have correlated deformations.

One of the ways to detect these deformations is to do a PCA on $\mathbf{X}$ and $\mathbf{Y}$ and then regress the eigenvectors corresponding to the largest eigenvalue onto each other. This method is called Principal Component Regression (PCR) [38]. However, when we try to apply PCR to a dataset with an unknown number of shared correlations, it becomes more challenging. It is difficult to know *a priori* by looking at the PCA of $\mathbf{X}$ and $\mathbf{Y}$ respectively, how many outliers are really correlated between both data sets and which outliers from $\mathbf{X}$ and $\mathbf{Y}$ should be chosen for regression. The standard way people deal with the problem is to have multiple different regressions for a combination of outliers from both $\mathbf{X}$ and $\mathbf{Y}$ and select the models with the highest prediction accuracy.

Here we use example of data related to finance (where the methods to correlate two different types of assets are widely used) to explain how PCA based methods work in

real world datasets where we are interested in finding correlations between 2 different data sets. In this example, we are using data about the prices of stocks in the past to predict their future price in two different stock markets (say US and China) and trying to find the shared signal between them. PCA based approaches are known to give contradictory results when trying to find shared correlations between two different stock markets, even when models have similar predictive accuracy [39, 40].This is partly due to the presence of multiple timescales in finance [41, 42, 43, 44]. These timescales arise from different kinds of periodicity and correlations that are part of the financial data. There are sectoral correlations, where tech stocks like Google, Meta, Apple have correlated price movements [45]. Similarly, because of options expiring on the 3^{rd} Friday of every month, there are systematic market movements in the 3^{rd} week of each month, as institutional investors often adjust large positions simultaneously near expiry. These create higher-than-average realized volatility and correlations [46, 47].

These periodic correlations affect how well PCA-based methods can detect signals. For instance, if we start with a $T \times N$ dataset with $N > T$, its rank is T. Repeating the dataset 12 times yields a $12T \times N$ matrix, but the rank remains T; hence, no new signal becomes detectable. This means taking a lot of correlated measurements does not help us get enough sampling to detect signal. Depending on the time scale we are interested in (these systems have multiple timescales), some of the periodicities will dominate. But we will have other periodicities which though not dominant introduce correlations, thus reducing the number of independent measurements we have in our data. This makes the detection of a weak signal very difficult because of insufficient sampling. Because of sectoral correlations, all the stocks in a particular sector can be coarse grained into a single effective latent feature [48, 49, 50]. This reduces the dimensionality of the problem and allows us to learn some of the statistics with the reduced amount of independent time points that we have available [51, 52]. Such

models, where we represent the dynamics of our systems in terms of coarse-grained latent features, are called structure factor models [53, 54, 55, 56]. But then depending on the choice of latent features and how we decide to deal with the different timescales, the interpretation of latent features and the number of latent features in our model can vary. Thus, the choice of different principal components and the resultant latent features can lead to subtle differences in the interpretation of the latent features and the economic meaning of the principal component coefficient weights, leading to contradictory conclusions when evaluated in real-world scenarios.

The problems related to PCR mean that more direct methods of obtaining cross-correlations such as Partial Least Square (PLS) [57] are sometimes used in real world problems. PLS uses the cross-covariance ($\frac{\mathbf{X}^T\mathbf{Y}}{T}$) instead of the covariance ($\frac{\mathbf{X}^T\mathbf{X}}{T}$), as is the case with PCA. PLS tries to find shared axes in both $\mathbf{X}$-space and $\mathbf{Y}$-space, to maximize the joint variability between the two datasets. But unlike for PCA, the calculations analogous to Marchenko Pastur distribution, which give the probability density of eigenvalues purely from noise-noise cross-correlation still does not exist. Because of which in the case of cross-covariance one does not know which of the observed correlations are significant. This makes a deeper understanding of cross-correlation based methods difficult.

In this dissertation, I use a combination of toy models, synthetic data, simulations and exactly solvable analytical methods to answer questions relating to the limits of of ML in high-dimensional physical systems. I summarize the key questions addressed and the primary finding of each of the Chapters of this Dissertation below.

In Chapter 2, we study the inverse problem of using ML methods to explain the physical properties of systems in the context of glassy materials. The complexity of glasses makes it challenging to explain their dynamics. ML has emerged as a promising pathway for understanding glassy dynamics by linking their structural features to rearrangement dynamics. We would like to understand the reliability of

SVM predictions of energy in the context of glasses. By numerical analysis of toy models, we explore under which conditions it is possible to infer the energy barrier to rearrangements from the distance to the separating hyperplane. We observe that such successful inference is possible only under very restricted conditions. Typical tests, such as the apparent Arrhenius dependence of the probability of rearrangement on the inferred energy and the temperature, or high cross-validation accuracy, do not guarantee success.

In Chapter 3, we extend the Marchenko Pastur result for the distribution of eigenvalues of empirical sample covariance matrices to singular values of empirical cross-covariances. For two large matrices $\mathbf{X}$ and $\mathbf{Y}$ with Gaussian i.i.d. entries and dimensions $T \times N_X$ and $T \times N_Y$, respectively, we derive the probability distribution of the singular values of $\mathbf{X}^T\mathbf{Y}$ in different parameter regimes. Our results will help to establish statistical significance of cross-correlations in all the parameter regimes by giving the exact analytic solution for the distribution of singular values purely on account of noise-noise cross-correlations in cross-covariance matrix.

In Chapter 4, we derive BBP type signal non-detectability to detectability phase transition for cross-correlations. One can find cross-correlations by evaluating individual PCAs and regressing them, by calculating the cross covariance and detecting the outlier eigenvalue or by concatenating $\mathbf{X}$ and $\mathbf{Y}$ and calculating the joint covariance of the concatenated quantity. We showed that we can detect a weaker signal, or detect the same signal with a lower amount of sampling, by using joint covariance or cross covariance instead of using two individual PCAs.

In Chapter 5, we summarize the results of previous chapters discuss how the calculations in Chapter 3 and Chapter 4 using RMT based methods be potentially applied to machine learning and data science.

Chapter 2

Machine learning that predicts well may not learn the correct physical descriptions of glassy systems

2.1 Introduction

[1]In recent years, there have been a number of attempts to use Machine Learning (ML) techniques to better understand physical phenomena [3]. One of the areas that has shown considerable promise is the use of classification algorithms to differentiate between different states of a physical system [5, 6, 7, 8, 9, 10, 13, 16, 17, 12, 18, 13, 19, 20, 21, 11, 14, 15, 59]. In some of these cases, ML techniques manage to go beyond classification, extracting physically interpretable low-dimensional descriptions, such as order parameters [5, 12, 11], topological invariants [60], or the energy

[1]This chapter presents the paper [58] "Swain, Arabind, Sean Alexander Ridout, and Ilya Nemenman. "Machine learning that predicts well may not learn the correct physical descriptions of glassy systems." Physical Review Research 6, no. 3 (2024): 033091.". The work was conducted in collaboration with Drs. Sean Alexander Ridout and Ilya Nemenman. I performed all simulations, conducted all analyses, and led writing of the manuscript. Dr. Nemenman conceived the model and led the research, while Dr. Ridout contributed to discussions regarding the procedures and analyses and had important inputs on the observables that were calculated. All authors participated in writing and reviewed the final manuscript.

barriers that determine the rate of rearrangements in a glassy liquid [13, 14, 15]. In other words, sometimes ML methods build accurate *physical* models of the studied system, even when the relevant variables describing the physics are not explicitly in the dataset. Traditionally, finding such low-dimensional, relevant descriptions requires specialized knowledge, *e. g.,* of conservation laws. Such successes without this specialized knowledge show the potential of ML techniques to discover new physics with minimal guidance by scientists. However, very little is known about when an ML method, trained to predict a certain aspect of the behavior of a physical system, constructs an accurate physical model, rather than a purely statistical one.

We will answer this question in a simplified, tractable model of the important physical problem of predicting rearrangements of glassy liquids using structural data [13, 14, 15]. Glassy liquids have heterogeneous rearrangement dynamics: in some regions particles rearrange quickly, while others are slow. The degree of heterogeneity, as well as length scales characterizing the range of dynamical correlations, grows as the temperature is lowered [24, 61, 62, 63]. Despite this, the structural order in a glass is hard to detect, making the origin of these correlations difficult to understand [25, 26]. In recent years, there has been considerable progress in linking the dynamics of glassy liquids to their structure using ML. Support Vector Machines (SVMs) [18, 13, 14, 19, 15, 59, 64], Neural Networks [20, 65, 66, 67, 68, 69], and linear regression [21, 70, 69] have been trained on large data sets generated through simulations. Local structural features were used to predict whether a particle rearranges in a specific time period Δt. All of these methods were shown to predict rearrangements with high accuracy. The classifiers could also predict rearrangements when applied to data from previously unseen temperatures. Thus, the classifiers learn local structural predictors of dynamics that generalize across temperatures. In the linear SVM case, the distance to the separating hyperplane, named softness S [13], has a simple interpretation as a local energy barrier to rearrangement $\Delta E(S)$. This is because the

probability for a particle to rearrange in some unit time Δt given S was numerically found to obey the Arrhenius law,

$$P(R|S) \propto \exp\left[\Sigma(S) - \Delta E(S)/T\right], \tag{2.1}$$

which is precisely the probability of rearrangement for a process that requires crossing a single energy barrier $\Delta E(S)$. In particular, $\Sigma(S)$ and $\Delta E(S)$ were found to be linear in S. Therefore, this simple linear classifier seems to have learned a physical description of the system, without being instructed to infer it.

Recent work has begun to use this learned dynamical description as the basis for simplified dynamical models of supercooled liquids and amorphous solids, using the inferred $\Delta E(S)$ and $\Sigma(S)$ as parameters in these models [71, 72, 73, 28]. However, there has been no explicit study showing if the success in making predictions signifies that the inferred physical description agrees with the true one. Understanding when the two match is the goal of this work. Specifically, assuming that there exists an underlying structural variable S such that Eq. (2.1) holds in a glassy liquid, we will explore when an SVM can learn the correct variable S. We focus on SVMs [74] (and, more specifically, linear SVMs) in our study because SVMs are interpretable, their performance compares well to other methods for this system, and the interpretation of statistical properties of the classifier (softness) as a physical quantity (linearly proportional to the Arrhenius energy barrier) was made for SVMs, and not other ML methods.

We devise a toy model where a true energy barrier, $\Delta E(\vec{x})$ describes the probability for a given configuration $\vec{x}$ to rearrange. We show numerically how the choice of structural variables given to the SVM affects the prediction accuracy and the ability of the trained model to predict the true energy barrier. We show that, if the SVM is given as the input only those features that contribute linearly to $\Delta E(\vec{x})$, then the

inferred softness (distance to the separating hyperplane), indeed, predicts the true $\Delta E(\vec{x})$. This is true even when the SVM is only trained to predict rearrangements, rather than $\Delta E(\vec{x})$ explicitly. However, we also show that, with a finite amount of training data, the energy barrier estimated through the softness inferred by the SVM can be strongly biased. Surprisingly, this is true even if the quality of prediction, measured by common statistical tests, such as cross-validation, is high. Thus, for our simple model, SVM does not necessarily learn the correct energy barriers, even when it seems that it does or should. Since, in real systems, structural variables determining the energy barrier are typically unknown, one usually provides ML algorithm with a large set of features, with only some of the features that can act as predictors of the rearrangement probability [13]. One then hopes that the machine distinguishes the features that directly contribute to the barrier height from those that are correlated with them, and from those that are irrelevant for the prediction. In this scenario, we show that the SVM becomes confused, so that its softness cannot be interpreted as the barrier in the presence of additional features correlated with components of the true energy function. Although the models we study are simple toy models, the fact that SVMs can fail to infer the true energy barriers even in a simple model suggests that their applications in real physical systems should be more carefully tested. Finally, we demonstrate methods to diagnose these problems and to fix them by systematic pruning of the structural features used to predict rearrangements.

2.2 Model and Simulations

We study a toy model, which still contains many of the features relevant for our analysis. In the previous work, Ref. [13], an SVM was used to identify a linear combination $S_i = S(\vec{x}_i) = \sum_{j=1}^{n} \alpha^j x_i^j$ of structural features $\vec{x}_i$, associated with a specific particle i, such that the probability of rearrangement for the particle is as in

Eq. (2.1). Specifically, in order to reproduce Eq. (2.1), we require a model where (i) each particle i is described by n structural variables $\vec{x}_i = \{x_i^1, x_i^2, \ldots, x_i^n\}$, which vary among the particles; (ii) each particle has a rearrangement energy barrier $\Delta E(\vec{x}_i)$, and (iii) the probability to rearrange depends on T and $\Delta E(\vec{x}_i)$ with a law that tends to the Arrhenius law for low temperatures. Additionally we investigated data from simulations from [73] and found that the distributions of the predictors and the inferred energy were largely gaussian (see Section 2.3.4). The simplest model with these properties is one where all n dimensions of $\vec{x}_i$ are drawn independently at random, and the true energy barrier is a linear function of the n-dimensional $\vec{x}_i$. Thus, for each particle $i = 1, \ldots, N$, we generate an n-dimensional coordinate vector $\vec{x}_i = \{x_i^1, x_i^2, \ldots, x_i^n\}$ as

$$x_i^j \sim \mathcal{N}(0, (\sigma^j)^2) \quad \forall \quad j = 1, \ldots, n \quad \text{and} \quad i = 1, \ldots, N. \tag{2.2}$$

We then assume that the energy barrier to rearrangement is a linear combination of these coordinates

$$\Delta E(\vec{x}_i) = \sum_{j=1}^{n} \alpha^j x_i^j. \tag{2.3}$$

This results in a Gaussian distribution of ΔE, consistent with the Gaussian distribution of S in supercooled liquids [13].

Finally, for each configuration, we determine whether or not it rearranges by sampling a binary random variable $R_i = \pm 1$ (where ± 1 stands for presence/absence of a rearrangement) from

$$P(R_i = 1 \mid \vec{x}_i) = \frac{e^{-\beta \Delta E(\vec{x}_i)}}{1 + e^{-\beta \Delta E(\vec{x}_i)}}, \tag{2.4}$$

which reduces to the Arrhenius form at low T while remaining below 1 at high T.

We then train a linear SVM [75] to predict R_i from $\vec{x}_i$, for all $i = 1, \ldots, N$. As is

the common practice, for the training, we standardize all xs to have zero mean and unit variance. Thus, drawing x^j from $\mathcal{N}(0, \sigma^2)$ is equivalent to drawing them from $\mathcal{N}(0, 1)$ and absorbing the standard deviation into the definition of α, which is what we do. Further, the results shown below are all evaluated at $\alpha^j = 1.2$. We verified separately that this choice does not change qualitative results from Sec. 2.3.1 and Sec. 2.3.2 (not shown, but also see Section 2.3.5 for some discussion).

After training the SVM, we define the *softness* S_i for state $\vec{x}_i$ as the signed distance to the separating hyperplane, as in previous work [13]. We then want to estimate the probability of rearrangement $P(R|S)$, to see if the softness defines it well. In Ref. [13], this probability was estimated as the frequency of rearrangements in a certain small bin of S. Instead, to remove artifacts caused by the finite bin width, we estimate $P(R|S)$ using a logistic regression model.

In a glass, energy barriers should be strictly positive, and the probability for a typical particle to rearrange is tiny. To remove biases in the inference, one typically balances the dataset used for training to have similar numbers of particles that do and do not rearrange [13]. In our model, Eq. (2.3), we achieve this balance by explicitly centering ΔE at zero. We checked numerically that this choice does not qualitatively affect the ability of the SVM to correctly predict the energy (Section 2.3.5).

A large number of structural features are used to train an SVM to predict glassy dynamics [13]. These features, however, are correlated. To observe the effect of these correlations on the ability of the SVM to predict the correct energy, for simulations in Sec. 2.3.3, we give as input to the SVM a $2n$-dimensional coordinate vector $(\vec{x}_i, \vec{z}_i)$, where $z_i^j = (x_i^j)^2 \sum_{j_1} x_i^{j_1}$, and all x's remain uncorrelated, as before. There is nothing particular about this choice of additional variables z^j correlated with x^j, besides that we wanted to preserve the same symmetry under parity (even order contributions would average out for symmetric xs). Further, we wanted these spurious extra dimensions to be non-linearly correlated with xs, modeling nonlinear correlations be-

tween values of different radial and angular density functions in [13]. We believe that our conclusions on the ability of the SVM to predict the correct energy will be qualitatively the same for other choices of spurious correlated variables obeying these conditions, and we have checked a few other cases (Section 2.3.7). We then train the SVM to predict rearrangements from this expanded set of coordinates and evaluate the effect of the correlated input variables on the quality of the model the SVM builds.

2.3 Results

2.3.1 Linear SVM can learn the true energy barrier in the infinite data limit

First, we test whether or not the softness S, inferred by the SVM from a very large sample, is a good approximation for $\Delta E(\vec{x})$ from Eq. (2.3). We use $N = 10^6$ training samples with 5×10^5 examples each of rearranging and non-rearranging configurations to train the SVM. The distribution of energies in the training sample is symmetric. We have 14 independently sampled input dimensions, with $\alpha^j = 1.2$ for $j = 1, \ldots, 10$ and $\alpha^j = 0$ for $j = 11, \ldots, 14$. Thus, 10 dimensions determine the energy, while the other 4 dimensions can be seen as Gaussian noise uncorrelated with any of the relevant input dimensions.

In Figure 2.1(a), we show the relationship between the probability for particles to rearrange, $P(R|S)$, and S by plotting $\text{logit}\, P(R|S) \equiv \log[P(R|S)/(1-P(R|S))]$ vs. S. $P(R|S)$ is calculated by fitting a logistic regression that predicts whether a particle is rearranging from its S. This plot is analogous to the $\log P(R|S)$ vs. S plots in earlier studies [13] since in our model $\text{logit}\, P\,(R|\Delta E)$ is linear in ΔE. The plot shows a similar linear relationship between $\text{logit}\, P(R|S)$ and S. When $\text{logit}\, P(R|S)$ is plotted as a function of $1/T$ for several values of softness (Figure 2.1b), we also see a linear relationship between $\text{logit}\, P(R|S)$ and $1/T$ as observed in earlier studies [13]. As in

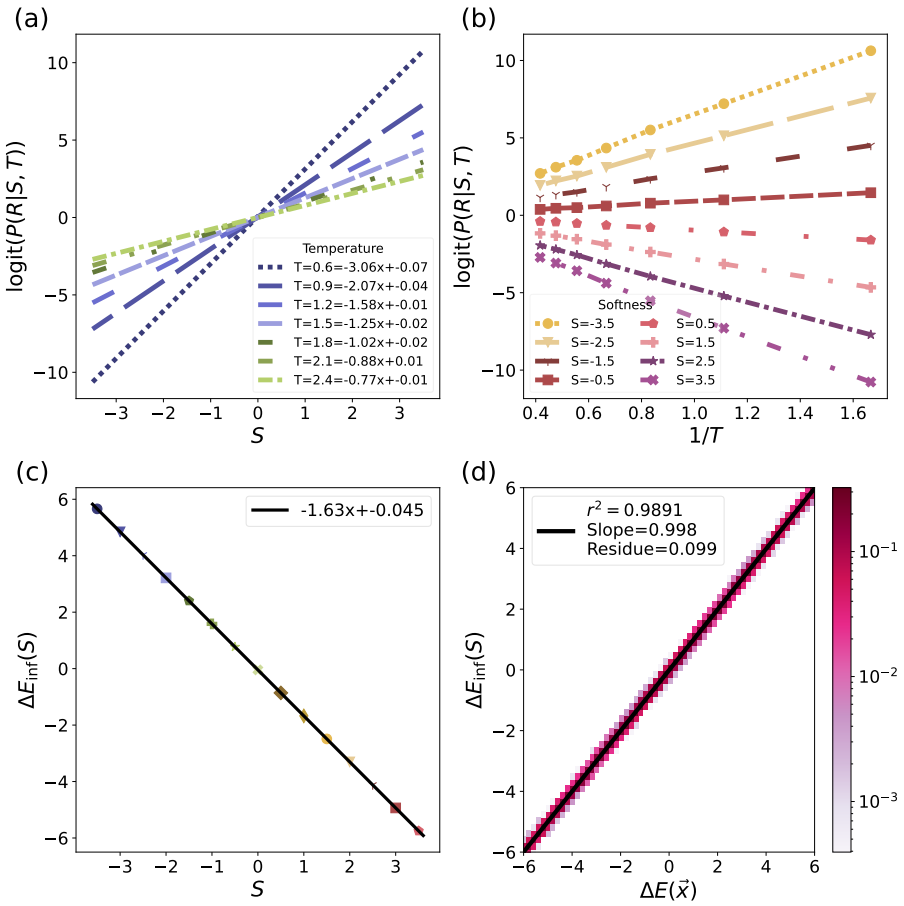

Figure 2.1: **Relationship between softness S and $\Delta E(\vec{x})$ for symmetric distribution of training energies for a large training set size, $N = 10^6$. (a)** logit $P(R|S)$ derived from fitting the logistic curve to the probability of rearrangement as a function of S for different temperatures T. **(b)** logit $P(R|S,T)$ vs. $1/T$ for 8 different values of softness. **(c)** The inferred $\Delta E_{\text{inf}}(S)$, calculated from logit $P(R|S,T)$, as a function of S. **(d)** 2D joint density plot and the linear fit of the true energy barrier $\Delta E(\vec{x})$ vs. the inferred energy barrier $\Delta E_{\text{inf}}(S)$ (we plot the joint density instead of the scatter for clarity of the visualisation).

the previous work [13], the slope of logit $P(R|S)$ vs. $1/T$ for each softness S is used to infer the corresponding energy barrier $\Delta E_{\text{inf}}(S)$ in Figure 2.1c. This $\Delta E_{\text{inf}}(S)$ is analogous to the barrier energy $\Delta E(S)$ in the Arrhenius rate equation, Eq. 2.1. As one can see, the inferred barrier energy, $\Delta E_{\text{inf}}(S)$, has a linear relationship with softness, S. Thus, our model, in this limit, reproduces the observations of previous work [13]: the probability of rearrangement is exponential in the distance S to the

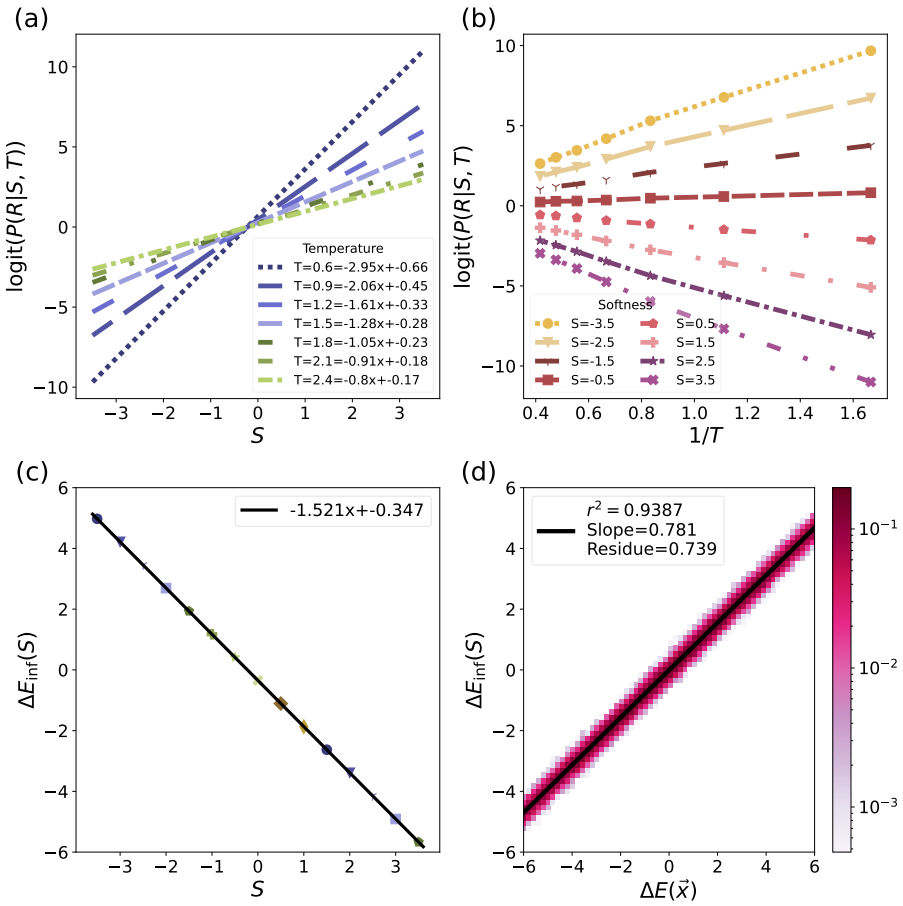

Figure 2.2: **Relationship between softness and $\Delta E(\vec{x})$ for symmetric distribution of training energies for a small training set size, $N = 10^3$. (a), (b), (c)** Same as in Figure 2.1. In **(d)** The true energy barrier $\Delta E(\vec{x})$ vs the inferred energy barrier from SVM $\Delta E_{\text{inf}}(S)$ is plotted. Note that, to the extent that the slope in **(d)** is not 1, the correct energy is not learned.

separating hyperplane, a.k.a. softness, and this distance has an interpretation as an inferred energy barrier $\Delta E_{\text{inf}}(S)$.

Unlike in past work, in our model, the true energy barriers are *known*. Thus, we then can compare the inferred energy barrier $\Delta E_{\text{inf}}(S)$ to the true energy barrier $\Delta E(\vec{x})$ for each configuration $\vec{x}_i$ in the test set. We plot the inferred energy vs. the true energy, as well as a linear regression line between the two in Figure 2.1(d). Since the slope of the fit is ≈ 1.0, and the scatter around the linear fit is small, we conclude that the SVM, indeed, learns the real energy barrier $\Delta E(\vec{x})$ with a high

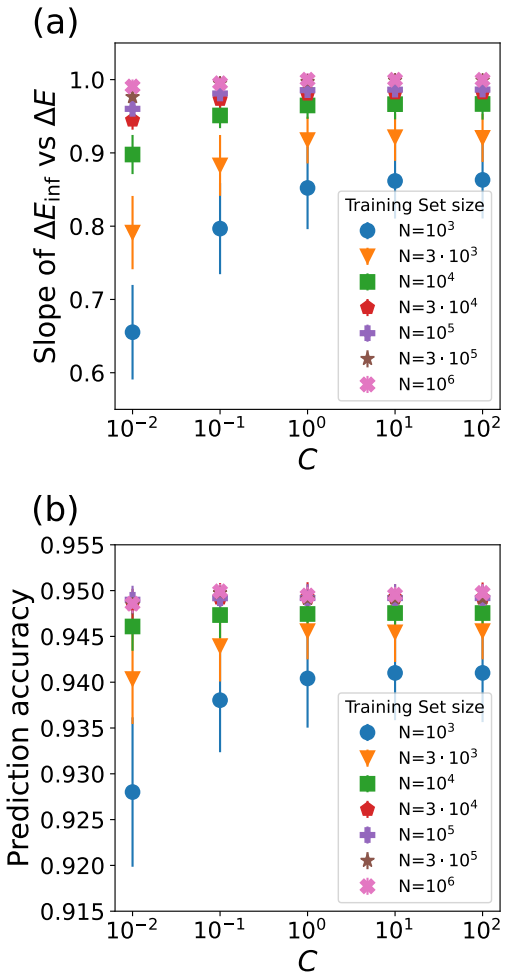

Figure 2.3: Slope of inferred energy **(a)** $\Delta E_{\text{inf}}(S)$ vs. real energy $\Delta E(\vec{x})$ and the prediction accuracy **(b)** for different sizes of training data as a function of the SVM cost parameter C. The training and test data were generated at $T = 0.4$.

degree of accuracy. We also find that the SVM captures the real energy when trained on unsymmetrical data where all energy barriers are positive. (Section 2.3.5).

2.3.2 Large training sets are required for SVM to learn true energy barriers

For real-world problems, we do not have access to an infinite (extremely large) amount of data. Thus, it is natural to ask whether inferred energies are still accurate for smaller training sets. For this, we repeated the analysis of section 2.3.1 with varied

training set size $N = 10^3, \ldots, 10^6$.

As shown in Figure 2.2(a–c), when $N = 10^3$, the inference procedure still seems to work. That is, logit $P(R|S,T)$ is still a linear function of S, and it still appears to be linear in $1/T$. This allows us again to infer the energy barrier $\Delta E_{\text{inf}}(S)$, which is linear in S. However, regressing $\Delta E_{\text{inf}}(S)$ against the true $\Delta E(\vec{x})$ shows that the inferred energy is *biased*, consistently underestimating the magnitude of the true energy by nearly 15%. Since the variance of the distribution of true energy $P(\Delta E)$ is a sum of the variance explained by S and the variance unexplained by S, the error must always have this sign: if the energy is inferred incorrectly, the variance of the distribution of inferred energies will be less than the variance of the distribution of true energies. This point is discussed further in Section 2.3.3.

Figure 2.3(a) show how this underestimation depends on N. Further, Figure 2.3(b) shows the N dependence of the classification (rearranged or not) prediction accuracy of our fitted model on a test set, different from the training one. To verify that fitting and prediction errors do not come from suboptimal choices during training, in this Figure, we also change the value of the SVM training hyperparameter C, which controls when the SVM treats data points that are labeled differently from their neighbors as outliers vs. true data that should be fitted [74, 75]. For small N, regardless of C, the true energy is underestimated. For large N, the quality of the fits improves, and the prediction accuracy as well as the error in slope become largely insensitive to C. C controls how misclassifications are treated in a SVM. Higher the value of C higher the penalty for misclassification though for noisy datasets this may lead to overfitting.

In practice, the true energy is rarely known. Thus detection of the bias shown in Figs. 2.2(d), 2.3 is nontrivial in experimental applications. Indeed, simple checks, such as verifying the linearity of plots in Figure 2.2(a,b,c), do not reveal this error. Further, the underestimation of the barrier magnitude is also difficult to diagnose by

looking at the prediction accuracy, Figure 2.3(b). When the true energy is underestimated by 15%, the prediction accuracy is still 94% ($C = 10^2$, $N = 10^3$), which is only 1% lower than the highest value obtained with large N. Since we do not have any prior information about the maximum possible prediction accuracy for specific experimental data sets, these figures suggest that, judging by the prediction accuracy only, one can never be sure if the learned energy is a good estimate of the true one: a seemingly high accuracy is not enough!

2.3.3 Presence of redundant features in the input data degrades the quality of the inference

In Ref. [13], 166 inputs were used for predicting rearrangements. However, many of these inputs were correlated with one another. To model this, we repeat our analysis using a higher-dimensional input vector. For this, as explained in Sec. 2.2, we train the SVM on a 20 dimensional input. Of these input dimensions, x_i^j, $j = 1, \ldots, 10$ were independently sampled from a Gaussian distribution, and the remaining inputs were strongly nonlinearly correlated with them. We again train an SVM on $N = 10^6$ balanced data points. The logit $P(R|S)$ vs. S plot (Figure 2.4a), logit $P(R|S,T)$ vs. 1/T plot (Figure 2.4b) and the inferred energy $\Delta E_{\rm inf}(S)$ vs softness plot (Figure 2.4c) again are linear, as in Figure 2.1 and the previous work [13]. However, plotting the inferred energy $\Delta E_{\rm inf}(S)$ vs. the true energy barrier $\Delta E(\vec{x})$ for each configuration and producing a linear fit between them, cf. Figure 2.4d, we see that the magnitude of the inferred energy is underestimated compared to the true energy even for very large N (cf. Figure 2.5). Looking at the optimal hyperplane learned by the SVM, we observe that the hyperplane contains contributions from the input variables that do not contribute to the true energy (not shown). One would not be aware of this problem from Figure 2.4(a-c) alone. We remind the reader that the true energy needed to produce Figure 2.4(d) is typically unknown.

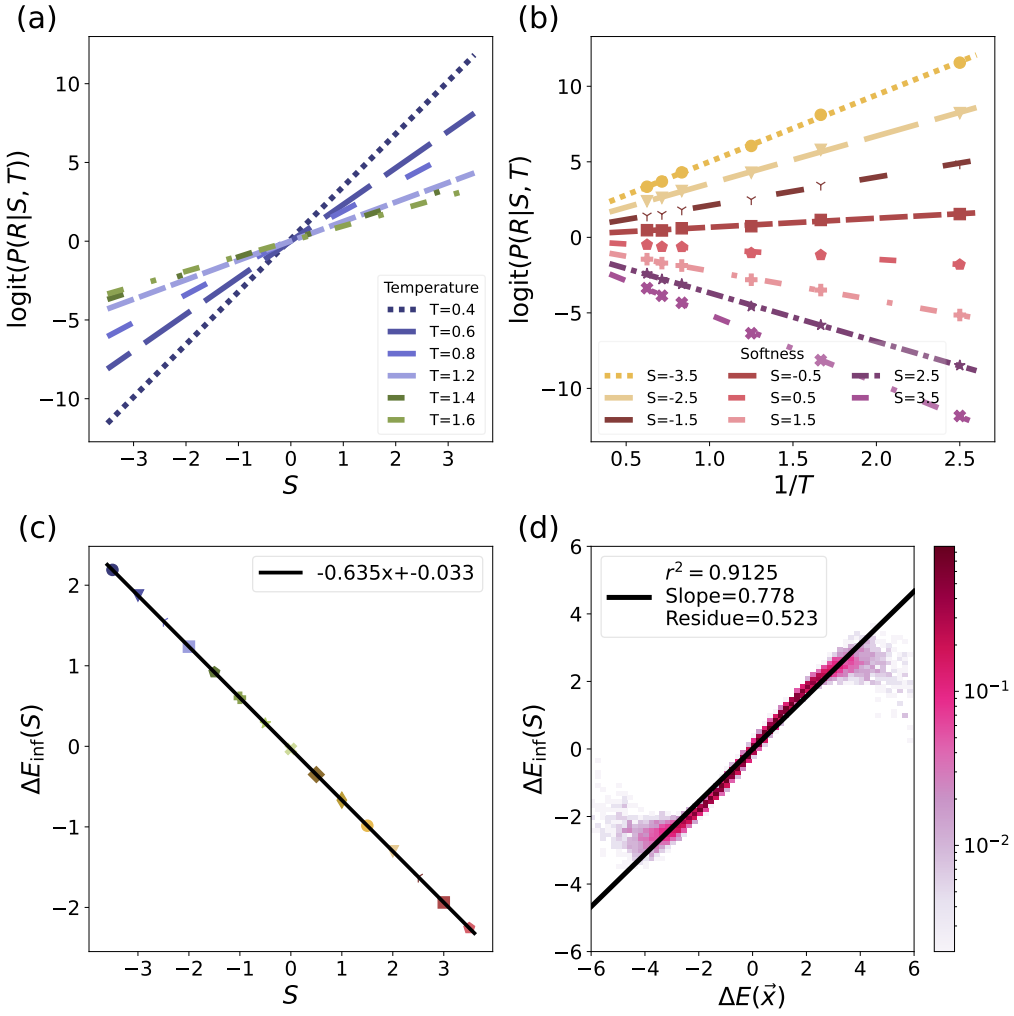

Figure 2.4: **Relationship between softness and $\Delta E(\vec{x})$ for a symmetric distribution of training energies and with spurious, correlated input terms.** Same plotting convention are used as in Figs. 2.1. In **(d)** the true energy barrier $\Delta E(\vec{x})$ vs the inferred energy barrier from SVM $\Delta E_{\text{inf}}(S)$ is plotted. The error to the fit is given by the purple semi transparent spread on both sides of the fit on a 2D density plot. Note that, to the extent that the slope in **(d)** is not 1, the correct energy is not learned. Also the deviation between the fit and the 2D density plot at the edges shows that even though a linear fit was used to fit the energy and softness and it fit has a high r^2 value the underlying function one is trying to fit is not really linear in S.

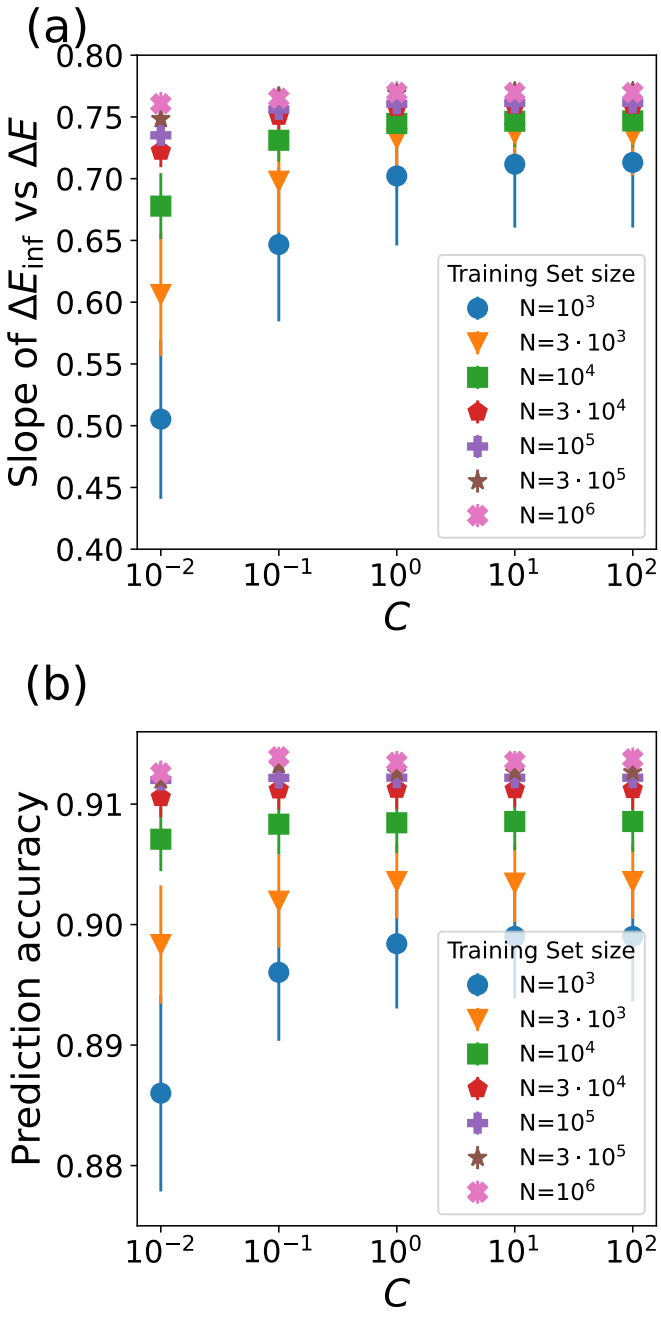

Figure 2.5: **(a)** Slope of the inferred energy, $\Delta E_{\text{inf}}(S)$, vs. the true energy, $\Delta E(\vec{x})$, and **(b)** the prediction accuracy for the model with spurious correlated inputs. Same plotting conventions as in Figure 2.3.

To design a method for identifying the bias from data, we note again that the variance of the true energy barrier distribution is a sum of the variance explained by S (i. e., the variance of $\langle \Delta E \rangle(S)$ over the distribution of S) and the variance conditional on S (i. e., the part of the energy barrier *not* captured by S). Thus, if we can find a different set of coordinates that allows the SVM to learn a different S that is *closer* to the true energy, this improvement should manifest as an increase in the variance of the distribution of inferred energies, $\mathrm{Var}[\Delta E_{\mathrm{inf}}]$. Our approach is then to reduce dimensionality of the input space, aiming to remove the correlated dimensions and increase the accuracy of the model at the same time. A particular version of this approach is known in the SVM literature as Recursive Feature Elimination (RFE) [76] procedure. RFE has been used in earlier work on predicting rearrangements [77, 78] for pruning the dimensionality of SVM inputs. Assuming that all input dimensions are normalized to the same variance, RFE works by removing the input dimension with the smallest magnitude contribution to the separating hyperplane. One then refits the SVM and continues the process iteratively. Figure 2.6a shows the variance of the inferred energy as a function of the number of inputs kept by the RFE procedure. The peak in $\mathrm{var}[\Delta E_{\mathrm{inf}}]$ clearly matches the true number of dimensions that contribute to the energy in our model. Figure 2.6b shows a corresponding (but broader) peak in the prediction accuracy as well. These analyses bode well for using RFE for pruning the input data and resulting in a more accurate inference of the energy barrier in real world problems.

2.3.4 The inferred energy and predictors in real glass simulations can be approximated by gaussian

We looked at the distribution of each of the 266 dimensions used to train the SVM in for the Kob-Anderson model supercooled liquid [73]. All the dimensions looked unimodal. We calculated the kurtosis of all the dimensions, which measures how

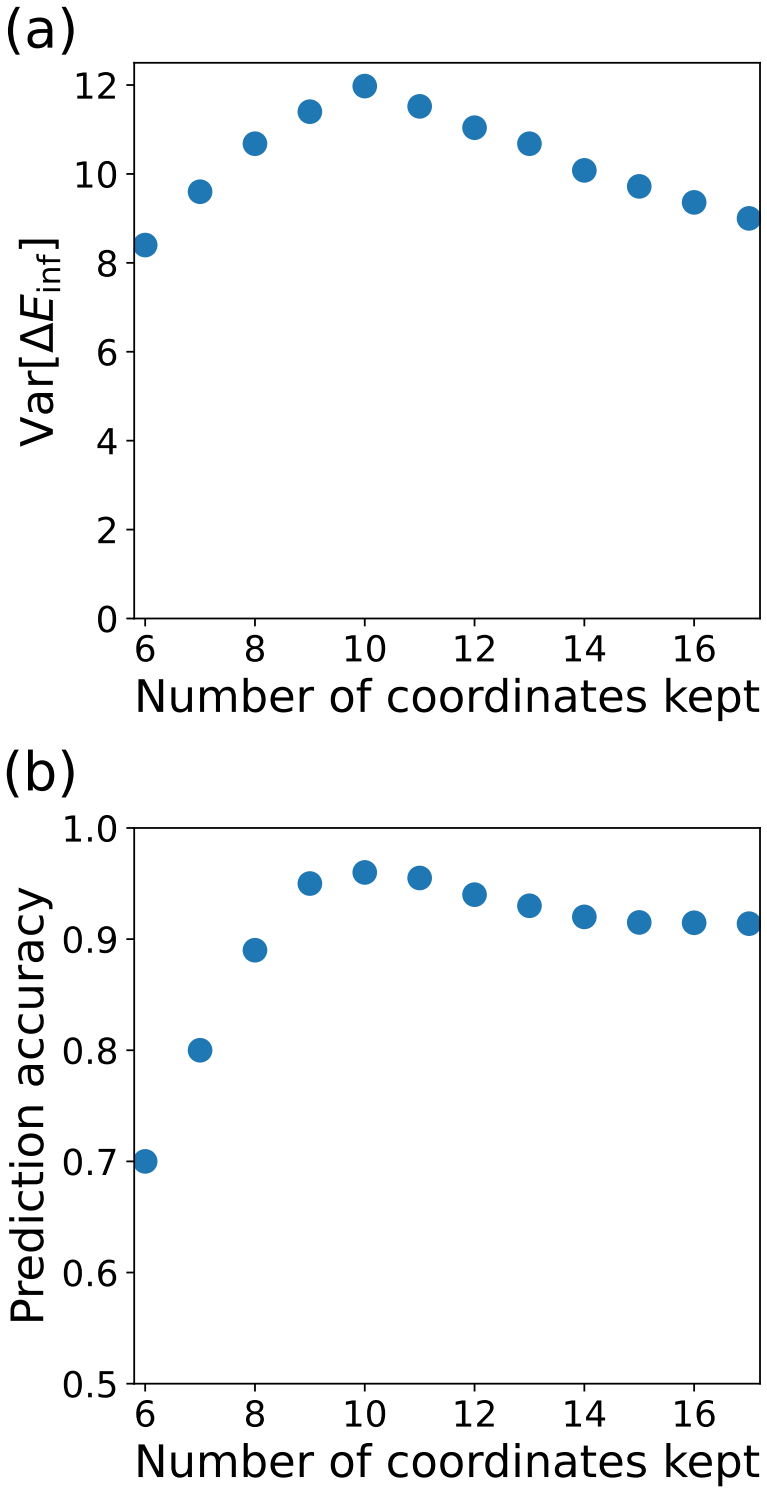

Figure 2.6: Plot of variance of ΔE_{inf} and the prediction accuracy as a function of number of coordinates kept. The variance of the distribution of ΔE_{inf} is more sensitive for detecting relevant dimensions. The variance of inferred energy at the peak matches well with the variance of the distribution of true energy (12 in our units).

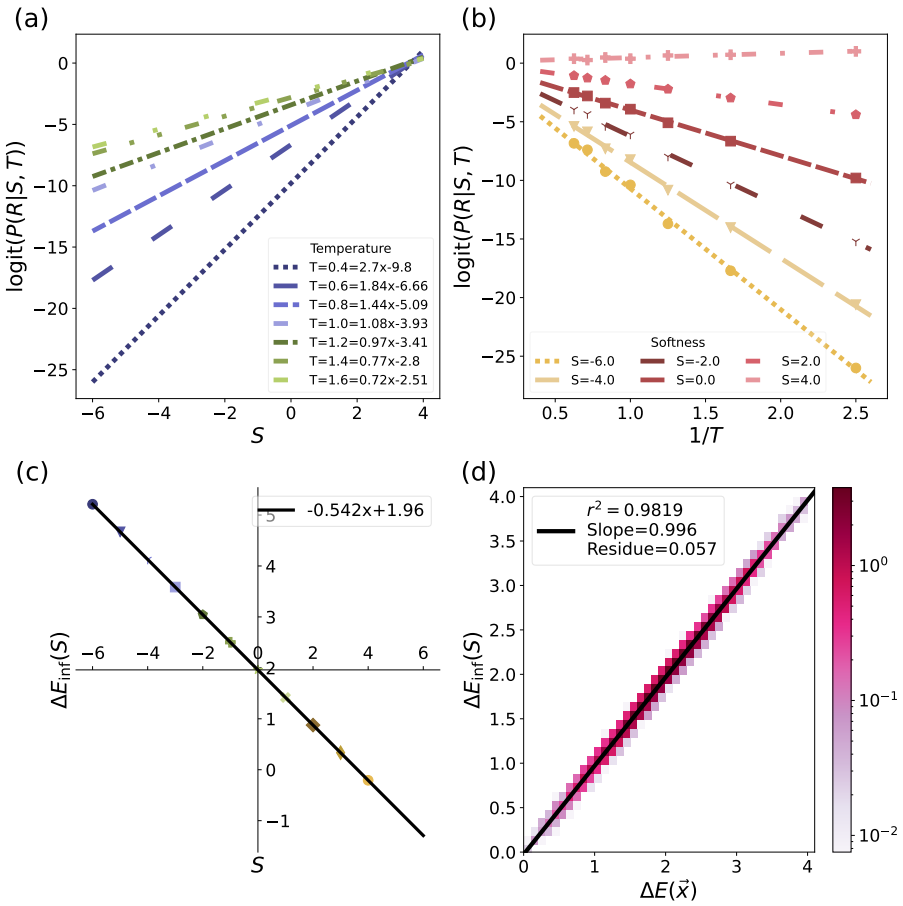

Figure 2.7: **Relationship between softness and $\Delta E(\vec{x})$ for positive energy barriers and a balanced dataset.** To balance our dataset, we choose 50% of samples where rearrangement was observed, and 50% where it was not. Plotting conventions are the same as in Figure 2.1. Note that the correct energy is learned (slope of 0.996), and the spread in the 2D density plot is minimal.

heavy tailed or light tailed a distribution is compared to a Gaussian distribution, which has kurtosis 0. 73% of the dimensions had a kurtosis in the range of (-0.3, 0.3) and 92% of dimensions had a kurtosis in the range of (-2,2). The values of kurtosis cutoffs acceptable for normality vary widely from ±2 to ±6 [79, 80, 81, 82, 83], and thus the true structural features have roughly Gaussian distributions.

2.3.5 Qualitative results remain unchanged when trained on data with non-centered distribution of energy barriers

Recall, as explained in the main text, that in the true system all energy barriers are positive. However, in the main text, we chose energy barriers to be symmetric around zero for simplicity. Figure 2.7 is the analogue of Figure 2.1, but now evaluated for a model where almost all energy barriers are positive. We balance the training set, similarly to Ref. [13], so that the number of rearranging and non-rearranging particles is the same.

We draw each of the dimensions from a Gaussian distribution with unit variance centered at zero. We have 10 independently sampled input dimensions, with $\alpha = 0.4$ for $j = 1, \ldots, 10$. Further, we add a constant to the energy so that the mean of the distribution is two standard deviation away from zero, and thus the energy is almost always positive. We use $N = 3 \times 10^5$ training samples with 1.5×10^5 examples each of rearranging and non-rearranging configurations to train the SVM. As seen in figure 2.7, the results for the probability of rearrangement and the inferred energy remain qualitatively unchanged from Figure 2.1 in the *Main text*. In particular, the correct energy barriers are learned.

Just as in the case with a centered ΔE distribution, with a non-centered ΔE distribution, the energy is not learned correctly at small training sample sizes. We generated a non-centered ΔE distribution as above, and generated training sets of different sizes, balancing them as above. As can be seen from a Figure 2.8 , these observations are qualitatively the same as in the centered case.

We also note that giving each variable x^j a nonzero mean μ_j has no effect except producing a nonzero mean ΔE, and is thus expected to be covered by the above checks. To see this, write $x^j = \mu_j + y^j$, where y^j has mean zero. We then have

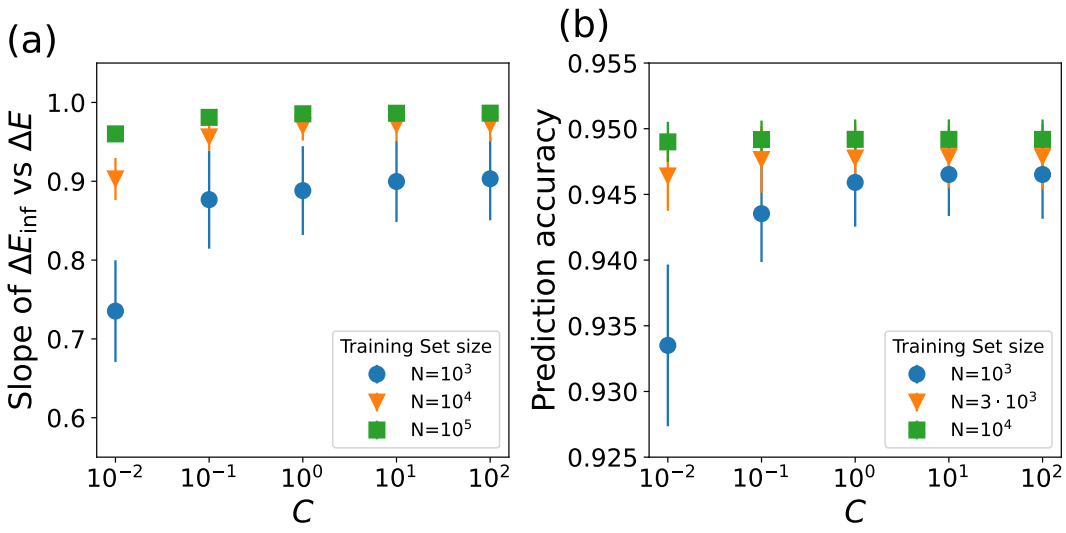

Figure 2.8: **(a)** Slope of the inferred energy, $\Delta E_{\text{inf}}(S)$, vs. the true energy, $\Delta E(\vec{x})$, and **(b)** the prediction accuracy for the model with non centered ΔE distribution. Same plotting conventions as in Figure 2.3.

$$\Delta E = \sum_j \alpha^j \mu_j + \sum_j \alpha^k y^j. \tag{2.5}$$

Thus, the only effect of giving x^j nonzero mean is to add a constant $\sum \alpha^j \mu_j$ to ΔE. Further, note that even in this case where $\mu_j \neq 0$, changing the sign of α^j only changes the mean ΔE: it has no effect on $\sum \alpha^j y^j$, since the distribution of y is symmetric around 0. Thus, qualitative results such as the above, which hold both when the mean ΔE is 0 and when it is positive, are expected to still hold when some of the α^j are negative.

2.3.6 Effect of missing features

In any realistic system, some of the features needed to express ΔE will be missing. Here we confirm that this prevents the correct energy from being learned. We use $N = 10^6$ training samples with 5×10^5 examples each of rearranging and non-rearranging configurations to train the SVM. The distribution of energies in the training sample

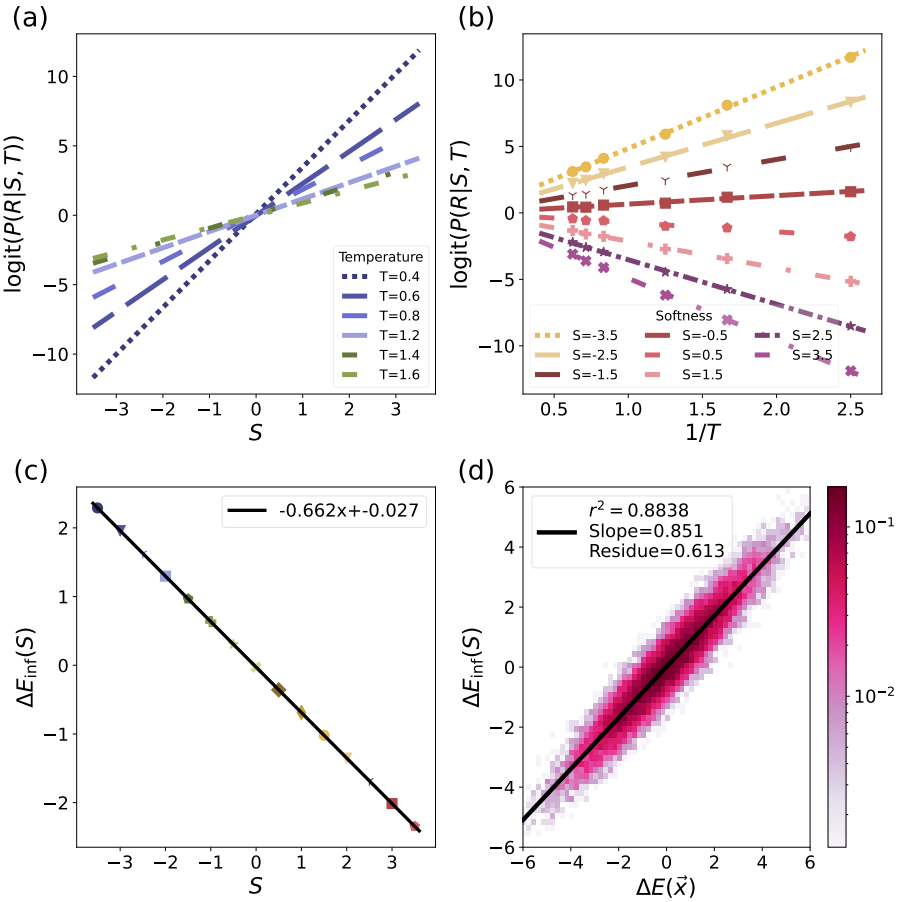

Figure 2.9: **Relationship between softness and $\Delta E(\vec{x})$ for symmetric distribution of training energies for a large training set size, $N = 10^6$, with one of the relevant feature missing.** **(a), (b), (c)** Same as in Figure 2.1. In **(d)**, the true energy barrier $\Delta E(\vec{x})$ vs the inferred energy barrier from SVM $\Delta E_{\text{inf}}(S)$ is plotted. Note that, to the extent that the slope in **(d)** is not 1, the correct energy is not learned.

is symmetric. We use 10 independently sampled input dimensions, with $\alpha^j = 1.2$ for $j = 1, \ldots, 10$, to determine the energy. Out of the 10 dimensions we train the SVM only with the first 9 dimensions and drop the last dimension. In this case, one ends up underestimating the variance of the true energy as can be seen from Figure 2.9

2.3.7 Effect of different choices of correlated features

As our model of correlated features in Section 2.3.3, we have chosen to add nonlinear functions of the "correct" input features to the input. Here we check that the results

of Section 2.3.3 generalize to other choices of correlated input features. In particular, we test two other options. Firstly, we consider addition of variables that, rather than being nonlinear functions of the "correct" input features, are simply linearly correlated with them. Secondly, we consider a set of input features that are nonlinearly correlated, but are not "redundant", in the sense that, in principle, all of them are required to express the true energy through a linear function. In both cases, we find that the results of Section 2.3.3 remain qualitatively unchanged.

2.3.8 Effect of redundant linear feature

We use $N = 10^6$ training samples with 5×10^5 examples each of rearranging and non-rearranging configurations to train the SVM. The distribution of energies in the training sample is symmetric. We have 10 independently sampled input dimensions, with $\alpha^j = 1.2$ for $j = 1, \ldots, 10$. Thus, 10 dimensions determine the energy. We train the SVM on a 12 dimensional input which consists of all the 10 dimensions and 1 extra copies each of $j = 1, 2$. This gives two extra, redundant features which are linear in the relevant coordinates. In this case, the SVM again underestimates the variance of the true energy, as can be seen from Figure 2.10.

2.3.9 Effect of non-redundant and non-linear correlated features

We use $N = 10^6$ training samples with 5×10^5 examples each of rearranging and non-rearranging configurations to train the SVM. The distribution of energies in the training sample is symmetric. We have 10 independently sampled input dimensions, with $\alpha^j = 1.2$ for $j = 1, \ldots, 10$. Thus, 10 dimensions determine the energy. Instead of giving the SVM $x^1 \ldots x^{10}$ as input features, we use the 14 input features

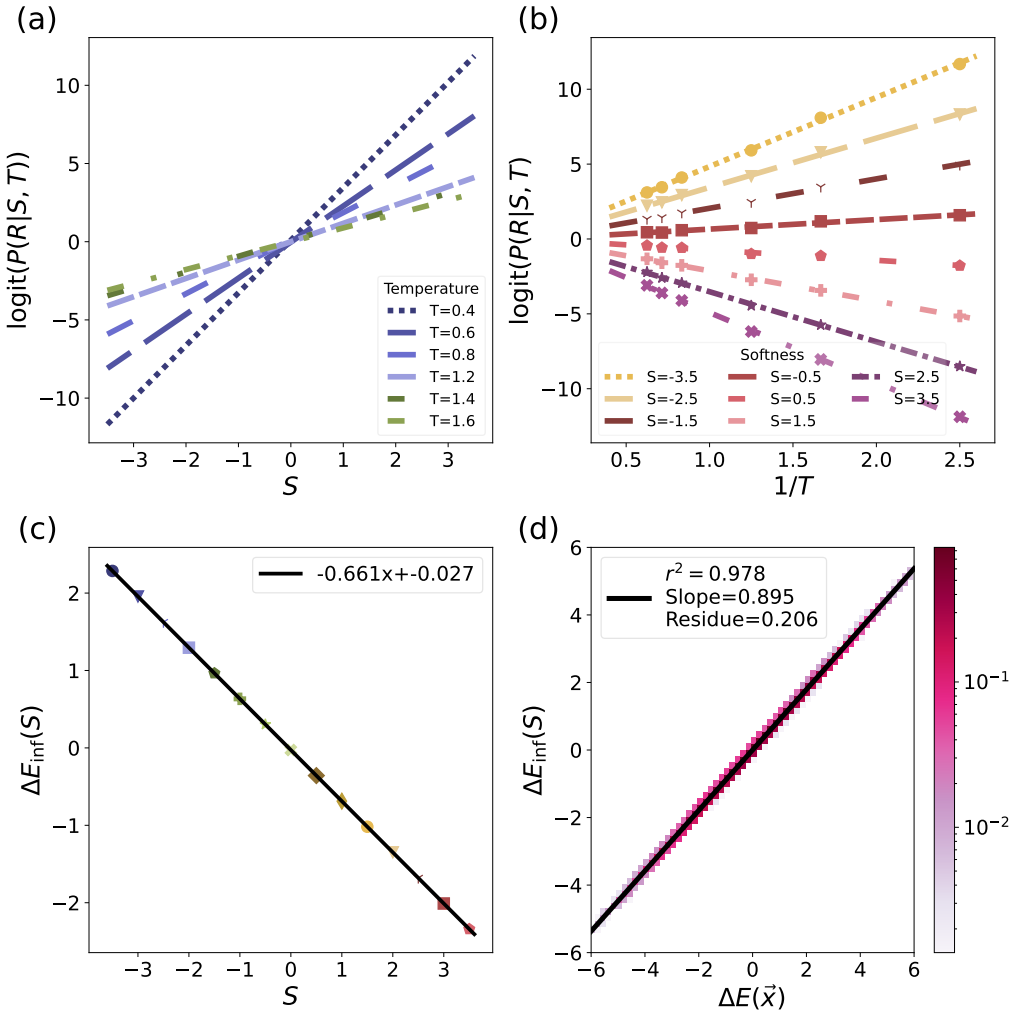

Figure 2.10: **Relationship between softness and $\Delta E(\vec{x})$ for symmetric distribution of training energies for a large training set size, $N = 10^6$, with additional linear features. (a), (b), (c)** Same as in Figure 2.1. In **(d)**, the true energy barrier $\Delta E(\vec{x})$ vs the inferred energy barrier from SVM $\Delta E_{\text{inf}}(S)$ is plotted. Note that, to the extent that the slope in **(d)** is not 1, the correct energy is not learned.

$$x^1 + (x^5)^3, x^2 + (x^6)^3, x^3 + (x^7)^3, x^4 + (x^8)^3, x^5, x^6, \ldots, x^{10}, (x^5)^3, (x^6)^3, (x^7)^3, (x^8)^3. \tag{2.6}$$

(There is nothing particular about these features, and we believe that other combinations of powers of predictors would deliver a similar point.) It should be possible for the SVM to learn a linear combination of these features that would cancel out the cubic terms and infer the true energy. Nonetheless, we observed that the SVM *does not* learn the correct energy, Figure 2.11. Thus, the presence of non-linearities as well as redundant features affects the ability of SVM to predict the correct energy.

2.4 Discussion

We have shown that, in our toy model, one can always use a linear SVM to predict rearrangements with a high accuracy, though the amount of data needed for this might be larger than what typical experiments would allow in realistic cases. However, even if the inference seems successful, the inferred energy barrier matches the true energy only in specific cases. Crucially, by observing a high prediction accuracy or high quality linear relationship between softness, log rearrangement probability, and $1/T$, one cannot conclude that the correct energy has been learned. The problem becomes severe—even in our simple model—when the input data has extra features, potentially nonlinearly correlated with true variables describing the model. Realistic systems, e. g. glasses, are likely to have different types of correlations between their input features than those we have considered here. Nonetheless, our results suggest a need to carefully scrutinize the use of ML methods, and specifically SVMs, for inference of energy barriers in glasses.

For our model, we have demonstrated a method to diagnose and fix this problem: recursive feature elimination (RFE) can be used to remove "confusing" input features. By tracking the variance of the inferred energy barriers or of $\log P(R|S)$,

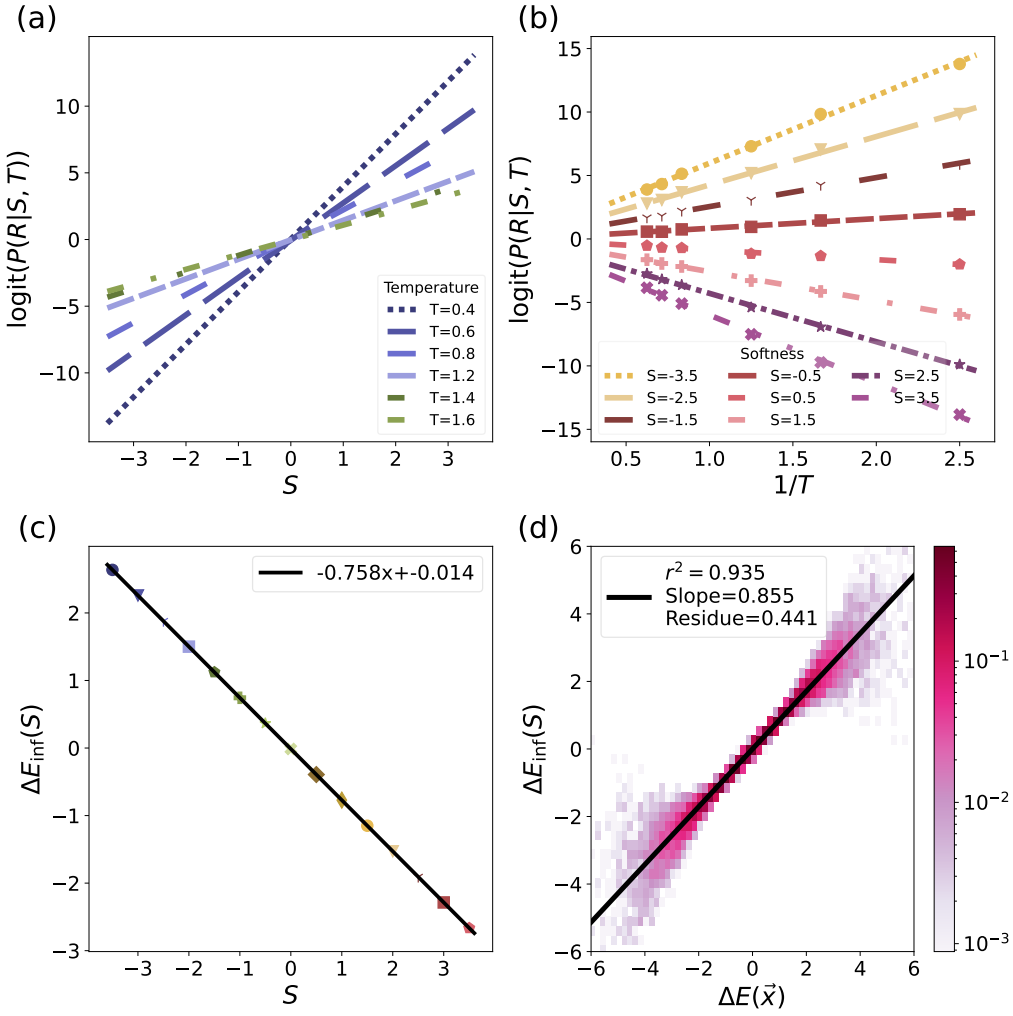

Figure 2.11: **Relationship between softness and $\Delta E(\vec{x})$ for a symmetric distribution of training energies and with spurious, correlated input terms.** Same plotting convention are used as in Figs. 2.1. In **(d)**, the true energy barrier $\Delta E(\vec{x})$ vs the inferred energy barrier from SVM $\Delta E_{\text{inf}}(S)$ is plotted. The fit error is illustrated by the purple semi transparent spread on both sides of the fit on the 2D density plot. Note that, to the extent that the slope in **(d)** is not 1, the correct energy is not learned. Also the deviation between the fit and the 2D density plot at the edges shows that, even though a linear fit was used to fit the energy and softness, and the fit had a high r^2 value, the underlying function we are trying to fit here is not linear in S.

which is maximal when the true barriers are learned, improvements in the inference of the barriers can be detected, even though the true barriers are unknown and the prediction accuracy may change little. RFE is particularly natural in our problem because there is a clear division between important and unnecessary input dimensions. For other systems, RFE may not be the best method for adjusting the set of input features. For example, Section 2.3.9 shows an example of a set of correlated features where no smaller set is sufficient to express the energy, and thus RFE cannot recover the true energy. As another example, if the input features are a discretization of the pair correlation function $g(r)$, it may be more natural to coarsen this discretization, or to change the choice of basis functions, than to eliminate specific input features. However, our criterion for comparing different choices of input features is general would still stand: features that produce a larger variance in the inferred energy barriers should be closer to predicting the true barriers. We expect it to be true in general that the choice of features for the inference will affect whether or not the true energy is learned, so that different possible choices should be compared using this criterion. The need to make such comparisons between different choices of input features and other hyperparameters, rather than only focusing on achieving the best possible prediction accuracy, is one of the main conclusions of our work.

In our simple model, the probability of particle rearrangement is purely a function of energy. However, when SVMs are used to predict rearrangements in real systems, the probability is a function of energy as well as of an entropic prefactor, both of which are found to depend on S [13], see Eq. (2.1). In addition, there are other complications not present in our toy model, such as ambiguity in the identification of rearrangements. We expect such complications to only strengthen our conclusion that a good prediction accuracy does not guarantee that the ML model learns the true values of the energy barriers.

It may seem surprising that the addition of extra coordinates degrades the pre-

diction accuracy and the quality of inference of ΔE_{inf}. Conventional wisdom is that such overcomplete representation should improve SVM accuracy by creating a higher-dimensional embedding space, in which the data become linearly separable [75]. It is possible that the failure of this intuition in our case comes from the probabilistic nature of rearrangements: for any $\vec{x}$, there are both rearranging and non-rearranging examples, at least in the $N \to \infty$ limit. Thus, the data are fundamentally not separable, irrespective of the space in which we embed them.

The process of adding more correlated coordinates explicitly to our input is similar to using some nonlinear kernel on the original data. SVM kernels allow us to create high dimensional embeddings that are nonlinear functions of the input coordinates without having to explicitly evaluate the embedding, and these embeddings are often even infinite-dimensional. Thus, our results seem to imply that using a kernel may also prevent the true energy barriers from being learned.

In our work, we have focused specifically on linear SVMs, rather than other ML methods, because this is the only method, which has been used in the past to explicitly deduce the underlying energy barriers from the inferred statistical model. However, note that we have chosen the true energy function to be expressible by a linear SVM. Further, note that more complex ML methods are generally thought to behave similarly to kernel methods [84, 85]. Thus we expect that our results are not caused by the simplicity of linear SVMs, and they will generalize to other ML approaches to the problem of learning energy barriers in glassy systems.

Our results may have implications for many systems beyond supercooled liquids, for which the underlying "physics" must be learned from an ML model trained on the data. Indeed, we have shown that, even given a powerful ML model that can express the true underlying physics, an arbitrarily large amount of training data, and a good prediction accuracy, the model may fail to learn a correct physical description even in a relatively simple scenario. We suspect that, in real world applications, this

problem will become even more severe. One must then use independent methods—going beyond prediction accuracy—to evaluate the model quality.

Chapter 3

Distribution of singular values in large sample cross-covariance matrices

3.1 Introduction

[1] Many data-science applications require detecting correlations between two variables X and Y of dimensions N_X and N_Y, respectively, with $N_X, N_Y \gg 1$. When these variables are sampled T times, with $T \sim N_X, N_Y$, sampling fluctuations can produce spurious correlations, even when X and Y are truly uncorrelated. Characterizing these sampling-induced correlations is essential before isolating genuine signals in real datasets.

Marchenko and Pastur famously analyzed similar correlations in sample self-

[1] This Chapter presents the paper [86] "Swain, Arabind, Sean Alexander Ridout, and Ilya Nemenman. "Distribution of singular values in large sample cross-covariance matrices." arXiv preprint arXiv:2502.05254 (2025)." The work was conducted in collaboration with Drs. Sean Alexander Ridout and Ilya Nemenman. I performed all simulations, conducted all analyses, and led writing of the manuscript. Dr. Nemenman conceived the model and led the project, while Dr. Ridout contributed to discussions regarding the calculations, procedures and analyses. All authors participated in writing and reviewed the final manuscript, currently in review in *Phys. Rev. E.*

covariance matrices [32], deriving their eigenvalue distribution using now-classic methods of Random Matrix Theory (RMT) [87]. For $T > N_X, N_Y$, later work generalized these results to cross-correlations of whitened variables (linearly transform the data so that all the resulting variables have zero mean, are uncorrelated such that their covariance matrix becomes an identity matrix) [88, 89, 90]. However, to our knowledge, no comparable results exist for the unwhitened cross-covariance between X and Y and arbitrary relations between T, N_X, and N_Y, though some related results have been calculated [91, 92, 93].

In this paper, we derive the eigenvalue spectra of unwhitened cross-covariance matrices for uncorrelated Gaussian i.i.d. data and arbitrary relations among T, N_X, and N_Y. We hope that these can then be used to distinguish signal from sampling noise in data science applications. For example, we hope to use these methods to understand noise-noise correlations between neural recordings (let's consider it to be $\mathbf{Y}$) and motor recordings (let's consider it to be $\mathbf{X}$). The motor recordings are generally well sampled as compared to the neural recordings and this would correspond to the case $N_Y > T > N_X$.

3.2 Model and methods

We consider T samples of random variables X and Y combined into matrices $\mathbf{X}$ and $\mathbf{Y}$, with dimensions $T \times N_X$ and $T \times N_Y$, respectively. The entries of $\mathbf{X}$ and $\mathbf{Y}$ are i.i.d. Gaussian random variables with zero mean and variances σ_X^2 and σ_Y^2 respectively,

$$X_{t\mu} \sim \mathcal{N}(0, \sigma_X^2)\,, \quad Y_{t\nu} \sim \mathcal{N}(0, \sigma_Y^2)\,, \tag{3.1}$$

$$t = 1, \ldots, T,\ \mu = 1, \ldots, N_X,\ \nu = 1, \ldots, N_Y, \tag{3.2}$$

so that the measured correlations in X and Y vanish asymptotically, as $N_X/T, N_Y/T \to 0$.

We define normalized matrices as

$$\widetilde{\mathbf{X}} = \frac{\mathbf{X}}{\sigma_X}, \qquad \widetilde{\mathbf{Y}} = \frac{\mathbf{Y}}{\sigma_Y}\,. \tag{3.3}$$

For $T \gg 1$, each column in these matrices has variance of nearly one. Note that, in typical applications, σ_X and σ_Y would be estimated from samples as well, and the estimates might be different from their true value. Here we disregard this distinction, as in [91], arguing that sampling fluctuations in estimating scalar parameters are negligible compared to sampling effects on the infinitely many singular values.

The normalized empirical cross-covariance matrix (NECCM) $\mathbf{C}$ is then

$$\mathbf{C} = \frac{1}{T}\widetilde{\mathbf{Y}}^T\widetilde{\mathbf{X}}, \tag{3.4}$$

which has dimensions $N_Y \times N_X$. If $N_X \neq N_Y$, this matrix is not square, but it obviously has the same nonzero singular values as its transpose. Without loss of generality, in all calculations, we take $N_X \leq N_Y$.

We want to calculate the distribution of these singular values. To utilize RMT methods, most of which only work for square symmetric matrices, we focus instead on eigenvalues of

$$\mathbf{C}^T\mathbf{C} = \frac{1}{T^2}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\widetilde{\mathbf{X}}. \tag{3.5}$$

The nonzero eigenvalues of $\mathbf{C}^T\mathbf{C}$, which we denote as λ, are the same as nonzero eigenvalues of $\mathbf{C}\mathbf{C}^T$, and their distribution is related to the distribution of nonzero singular values of $\mathbf{C}$, denoted as γ, via

$$\rho_C(\gamma) = 2\sqrt{\lambda}\rho_{C^TC}(\lambda), \quad \gamma = \sqrt{\lambda}. \tag{3.6}$$

The distribution further contains a delta function at zero, corresponding to the zero eigenvalues of $\mathbf{C}^T\mathbf{C}$ when $N_X \leq N_Y$.

To explore the problem in different regimes, we define:

$$p_X \equiv T/N_X,\ p_Y \equiv T/N_Y,\ p_X \equiv 1/q_X,\ p_Y \equiv 1/q_Y. \tag{3.7}$$

Eigenvalue density. We compute the eigenvalue density of the square of the NECCM, Eq. (3.5), by computing its Stieltjes transform, as is the standard approach [87]. The Stieltjes transform of a matrix A is defined as

$$g_{A,N}(z) = N^{-1}\mathrm{Tr}(z\mathbf{I} - \mathbf{A})^{-1}, \tag{3.8}$$

where z is a complex number. We denote the large-N limit of $g_{A,N}$ by $\mathfrak{g}_{\mathbf{A}}$ [87]. The eigenvalue density is obtained from the Sokhotski–Plemelj formula

$$\rho_A(\lambda) = \frac{1}{\pi}\lim_{\eta\to 0^+} \Im\mathfrak{g}_{\mathbf{A}}(z = \lambda - i\eta)\,, \tag{3.9}$$

where $\Im$ denotes the imaginary part. We use a series of relatively common random matrix operations to obtain the Stieltjes transform of the square of NECCM, in the limit where $N_X, N_Y, T \to \infty$ with p_X and p_Y held fixed. These steps are outlined in the *Section* 3.5.

As the imaginary part of the Stieltjes transform gives us the eigenvalue density of the square of the NECCM, evaluating the discriminant involved in solving an algebraic equation for the Stieltjes transform (see *Section*) gives the boundaries of the range, in which the eigenvalue density is nonzero. These boundaries are denoted by $\lambda_\pm$. The corresponding values for nonzero singular values of NECCM are denoted by $\gamma_\pm$. Analytical expression for these boundaries for the cross-covariance spectrum of pure uncorrelated noise are one of the central results of this chapter.

Numerical simulations. We confirm our results by simulating the model, Eq. (3.1), numerically. Although the eigenvalue density is expected to be self-averaging, and

thus our calculations for $\rho(\gamma)$ will be exact for SVD of an *individual* matrix for sufficiently large T, making T very large substantially increases the computational costs. Thus, we simulate matrices with $T = 1000$, and more precisely test our predictions by averaging over 500 independent realizations.

3.3 Equation for Stieltjes transform and singular value density bounds

We calculate the density of eigenvalues of the square of NECCM in 3 cases, covering all possible relationships between T, N_X, N_Y: (1) $T > N_X,\ N_Y$, (2) $N_Y \geq T \geq N_X$, and (3) $T < N_X,\ N_Y$. For analyzing these different cases, we note that the square of the NECCM can be written as an $N_X \times N_X$ matrix $\frac{1}{T^2}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\widetilde{\mathbf{X}}$ or an $N_Y \times N_Y$ matrix $\frac{1}{T^2}\widetilde{\mathbf{Y}}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}$. Both of these matrices will have the same nonzero eigenvalues. Similarly, the $T\times T$ matrix $\mathbf{H} = \frac{1}{T^2}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T$ will have the same nonzero eigenvalues.

While nonzero eigenvalues of all of these matrices are the same, the total number of eigenvalues is different. For example, the Stiltjes transform $\mathfrak{h}$ of $\mathbf{H}$ gives the density for T eigenvalues of the $T \times T$ matrix, of which only $\min(T, N_X)$ are nonzero. The total number of eigenvalues in $\mathbf{C}^T\mathbf{C}$ and $\mathbf{C}\mathbf{C}^T$ is N_X and N_Y, respectively. Thus, the eigenvalue densities of the three matrices are not the same. To relate densities to each other, we need to subtract the δ functions at zero, and then rescale the densities of nonzero eigenvalues to one in all three cases.

With this, we write the finite size Stieltjes transform of $\mathbf{C}^T\mathbf{C}$:

$$g_{\mathbf{C}^T\mathbf{C},N_X}(z) = \frac{1}{N_X}\left(T\,\frac{1}{T}\sum_{\mu=1}^{T}\frac{1}{z-\lambda_\mu} + \frac{N_X - T}{z}\right) \tag{3.10}$$

$$= \frac{1}{N_X}\left(T\,h_T(z) + \frac{N_X - T}{z}\right) \tag{3.11}$$

$$= p_X h_T(z) + (1 - p_X)\,\delta(z), \tag{3.12}$$

where λ_μ are the T eigenvalues of $\mathbf{C}^T\mathbf{C}$ and $h_T(z) \equiv g_{\mathbf{H},T}(z)$. A similar expression, with N_X and N_Y swapped, holds for $\mathbf{C}\mathbf{C}^T$. Eq. 3.10 has 2 terms. The first term is the contributions of the T non zero eigenvalues to the Stieltjes transform defined in Eq. 3.8. The second term is $N_X - T$ delta functions. We get one delta function as a contribution of each of the zero eigen values to the Stieltjes transform defined by Eq. 3.8 and we have $N_X - T$ zero eigenvalues. $\mathfrak{h}$ in Eq. 3.12 is the Stieltjes transform of only for the $T \times T$ matrix which we can calculate from the RMT and g is the Stieltjes transform of the $N_X \times N_X$ we are interested in calculating.

An RMT calculation (Section 3.5) then shows that the Stieltjes transform $\mathfrak{h}$ of $\mathbf{H}$ satisfies a cubic equation

$$a\mathfrak{h}^3 + b\mathfrak{h}^2 + c\mathfrak{h} + d = 0, \tag{3.13}$$

where

$$a = z^2 p_X p_Y, \tag{3.14}$$

$$b = z\left(p_Y(1-p_X) + p_X(1-p_Y)\right), \tag{3.15}$$

$$c = \left((1-p_X)(1-p_Y) - z p_X p_Y\right), \tag{3.16}$$

$$d = p_X p_Y. \tag{3.17}$$

Thus, solving Eq. (3.13), and then using Eq. (3.12), gives the eigenvalue density of $\mathbf{C}^T\mathbf{C}$, which can be used to compute the density of the nonzero singular values of the cross-covariance using Eq. (3.6).

3.3.1 Spectrum of empirical cross covariance matrix when $T < N_X,\ N_Y$

The cubic polynomial given by Eq. (3.13) can be solved, numerically or analytically, for the imaginary part of $\mathfrak{h}$ at any parameter values. Taking its imaginary part then

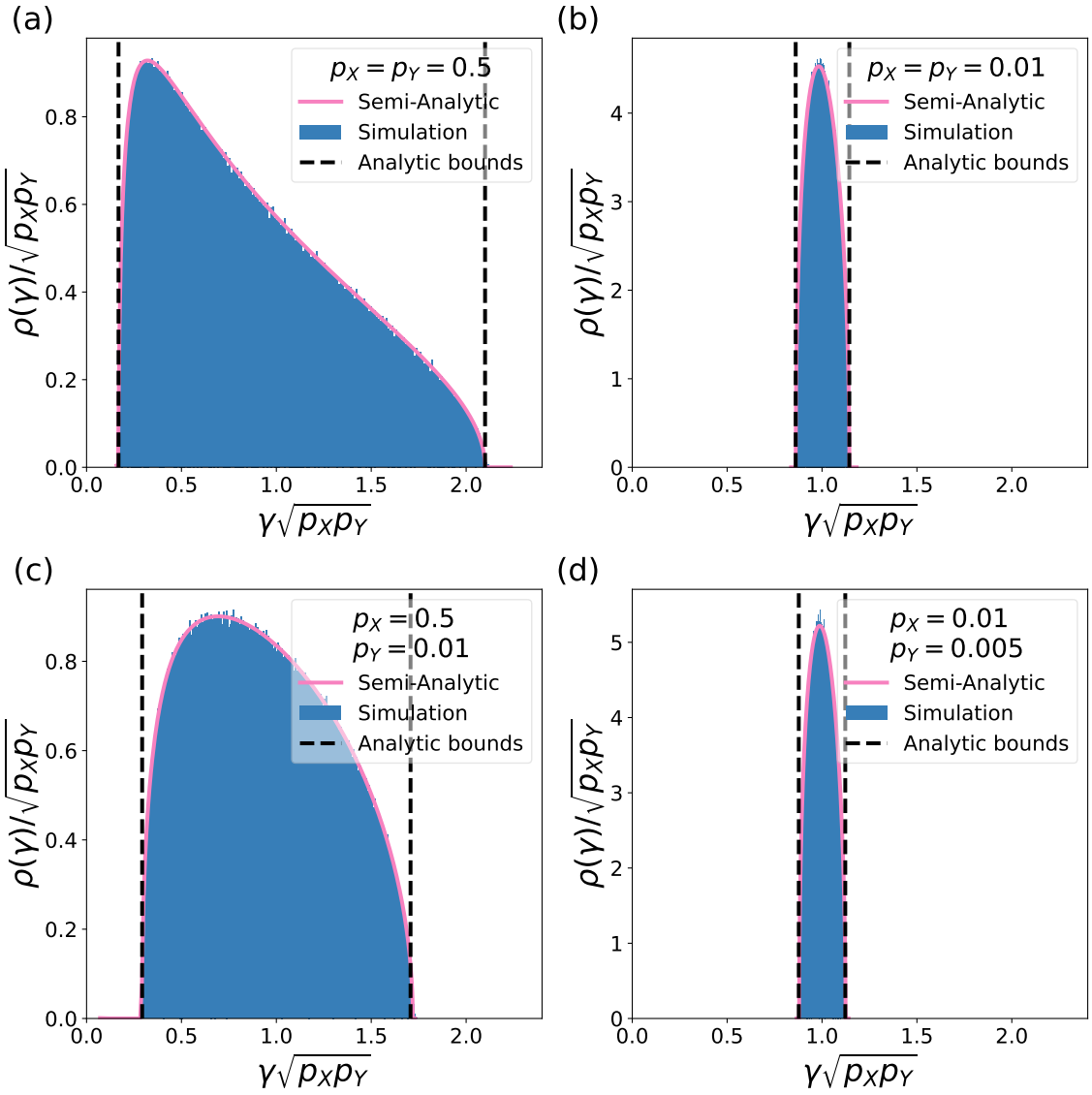

Figure 3.1: Distribution of nonzero eigenvalues scaled by a factor $\sqrt{p_X p_Y}$ for $T < N_X,\ N_Y$ for $T < N_X,\ N_Y$. **(a)** $p_X = p_Y = 0.5$, **(b)** $p_X = p_Y = 0.01$, **(c)** $p_X = 0.5$, $p_Y = 0.01$, and **(d)** $p_X = 0.01$, $p_Y = 0.05$. The blue bars are the histogram of the simulated data. The magenta curve is computed from the numerical solution of the exact cubic equation for the Stiltjes transform. The black dotted lines show edges of the nonzero part of the density in simplifying limits, evaluated analytically. Here, $T = 1000$, and the the numerical simulation for spectrum consists of 500 independent model realizations.

gives us the density of nonzero eigenvalues.

Here, we solve the equation numerically (which we refer to as the "semi-analytic" solution, since it solves numerically the analytical expression, Eq. (3.13)), and study the spectrum for a variety of parameter regimes. The spectrum has compact support, showing a single band of eigenvalues with upper and lower edges. The edges can be calculated by finding the condition under which the the discriminant of the cubic equation, Eq. (3.13), becomes zero. To get easily interpretable formulas for the edges $\lambda_\pm$ (and hence $\gamma_\pm$), we take various simplifying limits where the discriminant equation for the cubic polynomial is exactly solvable.

For the case where $p_X = p_Y$ (same-size data matrices), the bounds for the nonzero singular value density then become

$$\gamma_\pm = \sqrt{\frac{8p_X^2 + 20p_X^3 - p_X^4 \pm p_X^{5/2}(8+p_X)^{3/2}}{8p_X^4}}. \tag{3.18}$$

Assuming $p_X = p_Y \to 0$ (so that we are in the severely undersampled regime, where the number of samples is *much* smaller than the number of dimensions in X and Y), the edge values become

$$\gamma_\pm \approx \frac{1}{p_X}(1 \pm \sqrt{2p_X}). \tag{3.19}$$

For the case where $p_Y = \epsilon p_X$, where $\epsilon \to 0$, but $p_X = O(1) < 1$ the bounds are

$$\gamma_\pm \approx \sqrt{\frac{1 + p_X \pm 2\sqrt{p_X}}{\epsilon p_X^2}}. \tag{3.20}$$

Finally, for $p_Y = \alpha p_X$, where $p_X \to 0$ and $\alpha < 1$ (both X and Y are extremely undersampled, but unequal in size), the bounds are

$$\gamma_\pm \approx \frac{1 \pm \sqrt{p_Y + p_X}}{\sqrt{p_Y p_X}}. \tag{3.21}$$

We see that, in all of these limits, the center of the singular value distribution is approximately the geometric mean of the inverse aspect ratios, $\sqrt{\frac{1}{p_X p_Y}} = \sqrt{q_X q_Y}$. This sets the typical scale of sampling noise singular values at a given sample size T. The noise eigenvalues of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}/T$ and $\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}/T$ individually scale like q_x and q_y [32]. Thus, this scaling is plausible if each eigendirection is poorly-sampled enough that they can be found to correlate with each other by chance.

Figure 3.1 compares our analytical results to numerical simulations for the density of singular values γ of $\mathbf{C}$. We scale the singular values by the scale factor $\sqrt{\frac{1}{p_X p_Y}}$. We see that the semi-analytic solution for the density is in excellent agreement with our numerical results. Further, we see that the analytical solutions for the bounds, in appropriate limits, also agree well with simulations.

The simulations and the semi-analytic solutions also agree for other parameter values where simple analytic bounds for the edges could not be evaluated exactly (see Section 3.5).

3.3.2 Spectrum of empirical cross covariance matrix when $N_Y \geq T \geq N_X$

Solving for the roots of the cubic polynomial in Eq. (3.13) and taking its imaginary part again gives us the density of nonzero eigenvalues.

In this case, we can evaluate the edges of the spectrum exactly in the limit $p_Y = \epsilon p_X$, where $\epsilon \to 0$, and $p_X = O(1) \geq 1$. In this case, the bounds are

$$\gamma_\pm = \sqrt{\frac{1 + p_X \pm 2\sqrt{p_X}}{\epsilon p_X^2}}. \tag{3.22}$$

This limit is the same as in the case when $T \leq N_X,\ N_Y$ and $N_X \ll N_Y$. Though the rank of the matrix is now N_X instead of T (in the case when $T \leq N_X,\ N_Y$).

Figure 3.2 shows that the semi-analytic solution for the density, and the analytic

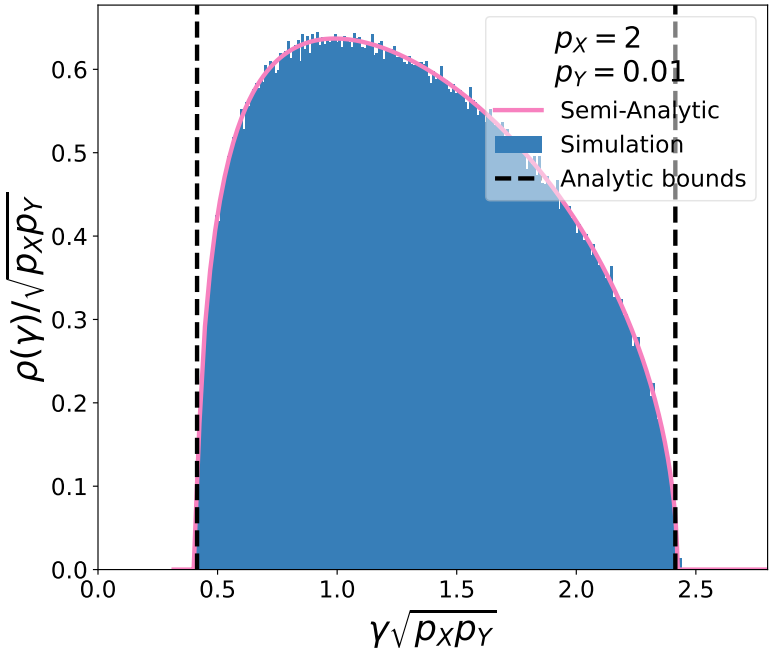

Figure 3.2: Distribution of nonzero eigenvalues for $N_Y \geq T \geq N_X$, specifically, $p_X = 2$, $p_Y = 0.01$. Plotting conventions are the same as in Fig. 3.1. Here, again, $T = 1000$, and the numerical simulation for spectrum consists of 500 independent model realizations.

solution for the edges, match our numerical simulations in this case as well.

The tail which lie outside of the analytic equation to the right of the density plot in Fig 3.2 follows the Tracy-Widom distribution [94]. This is because all the outliers are because of the largest eigenvalue, and the largest eigenvalue of a product of Wishart matrices follows Tracy-Widom distribution as well.

3.3.3 Spectrum of empirical cross covariance matrix for $T > N_X, N_Y$

Solving for the roots of the cubic polynomial, Eq. (3.13), and taking its imaginary part again gives us the density of nonzero eigenvalues. We then obtain simplified formulas for $\gamma_\pm$ in limiting cases.

For the case where $p_X = p_Y$, the discriminant of the cubic equation for $\mathfrak{h}$ is a 5^{th}-order polynomial with three zero solutions and two nonzero solutions, given by $z_\pm = \frac{8p_X^2+20p_X^3-p_X^4 \pm p_X^{5/2}(8+p_X)^{3/2}}{8p_X^4}$. Now because $z_- < 0$ and the squares of singular values are always positive, the upper bound of the non-zero density is z_+ but the lower bound is 0. Thus the bounds for the nonzero eigenvalue density are

$$\gamma_+ = \sqrt{\frac{8p_X^2 + 20p_X^3 - p_X^4 + p_X^{5/2}(8+p_X)^{3/2}}{8p_X^4}}, \quad \gamma_- = 0. \tag{3.23}$$

In the limit $p_X \gg 1$ (extremely good sampling), this simplifies to $\gamma_+ \approx \sqrt{\frac{3}{2p_X}} = \sqrt{\frac{3q_X}{2}}$. Thus, in this limit the scaling of the edges agrees with those for the cross-correlations of whitened variables evaluated in Ref. [88], where $\gamma_+ = 2\sqrt{q_X}$, and $\gamma_- = 0$. Note, however, that the exact value of the upper edge is different for the whitened cross-correlation matrices, because the self-covariances used for whitening also fluctuate.

Figure 3.3 shows that these limiting formulas for the edges, and the semi-analytic solution for the spectrum match numerical simulations.

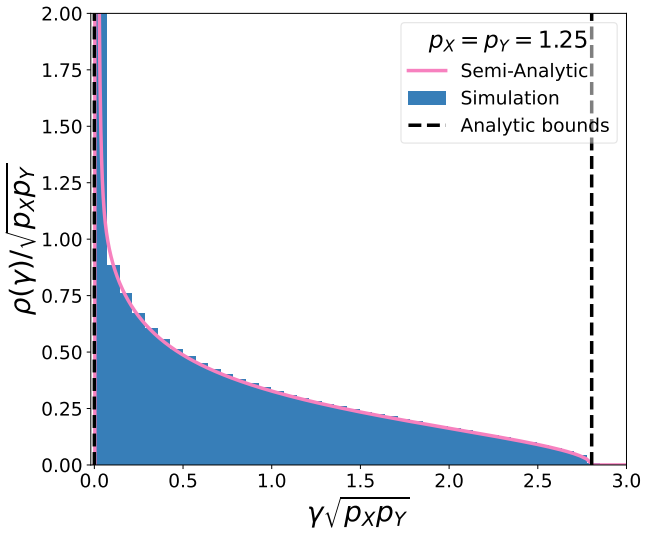

Figure 3.3: Distribution of nonzero eigenvalues for $T > N_X, N_Y$, specifically $p_X = p_Y = 1.25$. Plotting conventions are the same as in Fig. 3.1. Here, again, $T = 1000$, and numerical simulation for spectrum consists of 500 independent model realizations.

3.4 Discussion

We have used random matrix theory to calculate the density of singular values of normalized cross-correlation matrices. Further, in simplifying limits, we were able to obtain simple, exact formulas for the edges of the spectrum.

In all cases, the scale of the non-zero singular values is given roughly by $1/\sqrt{p_X p_Y} = \sqrt{N_X N_Y}/T$. Thus, the noise, unsurprisingly, decreases as more samples are collected, relative to the dimensions of the two observed variables. More surprisingly, however, this calculation in fact suggests that the cross-covariance can sometimes be used to detect a signal which is not detectable from either the covariance of X or that of Y alone, as recently observed numerically [95].

To see this, consider a naïve protocol for establish a correlation between high-dimensional X and Y: we first search for a low-dimensional signal in X (e.g., using principle component analysis), then search for a low-dimensional signal in Y, and finally correlate the low-dimensional signals. The edges of the empirical covariance

spectra of X and Y are of order $1/p_X$ and $1/p_Y$, respectively. Thus, a shared signal which has $O(1)$ magnitude in both X and Y will correspond to an outlier eigenvalue outside of the spectrum, and hence can be detected if $T > N_X, N_Y$. But if $N_Y > T > N_X$ (one variable is well sampled, and one variable is poorly sampled), the signal in Y cannot be detected. Since the noise spectrum of $\mathbf{C}$ depends on the geometric mean $\sqrt{p_X p_Y}$, however, the same signal may be detectable in $\mathbf{C}$, if X is sampled well enough to "make up for" the poor sampling of Y. Making this rough analysis precise requires a full calculation of the spectrum of a model with both a signal and noise, which we will present in a future work.

These results also suggest that a sufficiently strong signal can be detected even if $T < N_X, N_Y$.

In the limit $T \gg N_X, N_Y$, where the covariances of X and Y are both well sampled, the edges of the spectrum have the same scaling with aspect ratio (sample size) as those for the whitened cross-correlation matrix [88]. Thus, in this extremely well sampled limit, the cross-corelation and cross-covariance matrices can both be used to detect a signal. However, the prefactor of this scaling is smaller for the cross-covariance matrix, indicating that whitening using the inverse of the empirically sampled self-covariance matrices introduces additional noise in the spectrum. Further, for sparse data, the cross-correlation cannot be evaluated—even if only one of the two variables is undersampled, where our results suggest that a signal may still be detectable in the cross-covariance. Together, these results suggest that in many cases the cross-covariance may be the most effective tool for detecting the shared signal in a pair of high-dimensional observations.

3.5 Calculating the spectrum of the empirical cross-covariance matrix

Here we calculate the spectrum of the $N_X \times N_X$ normalized empirical cross-covariance matrix $\mathbf{C}^T\mathbf{C}$, given by Eq. (3.4). Given $N_X,\ N_Y,\ T \gg 1$, this spectrum can be evaluated using random matrix theory. Parts of this calculation, can be mapped onto previous calculations [91, 92, 93] by reinterpreting the meaning of various various. However, for completeness, we present a full, self-contained calculation here.

The nonzero eigenvalues of the NECCM $\mathbf{C}^T\mathbf{C}$ are the same as those of the matrix

$$\mathbf{H} = \frac{1}{\sigma_X^2\sigma_Y^2T^2}\left(\mathbf{X}\mathbf{X}^T\right)\left(\mathbf{Y}\mathbf{Y}^T\right) \tag{3.24}$$

$$= \frac{N_XN_Y}{T^2}W_{X^T}W_{Y^T} \tag{3.25}$$

$$= \frac{1}{p_Xp_Y}W_{X^T}W_{Y^T}. \tag{3.26}$$

Here $\mathbf{W_X}$ and $\mathbf{W_Y}$ are normalized Wishart matrices, given by

$$\mathbf{W_Y} = \frac{1}{T\sigma_Y^2}\mathbf{Y}^T\mathbf{Y}\,, \tag{3.27}$$

and similar for X. Crucially, $\mathbf{W_X}$ and $\mathbf{W_Y}$ are free matrices (the appropriate generalization of independence to noncommuting objects, such as matrices).

The spectrum of $\mathbf{H}$, $\rho_{\mathbf{H}}$, can be evaluated from its Stiltjes transform,

$$\mathfrak{h}(z) \equiv \mathfrak{h}_{\mathbf{H}}(z) \equiv \lim_{T\to\infty}\frac{1}{T}\mathrm{Tr}(z\mathbf{I}-\mathbf{H})^{-1}, \tag{3.28}$$

using the formula

$$\rho_{\mathbf{H}}(\lambda) = \frac{1}{\pi}\lim_{\eta\to 0^+}\Im\mathfrak{h}(z=\lambda-i\eta)\,, \tag{3.29}$$

To evaluate this Stiltjes transform, we must introduce the $\mathcal{T}$ and $\mathcal{S}$ transforms, which are useful for evaluating the Stiltjes transform of products of random matrices [87]. Their properties used in further calculations are summarized below.

The $\mathcal{T}$ transform of a matrix A is defined as

$$\mathcal{T}_{\mathbf{A}}(z) = z\mathfrak{g}_{\mathbf{A}}(z) - 1. \tag{3.30}$$

The $\mathcal{T}$ transform, in turn, is used to define the $\mathcal{S}$ transform:

$$\mathcal{S}_{\mathbf{A}}(t) = \frac{t+1}{t\mathcal{T}_{\mathbf{A}}^{-1}(t)}. \tag{3.31}$$

For free matrices $\mathbf{A}$ and $\mathbf{B}$, the $\mathcal{S}$-transform of a product is multiplicative:

$$\mathcal{S}_{\mathbf{AB}}(t) = \mathcal{S}_{\mathbf{A}}(t)\mathcal{S}_{\mathbf{B}}(t)\,. \tag{3.32}$$

Furthermore, for a scalar a,

$$\mathcal{S}_{a\mathbf{A}}(t) = a^{-1}\mathcal{S}_{\mathbf{A}}(t)\,. \tag{3.33}$$

To derive the Stieltjes transform of $\mathbf{H}$, we first evaluate its $\mathcal{S}$ transform. Using Eq. (3.32) and Eq. (3.33), we write

$$\mathcal{S}_{\mathbf{H}}(t) = \mathcal{S}\left(\frac{1}{p_X p_Y} W_{X^T} W_{Y^T}\right) \tag{3.34}$$

$$= p_X p_Y \mathcal{S}_{W_{X^T}} \mathcal{S}_{W_{Y^T}}. \tag{3.35}$$

The $\mathcal{S}$-transform of a Wishart matrix is well known [87]:

$$\mathcal{S}_{W_{X^T}} = \frac{1}{1+p_X t}. \tag{3.36}$$

Now, plugging in the relevant terms for $\mathcal{S}_{W_{X^T}}$ and $\mathcal{S}_{W_{Y^T}}$ in Eq. (3.34) using Eq. (3.36), we obtain:

$$\mathcal{S}_{\mathbf{H}}(t) = \frac{p_X}{1+p_X t}\frac{p_Y}{1+p_Y t}. \tag{3.37}$$

To calculate the spectral density of the matrix of interest we replace the $\mathcal{S}$-transform in Eq. (3.37) with the corresponding $\mathcal{T}$-transform by using the relationship in Eq. (3.31):

$$\mathcal{T}^{-1}_{\mathbf{H}}(t) = \frac{t+1(1+p_X t)(1+p_Y t)}{t p_X p_Y}. \tag{3.38}$$

We now solve the equation for the functional inverse, $\mathcal{T}^{-1}(\mathcal{T}(z)) = z$, using the definition of the $\mathcal{T}$-transform, Eq. (3.30). This gives a cubic equation for the Stieltjes transform:

$$\mathfrak{h}^3 z^2 p_X p_Y + \mathfrak{h}^2 z\left(p_Y(1-p_X)+p_X(1-p_Y)\right) \quad +\mathfrak{h}\left((1-p_X)(1-p_Y)-z p_X p_Y\right)+p_X p_Y = 0 \tag{3.39}$$

The imaginary part of the roots of the cubic equation give us the density of eigenvalues. The edges of the band $[\lambda_-, \lambda_+]$, for which the density is nonzero, are obtained from the zeros of the discriminant of the cubic equation. Thus Eq. 3.39 is of the form

$$a\mathfrak{h}^3 + b\mathfrak{h}^2 + c\mathfrak{h} + d = 0, \tag{3.40}$$

the discriminant is

$$D = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd, \tag{3.41}$$

where:

$$a = z^2 p_X p_Y, \tag{3.42}$$

$$b = z\left(p_Y(1-p_X) + p_X(1-p_Y)\right), \tag{3.43}$$

$$c = \left((1-p_X)(1-p_Y) - z p_X p_Y\right), \tag{3.44}$$

$$d = p_X p_Y. \tag{3.45}$$

The density $\rho(\lambda)$ and the edges $\lambda_\pm$ must then be transformed into the density of singular values $\rho(\gamma)$ and the edges $\gamma_\pm$. For this, to get the spectrum of the nonzero part of the SVD of $\mathbf{C}$, we use:

$$\rho_A(z) = 2z\rho_{A^2}(z^2), \tag{3.46}$$

and the edges obey $\gamma_\pm = \sqrt{\lambda_\pm}$.

3.5.1 Spectrum of the empirical cross covariance matrix for $T < N_X,\ N_Y$

Simplified solutions for $p_X = p_Y$

For $p_X = p_Y$, the cubic equation for the Stieltjes transform, Eq. (3.39), reduces to:

$$\mathfrak{h}^3 z^2 {p_X}^2 + \mathfrak{h}^2 z\left(p_X(1-p_X) + p_X(1-p_X)\right) \tag{3.47}$$

$$+ \mathfrak{h}\left((1-p_X)(1-p_X) - z{p_X}^2\right) + p_X^2 = 0\,, \tag{3.48}$$

and the discriminant (Eq. 3.41) simplifies to

$$D = (4p_X^4 - 12p_X^5 + 12p_X^6 - 4p_X^7)z^3 \tag{3.49}$$

$$+ (-8p_X^6 - 20p_X^7 + p_X^8)z^4 + 4p_X^8 z^5. \tag{3.50}$$

Solving Eq. (3.50) for zeros we find that there are three zeros at $z = 0$ and two zeroes at $\lambda_\pm$. In between these values, the discriminant is negative and thus the solution for $\mathfrak{h}$ has a nonzero imaginary part, giving a nonzero density of eigenvalues. We obtain:

$$\lambda_\pm = \frac{8p_X^2 + 20p_X^3 - p_X^4 \pm p_X^{5/2}(8+p_X)^{3/2}}{8p_X^4}. \tag{3.51}$$

For $p_X \to 0$, Eq. (3.51) reduces to

$$\lambda_\pm = \frac{1}{p_X^2} \pm \frac{2\sqrt{2}}{p_X^{3/2}}. \tag{3.52}$$

The singular values of $\mathbf{C}$ have nonzero density between $\gamma_\pm$, where $\gamma_\pm = \sqrt{\lambda_\pm}$. Thus, for small p_X,

$$\gamma_\pm = \sqrt{\frac{1}{p_X^2} \pm \frac{2\sqrt{2}}{p_X^{3/2}}} = \frac{1}{p_X}\sqrt{1 \pm 2\sqrt{2}\sqrt{p_X}} \tag{3.53}$$

$$\approx \frac{1}{p_X}(1 \pm \sqrt{2p_X}). \tag{3.54}$$

Simplified solutions for $p_X < 1$, $p_Y \ll p_X$

For $p_Y = \epsilon p_X$ under the condition $\epsilon \to 0$, the cubic equation for the Stieltjes transform Eq. (3.39) reduces to:

$$\epsilon\mathfrak{h}^3 z^2 p_X^2 + \mathfrak{h}^2 z p_X\left(\epsilon(1-p_X) + (1-\epsilon p_X)\right) \tag{3.55}$$

$$+ \mathfrak{h}\left((1-p_X)(1-\epsilon p_X) - z\epsilon p_X^2\right) + \epsilon p_X^2 = 0\,. \tag{3.56}$$

The discriminant of Eq. (3.55) is calculated using Eq. (3.41). We then organize this discriminant as a polynomial in z, giving

$$D = 4z^5\epsilon^4 p_X^8 + z^4(\epsilon^2 p_X^6 + \epsilon^3(-10p_X^6 - 10p_X^7)$$

$$+\epsilon^4(p_X^6-10p_X^7+p_X^8))+z^3(\epsilon(-2p_X^4-2p_X^5)$$
$$+\epsilon^2(8p_X^4-4p_X^5+8p_X^6)+\epsilon^3(-2p_X^4-4p_X^5-4p_X^6-2p_X^7)$$
$$+\epsilon^4(-2p_X^5+8p_X^6-2p_X^7))+z^2(p_X^2-2p_X^3+p_X^4$$
$$+\epsilon(-2p_X^2+2p_X^3+2p_X^4-2p_X^5)$$
$$+\epsilon^2(p_X^2+2p_X^3-6p_X^4+2p_X^5+p_X^6)$$
$$+\epsilon^3(-2p_X^3+2p_X^4+2p_X^5-2p_X^6)+\epsilon^4(p_X^4-2p_X^5+p_X^6)). \quad (3.57)$$

Each term is of the form $f_n(\epsilon)z^n$. As $\epsilon \to 0$, we may expand each $f_n(\epsilon)$ to the lowest nontrivial order in ϵ. Collecting the lowest-order terms for each power of z, the discriminant in Eq. (3.57) reduces to:

$$D \quad \approx \quad z^2\left[p_X^2(1-p_X)^2-2(p_X^4+p_X^5)\epsilon z+p_X^6\epsilon^2z^2 \quad + \quad 4p_X^8\epsilon^4z^3\right]. \quad (3.58)$$

We seek positive roots $z_\pm(\epsilon)$ of the right-hand group of terms (the equation has a single negative root, but since the eigenvalues of $\mathbf{H}$ are positive by construction, this corresponds to a spurious root of the equation for $\mathfrak{h}$). This requires cancellation of at least two terms. That is, at least two terms of opposite signs must be of the same order in ϵ. We see that this can only happen if $z \sim \epsilon^{-1}$ or $z \sim \epsilon^{-3/2}$. In both of these possible cases, the final term is subleading and can be neglected. Thus, in this limit, we seek the roots of

$$D \approx (p_X^2-2p_X^3+p_X^4)z^2-2(p_X^4+p_X^5)z^3\epsilon+z^4p_X^6\epsilon^2. \quad (3.59)$$

We solve Eq. (3.59) for zeros. The 4^{th}-order equation has four zeroes. Two of the zeros are $z=0$, and the other two, $\lambda_\pm$, are

$$\lambda_\pm = \frac{1+p_X \pm 2\sqrt{p_X}}{\epsilon p_X^2} \quad (3.60)$$

$$\approx \frac{1 + p_X \pm 2\sqrt{p_X}}{p_X p_Y}. \tag{3.61}$$

Thus the density of eigenvalues for for SVD of $\mathbf{C}$ will be nonzero between $\gamma_\pm = \sqrt{\lambda_\pm}$, such that

$$\gamma_\pm \approx \sqrt{\frac{1 + p_X \pm 2\sqrt{p_X}}{p_X p_Y}}. \tag{3.62}$$

Simplified solutions for $p_X, p_Y \ll 1$

For $p_Y = \alpha p_X$ under the condition $p_X \to 0$ and $\alpha < 1$, the cubic equation for the Stieltjes transform Eq. (3.39) reduces to:

$$\alpha \mathfrak{h}^3 z^2 p_X^2 + \mathfrak{h}^2 z p_X (\alpha + 1) + \mathfrak{h}\left(1 - z\alpha p_X^2\right) + \alpha p_X^2 = 0\,. \tag{3.63}$$

The discriminant of Eq. (3.63) is calculated using Eq. (3.41). Written as a polynomial in z, it is

$$\begin{aligned} D = 4z^5\alpha^4 p_X^8 + z^4(\alpha^2 p_X^6 - 10\alpha^3 p_X^6 + \alpha^4 p_X^6 - 18\alpha^3 p_X^7 \\ - 18\alpha^4 p_X^7 - 27\alpha^4 p_X^8) + z^3(-2\alpha p_X^4 + 8\alpha^2 p_X^4 - 2\alpha^3 p_X^4 \\ - 4\alpha p_X^5 + 6\alpha^2 p_X^5 + 6\alpha^3 p_X^5 - 4\alpha^4 p_X^5) \\ + z^2(p_X^2 - 2\alpha p_X^2 + \alpha^2 p_X^2). \end{aligned} \tag{3.64}$$

As $p_X \to 0$, the contribution of higher-order terms for each power of z to the final solution will be negligible. Collecting the lowest order terms in p_X for each power of z, the discriminant in Eq. (3.64) reduces to

$$\begin{aligned} D = 4z^5\alpha^4 p_X^8 + z^4(\alpha^2 p_X^6 - 10\alpha^3 p_X^6 + \alpha^4 p_X^6) \\ + z^3(-2\alpha p_X^4 + 8\alpha^2 p_X^4 - 2\alpha^3 p_X^4) + \end{aligned}$$

$$z^2(p_X^2 - 2\alpha p_X^2 + \alpha^2 p_X^2). \quad (3.65)$$

We solve Eq. (3.65) for zeros. The 5^{th}-order equation has 5 zeroes (counting their multiplicities). Two of the zeroes are at $z = 0$, one is at $z = \frac{-(1-\alpha)^2}{4\alpha^2 p_X^2} < 0$. Thus, the other two are $\lambda_\pm$. Taking the condition $D < 0$, we find that nonzero density requires $\lambda \in [\lambda_-, \lambda_+]$. In particular, we find the solution

$$\lambda_\pm = \frac{1 \pm 2\sqrt{p_X(1+\alpha)}}{\alpha p_X^2}. \quad (3.66)$$

The nonzero density of eigenvalues for SVD of $\mathbf{C}$ will be between $\gamma_\pm = \sqrt{\lambda_\pm}$:

$$\gamma_\pm = \sqrt{\lambda_\pm} = \sqrt{\frac{1 \pm 2\sqrt{p_X(1+\alpha)}}{\alpha p_X^2}} \quad (3.67)$$

$$= \sqrt{\frac{1 \pm 2\sqrt{p_X + p_Y}}{p_Y p_X}} \quad (3.68)$$

$$\approx \frac{1 \pm \sqrt{p_Y + p_X}}{\sqrt{p_Y p_X}}. \quad (3.69)$$

3.5.2 Spectrum of the empirical cross covariance matrix when $T > N_X, N_Y$

Simplified solutions for $p_X = p_Y$

For $p_X = p_Y$, the discriminant takes the same form as in the case $T < N_X, N_Y$ (Eq. 3.50). That is,

$$D = (4p_X^4 - 12p_X^5 + 12p_X^6 - 4p_X^7)z^3 + (-8p_X^6 + 20p_X^7 + p_X^8)z^4 + 4p_X^8 z^5. \quad (3.70)$$

In this case, however, the identities of the roots that determine $\lambda_\pm$ are different from the case $T < N_X, N_Y$. Specifically, Eq. (3.70) has three zeros at $z = 0$ and one zero each at $z_- = \frac{8p_X^2+20p_X^3-p_X^4-p_X^{5/2}(8+p_X)^{3/2}}{8p_X^4}$ and $z_+ = \frac{8p_X^2+20p_X^3-p_X^4+p_X^{5/2}(8+p_X)^{3/2}}{8p_X^4}$. The root $z_- < 0$, and the squares of the singular values are always non-negative. Thus the lower edge is $\lambda_- = 0$, and $\lambda_+ = z_+ = \frac{8p_X^2+20p_X^3-p_X^4+p_X^{5/2}(8+p_X)^{3/2}}{8p_X^4}$. Thus the upper edge of the SVD spectrum is $\gamma_+ = \sqrt{\lambda_+}$. For $p_X \gg 1$,

$$\lambda_+ \approx -\frac{1}{8} + \left(\frac{(8+p_X)^{3/2}}{8p_X^{3/2}}\right) \tag{3.71}$$

$$\approx \frac{3}{2p_X}. \tag{3.72}$$

Thus,

$$\gamma_+ = \sqrt{\frac{3}{2p_X}}. \tag{3.73}$$

Chapter 4

Statistical properties of spiked joint covariance and cross covariance matrices

4.1 Introduction

[1]Recent experiments measure increasingly large numbers of variables simultaneously, giving rise to extraordinarily large datasets. Examples include recordings from populations of neurons [96, 97], movies of animal postures [98, 99], 'omics datasets [100, 101], collective behavior [102], particle positions in soft matter systems [13], ecological data [103], etc. In many of these cases, one wants to understand the relationship between two high-dimensional variables—e.g., neural activity and behavior, or gene expression and cellular phenotypes. Such correlations are inferred from the singular value decomposition of the cross covariance matrix. In order to determine

[1]The work in this chapter was conducted in collaboration with Drs. Sean Alexander Ridout and Ilya Nemenman. I performed all simulations, conducted all analyses, and led writing of the manuscript. Dr. Nemenman conceived the model and led the project, while Dr. Ridout contributed to discussions regarding the calculations, procedures and analyses. All authors participated in writing and reviewed the chapter which will become the final manuscript.

whether a given singular value corresponds to a true signal or merely to sampling noise, we must first understand the singular value spectra produced by uncorrelated data due to finite-sampling effects (see previous Chapter), and how these sampling effects affect the detectability of signals. Random Matrix Theory (RMT) provides an interpretable framework that allows us to do this.

Data will be modeled as T independent samples of the state of the system, specified through $N_X \gg 1$ variables x_i, $i = 1 \dots N_X$, and $N_Y \gg 1$ variables y_i, $i = 1 \dots N_Y$. Thus, the data comprise two matrices $\mathbf{X}$ and $\mathbf{Y}$ with dimensionality $T \times N_X$ and $T \times N_Y$, respectively. RMT has previously been used to study finite-sampling induced correlations within $\mathbf{X}$ or $\mathbf{Y}$ separately and for calculating the conditions on the strength and the structure of the signal that would allow the signal to be distinguishable from the spurious correlations emerging due to sampling fluctuations. Such calculations were made for different models, including the latent features model [91], and the 'spiked covariance matrices' model [104, 105, 106, 107]. The former is a generative model, which specifies the distribution of data as coming from a combination of signal and noise. The latter assumes that the sample covariance matrices, e.g. $\mathbf{X}^T\mathbf{X}$, have a low-dimensional signal contribution, a 'spike', without providing a generative model for how the said spike appears.

Similar RMT-based analyses for cross covariance, which would analyze when a shared signal between $\mathbf{X}$ and $\mathbf{Y}$ can be detected from the two data matrices together, are largely missing. Some attempts have been made to study the detectability of a signal in the cross covariance by considering a concatenation of $\mathbf{X}$ and $\mathbf{Y}$ matrices into a single matrix $\mathbf{Z}$ [108, 109]. Then the cross covariance estimation is a sub-problem of covariance estimation of $\mathbf{Z}$, so that much of the previous work on covariance within individual data modalities applies. However, to use these methods, one must make strong assumptions about the structure of the covariance matrix like both $\mathbf{X}$ and $\mathbf{Y}$ having the same covariance for the pure noise-noise covariance, so that the methods

have only limited applications. A complementary approach is to whiten the variables $\mathbf{X}$ and $\mathbf{Y}$, so that there are no within-modality correlations remaining, and all correlations thus represent the cross-correlations between the two data types [88, 89, 90]. These approaches are powerful if $T > \{N_X, N_Y\}$, so that whitening is possible. However, understanding the limits of detectability of shared signals in datasets remains rudimentary in the opposite regime of a small number of samples.

In the context of spiked models, where a signal perturbation is added directly to the otherwise sampling-noise induced covariance matrix, the Johnstone model deserves special attention [34]. It was found that the largest eigenvalue undergoes a phase transition [35] (the BBP phase transition) as the spike strength is varied. As the spike strength increases, there is a critical threshold where the top eigenvector of the sample covariance matrix starts to align with the spike added to the covariance matrix as a perturbation. If the strength of the spike is smaller than the critical value, then the largest eigenvalue of the spiked sample covariance matrix will be the largest eigenvalue of the bulk eigenvalue spectrum (and hence is not an outlier), and the corresponding sample eigenvector will be delocalized (or be effectively random). If the strength of the spike is larger than the critical value, then the associated eigenvalue will jump out of the bulk eigenvalue spectrum induced by the sampling noise, and the outlier sample eigenvector will have a nonzero overlap with the spike (signal). Similar results have been derived for different structures of distribution of bulk covariance, as well as for a multiplicative spike [110, 111, 112, 113, 114]. However, currently no models exist to understand the effect of additive perturbations in cross-correlation matrices.

Here, we define a model that can be used to study (i) the effect of an additive spike on individual covariances of $\mathbf{X}$ and $\mathbf{Y}$ with different strengths of spikes associated with $\mathbf{X}$ and $\mathbf{Y}$, (ii) the effect of this spike on the eigenspectrum of $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, and finally (iii) the effect of the same spike on the spectrum of the cross covariance $\mathbf{X}^T\mathbf{Y}$. We

calculate analytically the critical spike strength to detect an outlier eigenvalue in all three cases and calculate the overlap of the largest eigenvectors with the true signal vector.

We show that, for certain parameter values, a shared signal that *cannot* be detected using in the eigenvalue spectrum of either $\mathbf{X}$ or $\mathbf{Y}$ *can* be detected as an outlier in the eigenvalue spectrum of the concatenated data $\mathbf{Z}$ covariance, $\mathbf{Z}^T\mathbf{Z}$, or as an outlier in the singular value spectrum of the cross covariance $\mathbf{X}^T\mathbf{Y}$. We further analyze under which conditions outliers are easier to detect in the spectra of $\mathbf{Z}^T\mathbf{Z}$ or of $\mathbf{X}^T\mathbf{Y}$.

4.2 Model and methods

We are interested in understanding the ability of different methods to detect a shared signal between two large dimensional datasets as an outlier from the bulk of finite sampling-induced correlations. For this, we first set up spiked covariance matrix models for all of these cases, as is a standard approach in the literature, so that the corresponding covariance matrices have a deterministic low-rank additive contribution [34, 35], in addition to the sampling-induced structure. These spikes are constructed to be equivalent, representing the same signal in each case. We then compare the limits of detectability of these low-rank spikes along all models.

We represent data as matrices, $\mathbf{X}$ and $\mathbf{Y}$, with dimensions $T \times N_X$ and $T \times N_Y$, respectively. To study correlations due to sampling only, we assume that the entries of $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated Gaussian random variables with zero mean and variances σ_X^2 and σ_Y^2, respectively.

$$X_{t\mu} \sim \mathcal{N}(0, \sigma_X^2)\,, \quad Y_{t\nu} \sim \mathcal{N}(0, \sigma_Y^2)\,, \tag{4.1}$$

$$t = 1, \ldots, T,\ \mu = 1, \ldots, N_X,\ \nu = 1, \ldots, N_Y. \tag{4.2}$$

In what follows, we will often assume $\sigma_X^2 = \sigma_Y^2 = 1$ for simplicity, but this does not

result in the loss of generality as each of the variables can be normalized easily by its empirical variance.

We additionally define $\mathbf{Z}$ as the concatenation $(\mathbf{X}, \mathbf{Y})$.

$$\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \,. \tag{4.3}$$

The dimensionality of $\mathbf{Z}$ is then $T \times (N_X + N_Y)$.

We consider a spike—a low-rank perturbation—in the concatenated covariance, which points in the direction $\hat{\mathbf{u}}$ in X and $\hat{\mathbf{v}}$ in Y. The vectors ($\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$) are $1 \times N_X$ and $1 \times N_Y$ dimensional random (but quenched) unit vectors, respectively. The magnitude of the spike is a and b in the X and Y spaces, respectively. We further define $c^2 = a^2 + b^2$, so that the direction of the joint spike in the Z space corresponds to $\hat{\mathfrak{z}} = \left(\frac{a}{c}\hat{\mathbf{u}}, \frac{b}{c}\hat{\mathbf{v}}\right)$, and the spike's magnitude is c. With this choice of variables, the vector $\hat{\mathfrak{z}}$ vector is a random unit vector with dimensionality $1 \times (N_X + N_Y)$. The spiked covariance model for the $\mathbf{Z}^T\mathbf{Z}$ data is given by

$$\mathbf{H}_{\mathbf{Z^T Z}} = \frac{\mathbf{Z^T Z}}{T} + \left(a^2 + b^2\right) \hat{\mathfrak{z}}^T \hat{\mathfrak{z}}. \tag{4.4}$$

The first term in r. h. s. of Eq. (4.4) can be written in a block form as

$$\frac{\mathbf{Z^T Z}}{T} = \begin{bmatrix} \frac{\mathbf{X^T X}}{T} & \frac{\mathbf{X^T Y}}{T} \\ \frac{\mathbf{Y^T X}}{T} & \frac{\mathbf{Y^T Y}}{T}. \end{bmatrix} \tag{4.5}$$

In other words, this term consists of the square X and Y self-covariance matrices, as well as their rectangular cross covariance.

A similar block-form of the second term, the spike, in Eq. (4.4) is:

$$\left(a^2 + b^2\right) \hat{\mathfrak{z}}^T \hat{\mathfrak{z}} = \begin{bmatrix} a^2 \hat{\mathbf{u}}^T \hat{\mathbf{u}} & ab \hat{\mathbf{u}}^T \hat{\mathbf{v}} \\ ab \hat{\mathbf{v}}^T \hat{\mathbf{u}} & b^2 \hat{\mathbf{v}}^T \hat{\mathbf{v}} \end{bmatrix}. \tag{4.6}$$

Combining these expressions, we write:

$$\mathbf{H_{Z^TZ}} = \frac{\mathbf{Z^TZ}}{T} + \left(a^2 + b^2\right)\hat{\jmath}^T\hat{\jmath} = \begin{bmatrix} \frac{\mathbf{X^TX}}{T} + a^2\hat{\mathbf{u}}^T\hat{\mathbf{u}} & \frac{\mathbf{X^TY}}{T} + ab\hat{\mathbf{u}}^T\hat{\mathbf{v}} \\ \frac{\mathbf{Y^TX}}{T} + ab\hat{\mathbf{v}}^T\hat{\mathbf{u}} & \frac{\mathbf{Y^TY}}{T} + b^2\hat{\mathbf{v}}^T\hat{\mathbf{v}} \end{bmatrix}. \tag{4.7}$$

This shows that the block form of the spiked covariance matrix model consists of spikes (of different magnitudes) added to the self- and cross covariance terms. This now allows us to compare the different spiked models against each other on equal footing, as subproblems of Eq. (4.7). Specifically, we define the following models for comparison.

The **spiked covariance model for X**:

$$\mathbf{H_{X^TX}} = \frac{\mathbf{X^TX}}{T} + a^2\hat{\mathbf{u}}^T\hat{\mathbf{u}}. \tag{4.8}$$

The **spiked covariance model for Y**:

$$\mathbf{H_{Y^TY}} = \frac{\mathbf{Y^TY}}{T} + b^2\hat{\mathbf{v}}^T\hat{\mathbf{v}}. \tag{4.9}$$

.

The **spiked cross covariance model**:

$$\mathbf{H_{X^TY}} = \frac{\mathbf{X^TY}}{T} + ab\hat{\mathbf{u}}^T\hat{\mathbf{v}}. \tag{4.10}$$

The results for Eq. 4.8 and Eq. 4.9 have been calculated earlier [110, 87]. To explore the problem in different regimes, we define the following parameters, which measure the aspect ratios of different parts of the data matrix:

$$q_X \equiv N_X/T,\ q_Y \equiv N_Y/T,\ p_X \equiv 1/q_X,\ p_Y \equiv 1/q_Y. \tag{4.11}$$

In general, small qs at large T mean that the data are relatively well-samped, and small ps signal the opposite.

4.3 Results

For the spiked covariance models for $\mathbf{X}$ and $\mathbf{Y}$, defined in Eq. (4.8) and Eq. (4.9), respectively, the conditions on a and b to observe an outlier away from the bulk spectrum due to noise are well known [35]. These are evaluated by plugging in the value of the right most edge λ_+ of the noise bulk of the perturbed matrix M into its Stieltjes transform [87], i.e.,

$$a_{\text{crit}}^2 = \frac{1}{\mathfrak{g}_M(\lambda_+)} \tag{4.12}$$

The i.i.d. Gaussian matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{X}$ we consider are the well-studied Wishart matrices. The Stieltjes transform of a Wishart matrix obtained from T samples of an N-dimensional variable is [32]

$$\mathfrak{g}(z) = \frac{z - 1 + q - \sqrt{z - \lambda_+}\sqrt{z - \lambda_-}}{2\pi q x}. \tag{4.13}$$

Here $q = \frac{N}{T}$, $\lambda_\pm = (1 \pm \sqrt{q})^2$. λ_+ is the rightmost edge of the noise bulk and λ_- is the left most edge of the bulk. The minimum value of the additive perturbation due to the spike for the signal to be detectable is $\frac{1}{\mathfrak{g}(\lambda_+)} = \sqrt{q}(1 + \sqrt{q})$.

Thus, for a spiked covariance matrix for $\mathbf{X}$, detecting an outlier requires

$$a^2 \geq \sqrt{q_X}\,(1 + \sqrt{q_X})\,. \tag{4.14}$$

Similarly, for $\mathbf{Y}$, the condition for observability of an outlier eigenvalue is

$$b^2 \geq \sqrt{q_Y}(1 + \sqrt{q_Y}). \tag{4.15}$$

Thus, to be able to detect a shared signal between $\mathbf{X}$ and $\mathbf{Y}$ via correlating the eigenvectors corresponding to the spikes in the individual variables, both of the conditions, Eqs. (4.15, 4.15) must be satisfied simultaneously.

4.3.1 Spiked joint covariance model

The joint covariance spiked model is defined in Eq. (4.4). As entries of both $\mathbf{X}$ and $\mathbf{Y}$ have the same variance (assumed to be 1), the entries of the matrix $\mathbf{Z}$ produced by concatenating $\mathbf{X}$ and $\mathbf{Y}$ will have the same variance as well. This means that $\mathbf{Z}$ is a $T \times (N_X + N_Y)$ dimensional Wishart matrix with a variance parameter of 1, and $\hat{\mathfrak{z}}$ vector is a unit vector with dimensionality $1 \times (N_X + N_Y)$. Thus, the spectrum of the joint covariance is caluclated in the same way as for the individual variables, similar to the previous section. Conditions for detectability of an additive perturbation are also similarly calculated, resulting in:

$$c^2 = a^2 + b^2 \geq \sqrt{q_X + q_Y}\left(1 + \sqrt{q_X + q_Y}\right). \tag{4.16}$$

Further, if $\lambda_{\max}^{\text{joint}}$ is the largest eigenvalue for the joint covariance matrix, and $\mathfrak{z}_{\max}^{\text{joint}}$ is the eigenvector associated with it, then the "joint overlap", or the dot product between $\mathfrak{z}_{\max}^{\text{joint}}$ and the original spike $\hat{\mathfrak{z}}$, can be derived from well-known results using the $\mathcal{R}$ transform of the Wishart matrix [32]:

$$\|\mathfrak{z}_{\max}^{\text{joint}} \cdot \hat{\mathfrak{z}}\| = \sqrt{1 - \left(\frac{1}{(a^2+b^2)^2}\right)\mathcal{R}'\left(\frac{1}{(a^2+b^2)}\right)} \tag{4.17}$$

The $\mathcal{R}$-transform of a random matrix $\mathbf{A}$ is

$$\mathcal{R}_{\mathbf{A}}(z) = \mathcal{B}_{\mathbf{A}}(z) - 1/z\,, \tag{4.18}$$

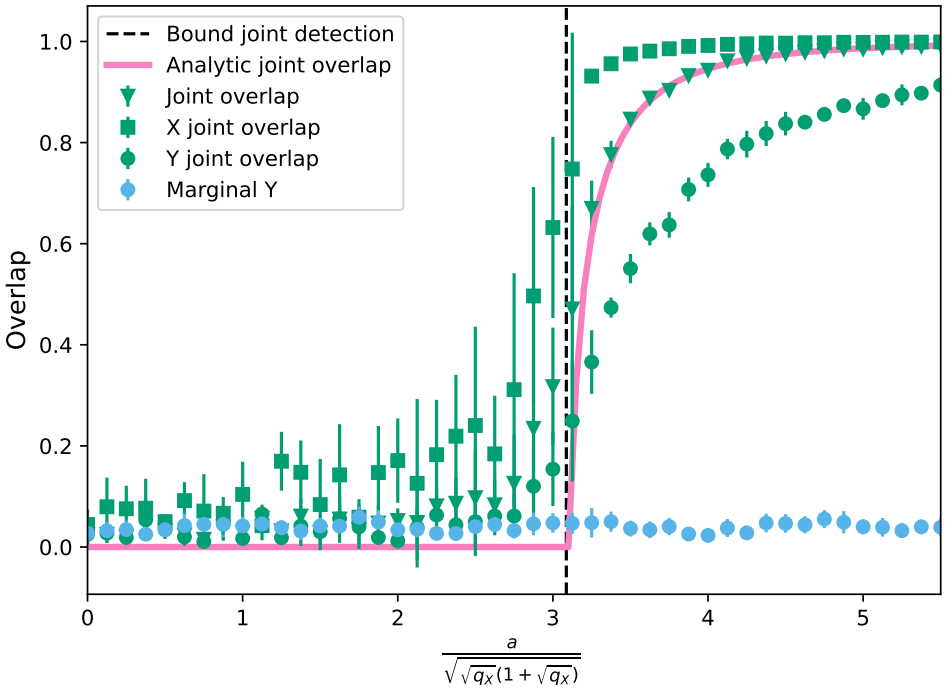

Figure 4.1: **Overlap-$\|\mathfrak{z}_{\max}^{\text{joint}} \cdot \hat{\mathfrak{z}}\|$ in the joint covariance.** Analytic solution for the joint overlap is represented by magenta line and the simulation by green triangles. X-joint overlap is the overlap between the X components of the spike and the largest eigenvector of the covariance matrix and is represented by green squares. Similarly, Y joint overlap is represented by green circles. Finally, the blue circles correspond to the the overlap in marginal Y covariance direction. For the simulations, $T = 100$, $N_X = 100$, $N_Y = 2 \times 10^3$. $b = 2.5$ and is fixed. We change the value of a and plot the overlaps. Once the value of a crosses the BBP bound (dotted line), the joint overlap goes up. More importantly as the joint overlap increases, the Y and the X joint overlaps also increase from zero, unlike the marginal Y overlap.

where the $\mathcal{B}$-transform is the functional inverse of the Stieltjes transform

$$\mathcal{B}_{\mathbf{A}}[\mathfrak{g}_{\mathbf{A}}] = z. \tag{4.19}$$

$R'_{\mathbf{A}}(z)$ is the derivative of the $\mathcal{R}$-transform of a random matrix $\mathbf{A}$.

Simplifying Eq. (4.17) gives us

$$\|\mathfrak{z}_{\max}^{\text{joint}} \cdot \hat{\mathfrak{z}}\| = \begin{cases} \sqrt{1 - \frac{q_X + q_Y}{(a^2+b^2-q_X+q_Y)^2}} & \text{if } a^2 + b^2 \geq \sqrt{q_X + q_Y}(1 + \sqrt{q_X + q_Y}), \\ 0 & \text{if } a^2 + b^2 < \sqrt{q_X + q_Y}(1 + \sqrt{q_X + q_Y}). \end{cases} \tag{4.20}$$

This overlap takes a nonzero value when the spike magnitude $c = \sqrt{a^2 + b^2}$ is above the threshold in Eq. (4.16).

We now need to verify if detection of the outlier eigenvalue in $\mathbf{Z}$ guarantees that both marginal outlier directions $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are correctly identified. To answer this, we compute numerically the X component of $\boldsymbol{\mathfrak{z}}_{\max}^{\text{joint}}$ and normalize its magnitude to 1. We denote this normalized X component by $\left(\boldsymbol{\mathfrak{z}}_{\max}^{\text{joint}}\right)_X$, and we calculate its dot product with $\hat{\mathbf{u}}$. We call the quantity $\| \left(\boldsymbol{\mathfrak{z}}_{\max}^{\text{joint}}\right)_X \cdot \hat{\mathbf{u}}\|$ the X-joint overlap, and we similarly define the Y-joint overlap.

In Fig. 4.1, we compare these overlaps in numerical simulations of a specific case, where $N_X > N_Y > T$. This is precisely the undersampled regime, where little analytical understanding exists. Yet, this regime is especially relevant for modern experimental datasets. We hold the Y signal strength b fixed, choosing it to be below the BPP transition threshold for the detectability of the spike in Y itself, Eq. (4.15). We vary the X signal strength a. We see that the joint overlap (green triangles) agrees well with the analytic prediction, Eq. (4.20) (magenta curve). Specifically, we start to reliably detect the signal (i. e., obtain a nonzero overlap) when $a^2 + b^2 > \sqrt{q_X + q_Y}(1 + \sqrt{q_X + q_Y})$, and the simulations match the analytics, although finite-size effects produce nonzero overlap even below the nominal detection threshold. The exact position of the new detection threshold can be calculated exactly using the Tracy-Widom distribution [94]. The fluctuations because of finite size effects scale as $T^{-2/3}$ if the ratios of q_X and q_Y are kept fixed. The joint X and Y overlaps both rise from zero at the same value of a as well. Note that, even though Y-joint overlap becomes nonzero when the threshold is crossed, the Y-marginal overlap (that is, the overlap of the eigenvector corresponding to the largest eigenvalue of the Y self-covariance, blue circles) remains zero as a increases. This shows that the joint covariance allows us to detect signal that is not detectable using individual variables alone.

Next we generalize these results and calculate the phase diagram for detectability of the spike (defined as a nonzero overlap between the largest singular vector(s) with both the X and the Y components of the spike, $\hat{\mathbf{u}}$, $\hat{\mathbf{v}}$) for different values of a and b (Fig. 4.2) using Eqs. (4.14, 4.15, 4.16) This phase diagram shows three possible regimes. When both the X and the Y components of the spike signal are weak (white area), correct identification of the spike is impossible from the marginal covariance matrices. Over a wide region where only one of a or b is large (green), both X and Y signals can be identified using the joint covariance $\mathbf{Z}^T\mathbf{Z}$. Yet, one of them and, in some cases, even both of them *cannot* be identified using the marginal covariance matrices. Finally, when both a and b are large enough (blue and green hatching), both individual spiked covariance matrices (blue) and joint covariance (green) can identify both components of the spike vector.

Importantly, detection of a spike is *always* easier in the joint covariance than in individual marginal covariances. In some cases, the difference is dramatic, so that existence of a strong signal component in, say, $\mathbf{X}$ makes detecting a weak signal in $\mathbf{Y}$ possible. Mathematically, this is because the constraint for a spike resulting in an outlier in the joint covariance, $a^2 + b^2 \geq \sqrt{q_X + q_Y}(1 + \sqrt{q_X + q_Y})$, is automatically satisfied if both $a^2 \geq \sqrt{q_X}(1 + \sqrt{q_X})$ and $b^2 \geq \sqrt{q_Y}(1 + \sqrt{q_Y})$ (the constraints for signal detection in the individual covariances). It is, however, a *weaker* constraint, and it can be satisfied when only one (or even none) of the marginal constraints are satisfied. This matches the intuition that the cross-covariance component of the concatenated covariance matrix provides information about the spike in addition to the self-covariance components, and this additional information can only improve the spike detection.

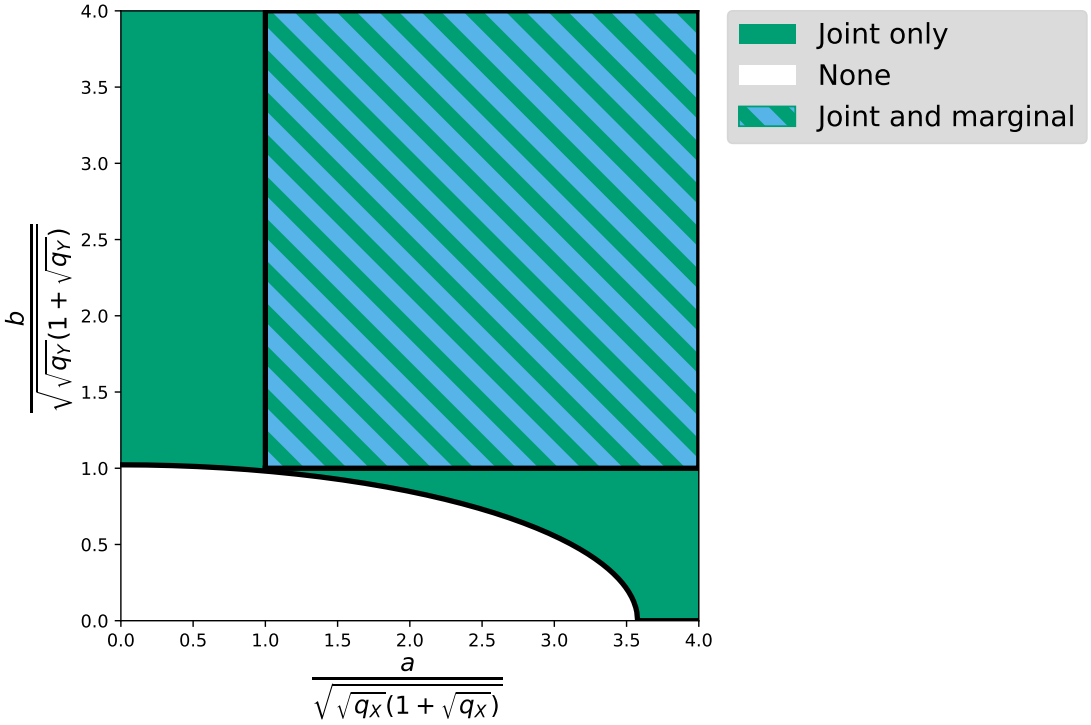

Figure 4.2: **Phase diagram for signal detection in the joint covariance and marginal covariances**. Solid green represents the region where a spike results in a detectable outlier in the joint covariance matrix. In the region with alternate blue and green hatching is the region where outliers emerge in both methods. For the white region in the plot, none of the methods are able to detect a signal. For this plot $q_X = 1$, $q_Y = 20$.

4.3.2 Spiked cross covariance model

We will take advantage of existing results for a rectangular matrix with a spike [107, 115, 116] in order to compute the conditions for detection of a signal in the cross covariance matrix. We will begin by reviewing these known results.

First, we define a spiked rectangular matrix model as

$$\tilde{\mathbf{X}}_{\mathbf{1}} = \mathbf{X}_{\mathbf{1}} + \theta \hat{\mathbf{u}}_1^T \hat{\mathbf{v}}_1. \tag{4.21}$$

Here $\mathbf{X}_1$ is a $T \times N_X$ dimensional matrix, whose entries are Gaussian random variables with zero mean and variance σ_X^2, and $\hat{\mathbf{u}}_{\mathbf{1}}$ is a $1 \times T$ dimensional unit vector. Similarly $\hat{\mathbf{v}}_{\mathbf{1}}$ is a $1 \times N_X$ dimensional unit vector. We focus on defining the value of θ_{crit}, for which there is a singular value outlier in the spectrum of $\tilde{\mathbf{X}}$. There are two types of outliers, depending on whether the outlier appears below the noise bulk (smaller than the the left edge of the noise bulk) or above the noise bulk (larger than its right edge). Here we are only interested in the second kind of outliers. To detect one, we evaluate the $\mathcal{D}$ transform for the noise matrix $\mathbf{X}_{\mathbf{1}}$ [115], which can be written as

$$\mathcal{D}_{X_1}(z) = z g_{X_1 X_1^T}(z^2) z g_{X_1^T X_1}(z^2). \tag{4.22}$$

Here, in turn, $g_{X_1 X_1^T}$ is the Stieltjes transform of $\mathbf{X_1 X_1^T}$. Similarly $g_{X_1^T X_1}$ is the Stieltjes-transform of $\mathbf{X_1^T X_1}$. Let λ_+ be the rightmost edge of the sampling noise bulk for singular values of $\mathbf{X}_{\mathbf{1}}$. Then θ_{crit} is determined by the equation

$$\mathcal{D}_{X_1}(\lambda_+) = \frac{1}{\theta_{\text{crit}}^2}. \tag{4.23}$$

For any value of $\theta > \theta_{\text{crit}}$, there exists a λ_1 that satisfies:

$$\mathcal{D}_{X_1}(\lambda_1) = \frac{1}{\theta^2}, \tag{4.24}$$

and this $\lambda_1(\theta)$ will be the outlier singular value that we expect to correspond to the spike of the magnitude θ. Letting $\mathbb{D}_{X_1}$ denote the inverse $\mathcal{D}$ transform, the largest eigenvalue of $\tilde{\mathbf{X}}$ is then

$$\lambda_1 = \begin{cases} \lambda_+ & \text{if } \theta < \theta_{\text{crit}}, \\ \mathbb{D}_{X_1}(\frac{1}{\theta^2}) & \text{if } \theta \geq \theta_{\text{crit}}. \end{cases} \tag{4.25}$$

Further, the overlaps of the left and right singular vectors $\hat{u}_{\lambda_1}$ and $\hat{v}_{\lambda_1}$ corresponding to the largest eigenvalue λ_1 with the corresponding elements of the spike using Eq. (4.25) are written as

$$\|\hat{\mathbf{u}}_\mathbf{1} \cdot \hat{u}_{\lambda_1}\|^2 = \begin{cases} 0 & \text{if } \theta < \theta_{\text{crit}}, \\ \frac{2\lambda_1 g_{X_1 X_1^T}(\lambda_1^2)}{\theta^2 \mathcal{D}'_{X_1}(\lambda_1)} & \text{if } \theta \geq \theta_{\text{crit}}, \end{cases} \tag{4.26}$$

$$\|\hat{\mathbf{v}}_\mathbf{1} \cdot \hat{v}_{\lambda_1}\|^2 = \begin{cases} 0 & \text{if } \theta < \theta_{\text{crit}}, \\ \frac{2\lambda_1 g_{X_1^T X_1}(\lambda_1^2)}{\theta^2 \mathcal{D}'_{X_1}(\lambda_1)} & \text{if } \theta \geq \theta_{\text{crit}}. \end{cases} \tag{4.27}$$

With this background, we can now analyze detectability of a spike in the cross-covariance matrix by replacing $\mathbf{X_1}$ in Eq. (4.21) with $\mathbf{X}^T\mathbf{Y}$, and the spike $ab\hat{\mathbf{u}}^\mathbf{T}\hat{\mathbf{v}}$ as in Eq. (4.10). To evaluate detectability of a spike along the lines described above, we now need to evaluate the $\mathcal{D}$ transform of $\mathbf{X}^T\mathbf{Y}$, which, in turn requires its Stieltjes transform. The latter, as well as the rightmost edge of the noise spectrum of $\mathbf{X^TY}$ have been evaluated in [86], see also Chapter 3 (3.39). The most general analytical solution is hard to obtain, so we resort to a few special cases.

Simplified solutions for $q_X \ll q_Y$

From Eq. (4.22), when the rectangular matrix is $\mathbf{X_1}$, the $\mathcal{D}$ transform is a product of Stieltjes transforms for $\mathbf{X_1}^T\mathbf{X_1}$ and $\mathbf{X_1}\mathbf{X_1}^T$. The rectangular matrix here is $\mathbf{X^TY}$. Hence the $\mathcal{D}$ transform will be function of the product of Stieltjes transforms for $\mathbf{X^TYY^TX}$ and the Stieltjes transform for $\mathbf{Y^TXX^TY}$. Using the formula for Stieltjes

transform for $\mathbf{X^T Y Y^T X}$ from Eq. (3.12), we derive the $\mathcal{D}$ transform for the model as

$$\mathcal{D}_{X^T Y}(z) = z g_{X^T Y Y^T X}(z^2) z g_{Y^T X X^T Y}(z^2) \tag{4.28}$$

$$= \left(p_X z \mathfrak{g}(z^2) + \frac{1-p_X}{z}\right)\left(p_Y z \mathfrak{g}(z^2) + \frac{1-p_Y}{z}\right) \tag{4.29}$$

Here the terms proportional to $1/z$ in both parentheses come because the matrices $\mathbf{X}$ and $\mathbf{Y}$ themselves have zero singular values. From [86] (see also Eq. (3.13)) we know that the Stieltjes transform $\mathfrak{g}(z)$ satisfies the equation:

$$a\mathfrak{g}^3 + b\mathfrak{g}^2 + c\mathfrak{g} + d = 0, \tag{4.30}$$

where

$$a = z^2 p_X p_Y, \tag{4.31}$$

$$b = z\left(p_Y(1-p_X) + p_X(1-p_Y)\right), \tag{4.32}$$

$$c = \left((1-p_X)(1-p_Y) - z p_X p_Y\right), \tag{4.33}$$

$$d = p_X p_Y. \tag{4.34}$$

To obtain simplified analytic results, we consider the case $p_Y = \epsilon p_X$, with $\epsilon \ll 1$. In this case,

$$\epsilon(\mathfrak{g}(z))^3 z^2 p_X^2 + (\mathfrak{g}(z))^2 z p_X \left(\epsilon(1-p_X) + (1-\epsilon p_X)\right) + \mathfrak{g}(z)\left((1-p_X)(1-\epsilon p_X) - z\epsilon p_X^2\right) + \epsilon p_X^2 = 0\,. \tag{4.35}$$

We need to solve this for $z\mathfrak{g}(z^2)$. For this, substituting z^2 for z, we get for $\mathfrak{g}(z^2)$:

$$\epsilon(\mathfrak{g}(z^2))^3 z^4 p_X^2 + (\mathfrak{g}(z^2))^2 z^2 p_X \left(\epsilon(1-p_X) + (1-\epsilon p_X)\right)$$

$$+ \mathfrak{g}(z^2)\left((1-p_X)(1-\epsilon p_X) - z^2\epsilon p_X^2\right) + \epsilon p_X^2 = 0\,. \quad (4.36)$$

Now after multiplying by z and then collecting terms with $z\mathfrak{g}(z^2) = \mathfrak{f}(z)$, we get:

$$\epsilon\mathfrak{f}(z)^3 z^2 p_X^2 + \mathfrak{f}(z)^2 z p_X\left(\epsilon(1-p_X) + (1-\epsilon p_X)\right) + \mathfrak{f}(z)\left((1-p_X)(1-\epsilon p_X) - z^2\epsilon p_X^2\right) + z\epsilon p_X^2 = 0. \quad (4.37)$$

In order to get the threshold for ab such that there is an outlier, we first solve for $\mathfrak{f}(\lambda_+)$ where λ_+ is the right most edge of the noise bulk, calculated in Ref. [86] as

$$\gamma_+ \approx \sqrt{\frac{1+p_X+2\sqrt{p_X}}{p_X p_Y}}. \quad (4.38)$$

This gives

$$\mathfrak{f}(\lambda_+) = \frac{-1+\sqrt{1+8\epsilon\sqrt{p_X}+8\epsilon p_X}}{2\sqrt{\epsilon}(1+\sqrt{p_X})} \quad (4.39)$$

Now, substituting the value of $\mathfrak{f}(\lambda_+)$ in $\mathcal{D}_{X^TY}(z)$ and replacing all z with λ_+ in Eq. (4.29) results in

$$\mathcal{D}_{X^TY}(\lambda_+) = \left(p_X\mathfrak{f}(\lambda_+) + \frac{1-p_X}{\lambda_+}\right)\left(p_Y\mathfrak{f}(\lambda_+) + \frac{1-p_Y}{\lambda_+}\right) \quad (4.40)$$

The condition, Eq. (4.25), to have an outlier transforms in this case into

$$ab \geq \sqrt{\frac{1}{\mathcal{D}_{X^TY}(\lambda_+)}}, \quad (4.41)$$

which evaluates to:

$$ab \geq \sqrt{(1+\sqrt{q_Y q_X})\sqrt{q_Y}}. \quad (4.42)$$

Barring λ_+, for other values of z, $\mathcal{D}_{X^TY}(z)$ was computed by solving the equation for $\mathfrak{f}$ (given in Eq. (4.37)) numerically and substituting this value in the $\mathcal{D}_{X^TY}(z)$.

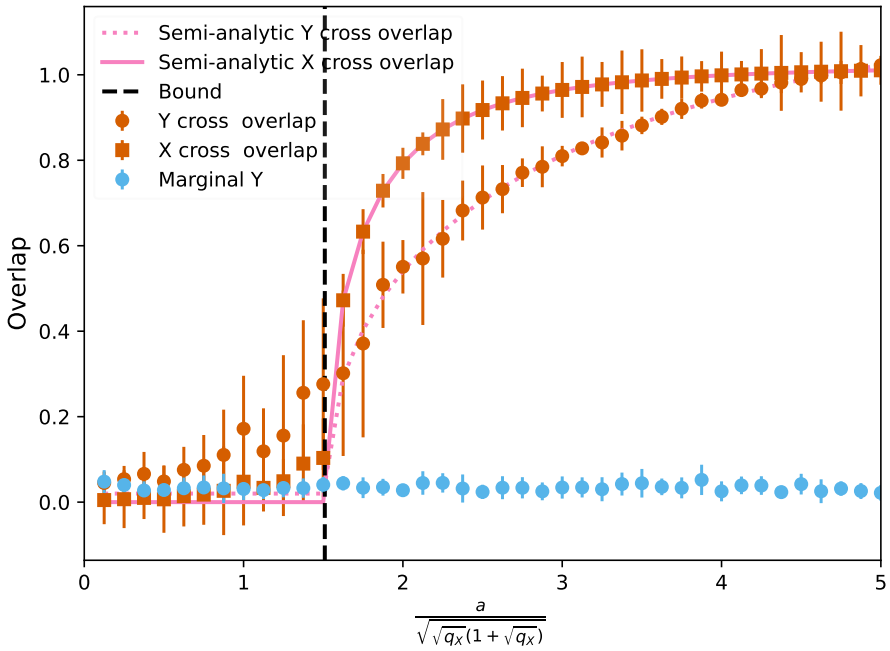

Figure 4.3: **Overlap for cross covariance.** $(\mathfrak{z}_{\max}^{\text{cross}})_X$ and $(\mathfrak{z}_{\max}^{\text{cross}})_Y$ are the left and right singular vectors, respectively, associated with the largest singular value. The quantity $\|(\mathfrak{z}_{\max}^{\text{cross}})_X \cdot \hat{\mathbf{u}}\|$ is the X-cross overlap, denoted with red squares and the semi-analytic solution for it is represented by solid magenta line. Similarly, $\|(\mathfrak{z}_{\max}^{\text{cross}})_Y \cdot \hat{\mathbf{v}}\|$ is the Y-cross overlap, denoted with red circles and the semi analytic solution for it is denoted by magenta dotted line. For comparison, we also show the marginal Y-overlap (blue circles), similar to Fig. 4.1. The dashed black line represents the analytically calculated BBP transition value for the cross overlap. Crucially, the cross overlap is much larger than the marginal one. Thus the cross-covariance is better at detecting the spike than the individual covariance matrices. For all these points, $T = 100$, $N_X = 100$, $N_Y = 2 \times 10^3$ and $b = 2.5$. b remains fixed, and overlaps are plotted as a function of a.

Let $(\mathfrak{z}^{\text{cross}}_{\max})_X$ and $(\mathfrak{z}^{\text{cross}}_{\max})_Y$ be the left and right singular vectors respectively associated with the largest singular value. We call $\|(\mathfrak{z}^{\text{cross}}_{\max})_X \cdot \hat{\mathbf{u}}\|$ the X-cross overlap. Given that $\frac{1}{a^2b^2} = \mathcal{D}_{X^TY}(\lambda_1)$, this overlap can be computed "semi analytically". That is, we solve the cubic equation for $\mathfrak{f}$ numerically and plug it into the analytical expression

$$\|(\mathfrak{z}^{\text{cross}}_{\max})_X \cdot \hat{\mathbf{u}}\| = \begin{cases} 0 & \text{if } ab < \sqrt{\frac{1}{\mathcal{D}_{X^TY}(\lambda_+)}}, \\ \frac{-2\left(p_X \mathfrak{f}(\lambda_1) + \frac{1-p_X}{\lambda_1}\right)}{a^2b^2\mathcal{D}'_{X^TY}(\lambda_1)} & \text{if } ab \geq \sqrt{\frac{1}{\mathcal{D}_{X^TY}(\lambda_+)}}. \end{cases} \tag{4.43}$$

Similarly, $\|(\mathfrak{z}^{\text{cross}}_{\max})_Y \cdot \hat{\mathbf{v}}\|$ is the Y-cross overlap, and its value can be obtained semi-analytically by plugging the numerical solution for $\mathfrak{f}$ into

$$\|(\mathfrak{z}^{\text{cross}}_{\max})_Y \cdot \hat{\mathbf{v}}\| = \begin{cases} 0 & \text{if } ab < \sqrt{\frac{1}{\mathcal{D}_{X^TY}(\lambda_+)}} \\ \frac{-2\left(p_Y \mathfrak{f}(\lambda_1) + \frac{1-p_Y}{\lambda_1}\right)}{a^2b^2\mathcal{D}'_{X^TY}(\lambda_1)} & \text{if } ab \geq \sqrt{\frac{1}{\mathcal{D}_{X^TY}(\lambda_+)}} \end{cases} \tag{4.44}$$

In Fig. 4.3, we compare these semi-analytical cross-overlaps to the empirical cross-overlaps in simulated data. We also compare them to marginal overlaps, similar to the analysis in the previous Section. The agreement between the theory and the simulations is excellent, showing a BBP-like detectability transition. Further, for these parameter values, it is clear that the cross-covariance matrix detects the spike a lot before both marginal covariance matrices do.

We formalize this superiority of the cross-covariance matrix by exploring the phase diagram of the signal detectability as a function of the marginal spike magnitudes, a and b, normalized such that the marginal covariances detect the spikes at exactly 1.0 on both axes, Fig. 4.4. We observe that the cross covariance is always better at detecting the spike than the individual marginal covariances in the undersampled regime, *i. e.*, when either $q_X \gg 1$ or $q_Y \gg 1$. As for the concatenated covariance, it seems to be possible to use the signal strength in the smaller-dimension component $\mathbf{X}$, where it is stronger, to make an effectively weaker signal in the larger dimensional

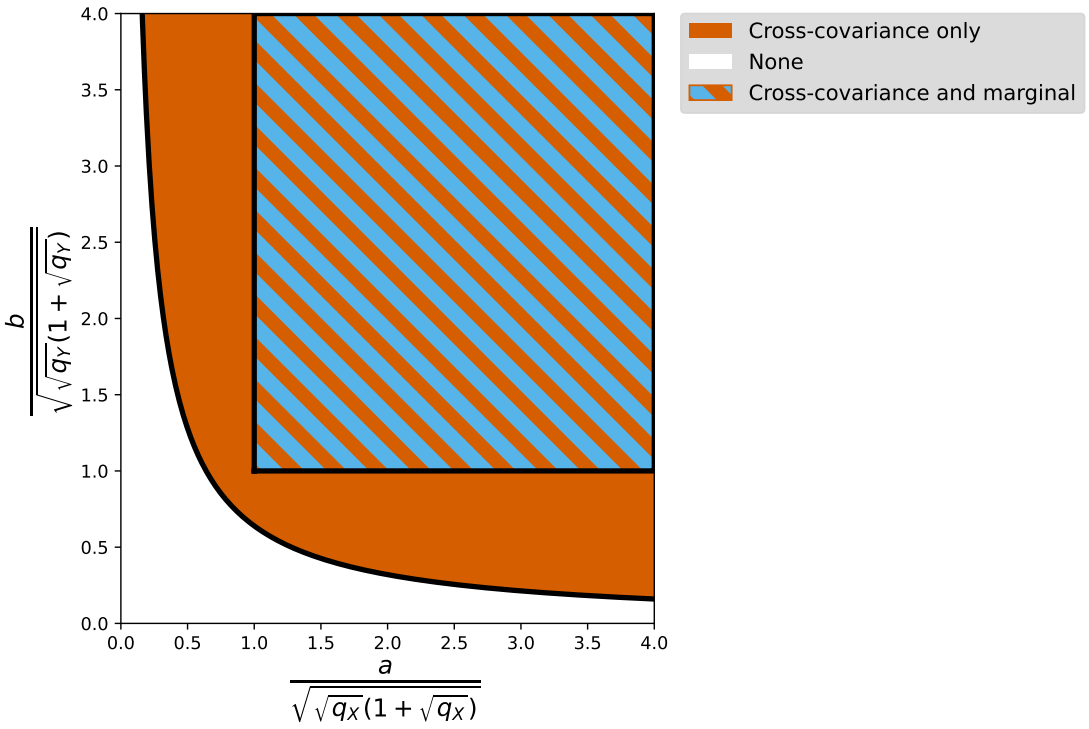

Figure 4.4: **Phase diagram spike detectability for cross covariance and marginal covariances.** The solid red region, given by Eq. (4.42), is where cross covariance is able to detect the outlier with a nonzero overlap with both the X and the Y components of the spike. Blue region is where the self-covariance of $\mathbf{X}$ and $\mathbf{Y}$ are both able to detect their marginal spike contributions, thus having information about the entire spike. Thus the region with alternating blue and red hatch is the region where both approaches have nonzero overlaps with the spike (though the magnitudes of the overlaps can be different). Crucially, cross-covariance may detect the spike when the marginal covariances cannot, but not the other way around. The white solid region is the region where neither the cross-covariance nor the marginal will be able to detect the signal. For this plot $q_X = 1$, $q_Y = 20$.

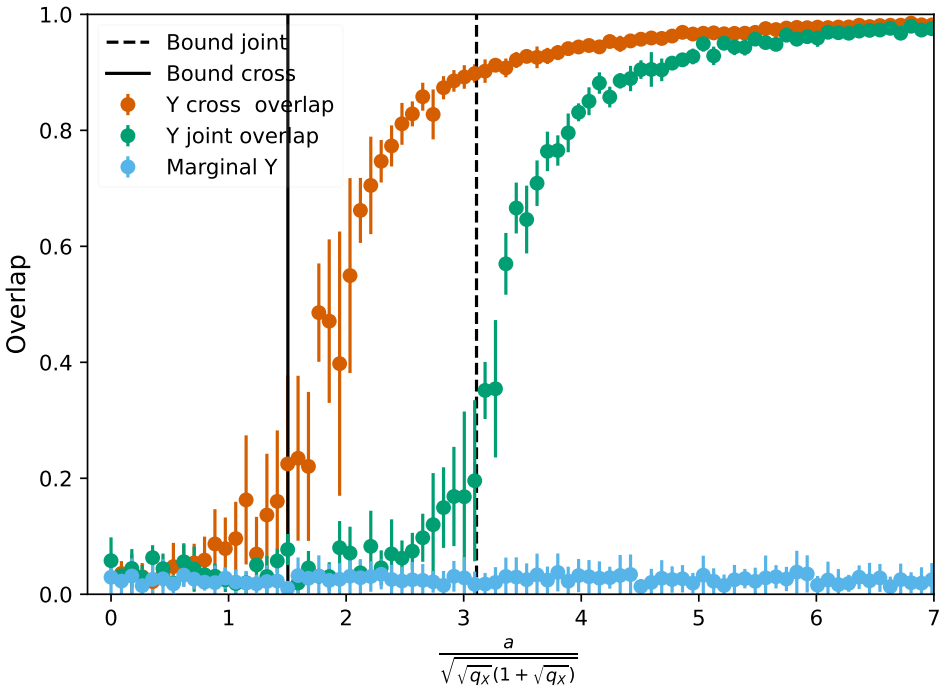

Figure 4.5: **Comparison between joint and cross overlaps.** The red circles represent the Y-cross overlap. The green circles represent the Y-joint overlap. The blue circles represent the Y-marginal overlap. For all these points, $T = 100$, $N_X = 100$, $N_Y = 2 \times 10^3$ and $b = 2.5$. b remains fixed, a is varied and the overlaps are plotted as a function of a. Solid and dashed black lines represent the analytically calculated BBP transition values for the Y-cross overlap and the Y-joint overlap, respectively.

component $\mathbf{Y}$ visible, even if it would nondetectable alone. Further, for some parameter combinations, the two signals can be detectable from the cross-covariance when neither will result in an outlier in the marginal covariances.

4.3.3 Comparison between cross covariance and joint covariance

In Fig. 4.5, we compare the overlaps for Y observed for different methods as a function of changing a for a fixed b. For the illustrated parameters, the value of b is small, and the marginal covariance of Y should have no outlier and thus no overlap with the spike. This is, indeed, the case. Crucially, the cross overlap is stronger than the joint

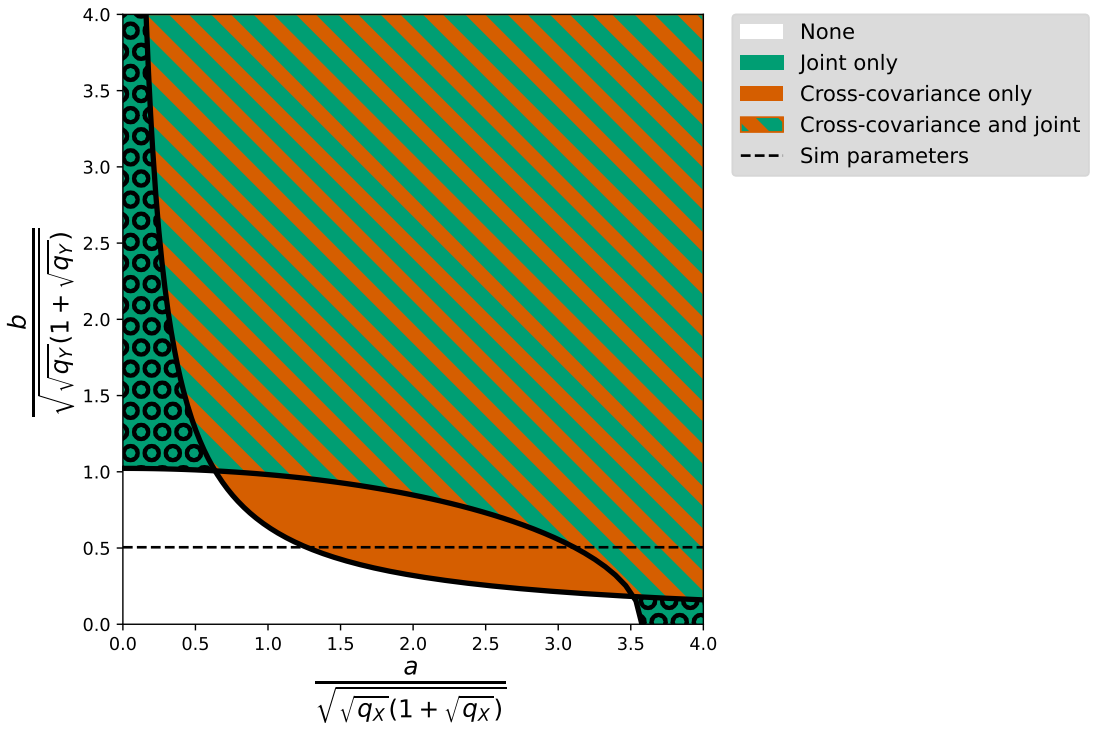

Figure 4.6: **Phase diagram spike detectability for cross covariance and joint covariance.** The solid red region, given by Eq. (4.42), is where cross covariance is able to detect the outlier with a nonzero overlap with both the X and the Y components of the spike. Spotted green region is where the joint covariance of $\mathbf{X}$ and $\mathbf{Y}$ is able to detect the outlier spike with a nonzero overlap with both the X and the Y components of the spike. Thus the region with alternating green and red hatch is the region where both approaches have nonzero overlaps with the spike (though the magnitudes of the overlaps can be different). The white solid region is the region where neither the cross-covariance nor the joint covariance will be able to detect the signal. For this plot $q_X = 1$, $q_Y = 30$. The dotted line labeled as sim parameters gives us the values of signal strength for $\mathbf{X}$ (labeled as a) and values of signal strength for $\mathbf{Y}$ (labeled as b) used for generating the plot Fig. 4.5

one. This is because the example in the figure is in the limited area of the phase spaces of Figs. 4.2, 4.4, where an outlier in the cross-covariance is expected to be easier to detect than in the joint covariance. That such region exists is surprising. Indeed, the cross-covariance matrix is only a subset of the joint covariance one. Naively, one would expect that, by adding more data, one should make spike detection easier, so that the joint covariance approach should never be inferior. An intuitive explanation for the phenomenon is still waiting to be found.

4.4 Discussion

In this study we created a set of spike models for joint covariance, cross-covariance and individual marginal covariances that allow us to understand when a low-dimensional signal can be detected despite sampling noise. Our model allows us to directly compare emergence of outliers in the spectra of the joint (concatenated) covariance, cross-covariance, and individual marginal covariances. We note that an outlier *always* emerges in the joint or cross covariance matrices for a weaker spike strength compared to individual marginal covariances. Thus, statistical methods exploiting cross covariance or joint covariance matrices are more data efficient, in that they should be able to detect a weaker signal or to detect a signal with fewer samples compared to individual marginal covariances.

While joint and cross covariances detect weaker signals than marginal covariances, neither is superior to the other, and both have their own strengths and weaknesses. Joint covariance can detect an outlier even if the spike is extremely small in one of the two datasets being concatenated into the joint matrix. This is not the case for the cross-covariance, for which the critical signal strength is given by the product of the spikes ab for a given T, N_X, N_Y. It thus fails to detect an outlier spike if the spike is extremely small in one of the two datasets. Yet, there are parameter regions

where cross-covariance approach is the obly one that works, and also regions where both cross- and joint methods work, but the overlap produced by the cross-covariance method is stronger. Understanding how this translates to practical statistical methods, and how would one know which of the two methods should be used in which situations is the next step in this research direction.

Chapter 5

Discussion

The key insights gained from the dissertation are as follows.

- In Chapter 2, I showed that when trying to learn the activation energy of rearrangements in glassy systems, prediction accuracy fails to capture if the energy function has been learned accurately. Similarly, a high quality linear relationship between softness, logarithm of the rearrangement probability, and $1/T$, cannot be used to conclude that the energy was learned correctly as well. Further, the ability to correctly predict the energy depended heavily on the features used to train the classifiers (SVMs in this case). Tracking the variance of the inferred energy across different choices of input features performed better than cross-validation accuracy in selecting the best possible set of input features to train the SVM.

- In Chapter 3, I evaluated the exact solution for the singular value spectrum of finite sampling-induced noise-noise cross-covariance between two datasets. This parallels classical results for the spectrum of sampling-induced self-covariance.

- In Chapter 4, I developed a spiked model for cross-covariance and joint-covariance, similar to classic spiked covariance matrix model for more traditional random

matrix data [34]. For these spike models, I derived and verified numerically the solution for the minimum spike amplitude for the spike to result in an outlier in the singular value spectrum, and hence be detectable. Using this result, I showed that, for reconstructing a shared rank one spike in two datasets X and Y, it is always better to calculate a joint covariance or the cross-covariance rather than try to calculate individual covariances of the variables.

These results open new research directions, which I will discuss here. For these directions, I will restrict myself to the latter part of the dissertation and will focus the potential applications of Random Matrix theory (RMT) methods in data science and machine learning.

5.1 Generative model with a shared signal

Instead of evaluating the spiked covariance matrix model, as in Chapter 4, an alternative approach could be to build generative models for shared features in two high-dimensional datasets. The simplest such model has one shared feature between $\mathbf{X}$ and $\mathbf{Y}$, and has the form

$$\hat{\mathbf{X}} = \frac{\mathbf{X} + a\mathbf{P}\hat{\mathbf{u}}}{\hat{\sigma}_X}, \tag{5.1}$$

$$\hat{\mathbf{Y}} = \frac{\mathbf{Y} + b\mathbf{P}\hat{\mathbf{v}}}{\hat{\sigma}_Y}. \tag{5.2}$$

Here, $\mathbf{X}$ and $\mathbf{Y}$ are the two uncorrelated, Gaussian i.i.d. datasets as elswhere in this dissertation; a and b are constants the strength of the contributions of the shared signal to each of the variables, and $\mathbf{P}$ is a matrix of dimensions $T \times 1$, representing samples of a latent variable driving this signal. We take entries of this matrix to be Gaussian random variables with zero mean and unit variance. Further, $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are $1 \times N_X$ and $1 \times N_Y$ dimensional unit vectors, giving the (fixed) projections of

the shared signal onto X and Y. Then $\hat{\sigma}_X^2 = \sigma_X^2 + \frac{a^2}{T}$ and $\hat{\sigma}_Y^2 = \sigma_U^2 + \frac{b^2}{T}$. In this calculation $T \to \infty$, so $\sigma_X^2 = \hat{\sigma}_X^2$ and $\sigma_Y^2 = \hat{\sigma}_Y^2$. That is, the existence of the shared signal does not change the variance of the observables variables much.

For $\hat{\mathbf{X}}$ given by Eq. (5.1), we can calculate its empirical covariance matrix:

$$\mathcal{H}_{X^T X} = \frac{1}{T}\hat{\mathbf{X}}^T\hat{\mathbf{X}} \tag{5.3}$$

$$= \frac{1}{T\sigma_X^2}(\mathbf{X} + a\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{X} + a\mathbf{P}\hat{\mathbf{u}}) \tag{5.4}$$

$$= \frac{1}{T\sigma_X^2}\left(\mathbf{X}^T\mathbf{X} + \mathbf{X}^T(a\mathbf{P}\hat{\mathbf{u}}) + (a\mathbf{P}\hat{\mathbf{u}})^T\mathbf{X} + a^2(\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{P}\hat{\mathbf{u}})\right). \tag{5.5}$$

This is not the same as the spiked model, Eq. (4.8), because of the two terms linear in $\mathbf{P}$. Analyzing Eq. (5.5) in full generality is hard because terms linear and quadratic in $\mathbf{P}$ are not mutually free. Thus, an interesting research question is under which conditions the linear terms in Eq. (5.5) can be neglected. Since these terms are symmetric, preliminary numerical simulations and analysis in Ref. [91] show that such regimes exist. If we are able to identify them, and we assume, with no loss of generality, that $\sigma_x^2 = \sigma_P^2 = 1$. Then, Eq. (5.5) reduces to

$$\mathcal{H}_{X^T X} \approx \frac{\mathbf{X}^T\mathbf{X}}{T} + \frac{\sigma_P^2}{\sigma_X^2}a^2\hat{\mathbf{u}}^{\mathbf{T}}\hat{\mathbf{u}} \tag{5.6}$$

$$\approx \frac{\mathbf{X}^T\mathbf{X}}{T} + a^2\hat{\mathbf{u}}^{\mathbf{T}}\hat{\mathbf{u}}, \tag{5.7}$$

which is the same as the more traditional spiked covariance matrix model. In other words, the question here would be to either solve for the spectrum of the generative model, or to investigate under which conditions it reduced to the spiked model, where calculations are easier.

Just like we can define a generative model that corresponds to the spiked covariance model, we can define generative models for spiked joint and cross-covariance matrices. Indeed, let us define $\hat{\mathbf{Z}}$ given by the concatenation of $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$. We now have

$\hat{\mathbf{Z}}$ that is a $T \times (N_X + N_Y)$ dimensional matrix

$$\hat{\mathbf{Z}} = \left(\frac{\mathbf{X} + a\mathbf{P}\hat{\mathbf{u}}}{\sigma_X}, \frac{\mathbf{Y} + b\mathbf{P}\hat{\mathbf{v}}}{\sigma_Y} \right) \tag{5.8}$$

Assuming $\sigma_X^2 = \sigma_Y^2$ without any loss of generality, and calculating the covariance of $\mathbf{Z}$, we would get

$$\mathcal{H}_{Z^T Z} = \frac{1}{T}\hat{\mathbf{Z}}^T\hat{\mathbf{Z}} \tag{5.9}$$

$$= \frac{1}{T\sigma_X^2}\left(\mathbf{Z}^T\mathbf{Z} + \mathbf{Z}^T(c\mathbf{P}\hat{\mathfrak{z}}) + (c\mathbf{P}\hat{\mathfrak{z}})^T\mathbf{Z} + c^2(\mathbf{P}\hat{\mathfrak{z}})^T(\mathbf{P}\hat{\mathfrak{z}})\right). \tag{5.10}$$

Here $c^2 = a^2 + b^2$ and $\hat{\mathfrak{z}} = \left(\frac{a}{c}\hat{\mathbf{u}}, \frac{b}{c}\hat{\mathbf{v}}\right)$. The second and third terms in Eq. (5.10) are the extra terms that we do not find in Eq. (4.4), and both the terms are symmetric. It has been shown earlier that, for latent feature models, where the rank of $\mathbf{P} \ll N_X\ ; N_Y\ ; T$, the second and third terms can be ignored [91] under various conditions. Though the cross terms do contribute to the final results, they do not qualitatively change either the overlap or the value of the outlier eigenvalue. Instead, these terms linear in $\mathcal{P}$ shift both the outlier eigenvalue and the overlap up (that is, make detection of the outlier easier) because of the additional signal-correlated contributions. Here, the rank of $\mathbf{P}$ is one, and we know from simulations that, for such rank one perturbation, the terms linear in $\mathbf{P}$ cannot be ignored, in general, if quantitative result are desired. Thus, as above, an interesting future research problem is to identify when these linear terms can be neglected quantitatively.

Finally, we may also try to define a generative model for the cross-covariance matrix of uncorrelated data with one shared signal:

$$\mathcal{H}_{X^T Y} = \frac{1}{T}\hat{\mathbf{X}}^T\hat{\mathbf{Y}} \tag{5.11}$$

$$= \frac{1}{T\sigma_X\sigma_Y}(\mathbf{X} + a\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{Y} + b\mathbf{P}\hat{\mathbf{v}}) \tag{5.12}$$

$$= \frac{1}{T\sigma_X\sigma_Y}\left(\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T(b\mathbf{P}\hat{\mathbf{v}}) + (a\mathbf{P}\hat{\mathbf{u}})^T\mathbf{Y} + ab(\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{P}\hat{\mathbf{v}})\right) \tag{5.13}$$

Unlike in Eq. (5.5) and Eq. (5.10), the second and third terms in Eq. (5.13) are no longer symmetric, which makes it much harder to find conditions when they can be neglected. Indeed, even in our preliminary simulations (not shown), dropping these non-symmetric cross terms has had a significant impact on the edge of the bulk of the sampling-induced eigenvalues, and hence on the detectability of the signal.

It is not surprising that the generative latent feature models are not exactly equal to spiked models. For example, if one studies spiked covariance models, one can detect not only outliers that are larger than the right edge of the noise bulk, but also outliers that are smaller than the left edge of the bulk. That is, very large *and* very small spikes are detectable. Conceptually, this should not possible for latent features, generative models. Indeed, for very small values of signal strength, it is the linear terms, which are cross terms between the noise $\mathbf{X}$ and the signal $\mathbf{P}\hat{\mathbf{u}}$, that contribute to the covariance and make it impossible to detect left outliers. Once that happens, we can no longer do the analysis using simple RMT methods since the four terms contributing to the covariance are no longer free (rotationally invariant with respect to each other). We can only neglect the second and third term if we are trying to find an outlier eigenvalue that is larger than the largest eigenvalue in the bulk [91]. But when can this be done for non-symmetric cross terms is unclear. Understanding the non-symmetric case is important because, unlike in the symmetric case, where the RMT-based predictions of overlap and the largest eigenvalue are relatively accurate, this accuracy degrades in the case of non-symmetric cross terms, sometimes resulting in qualitatively different results.

To get intuition for how to treat the non-symmetric case, I propose to study two different generative models, where the degree of nonsymmetry can be tuned. First such model would start with the same matrix $\mathbf{X}$, but with two different strengths of

latent signal in the two terms contributing to the covariance:

$$\mathcal{H}_1 = \frac{1}{T\sigma_X^2}(\mathbf{X} + a\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{X} + b\mathbf{P}\hat{\mathbf{u}}) \tag{5.14}$$

$$= \frac{1}{T\sigma_X^2}\left(\mathbf{X}^T\mathbf{X} + \mathbf{X}^T(b\mathbf{P}\hat{\mathbf{u}}) + (a\mathbf{P}\hat{\mathbf{u}})^T\mathbf{X} + ab(\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{P}\hat{\mathbf{u}})\right). \tag{5.15}$$

One can make, for example, a small and b large, which would change the relative contribution of different terms to the covariance, and match asymptotics for different regimes of a and b.

Another possibly useful generative model is would have two distinct matrices with the same signal strength:

$$\mathcal{H}_2 = \frac{1}{T\sigma_X\sigma_Y}(\mathbf{X} + a\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{Y} + a\mathbf{P}\hat{\mathbf{v}}) \tag{5.16}$$

$$= \frac{1}{T\sigma_X\sigma_Y}\left(\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T(a\mathbf{P}\hat{\mathbf{v}}) + (a\mathbf{P}\hat{\mathbf{u}})^T\mathbf{Y} + a^2(\mathbf{P}\hat{\mathbf{u}})^T(\mathbf{P}\hat{\mathbf{v}})\right). \tag{5.17}$$

If $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are sufficiently distinct, different terns in Eq. (5.17) will become free, making analysis easier. One then can hope to analytically continue to the same (or, at least, partially overlapping $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$.

Understanding how to treat non-symmetric cross terms in Eq. (5.15) and Eq. (5.17) will help us in understanding how to treat these terms in general. My current hypothesis, based on preliminary simulations, is that the spiked models give the worst-case performance. The latent feature model have larger outlier eigenvalues, as well higher overlap as compared to pure spike models—all because of contribution of the cross terms. This problem has to be mostly studied using simulations as, once the cross terms start contributing, the problem is no longer analytically tractable with common methods.

5.2 Increasing the signal rank

Once we understand how to deal with cross terms in the generative models, the next natural extension is to allow the $\mathbf{P}$ matrix to have a rank $m \gg 1$. This allows for richer generative models. Such models would be written as

$$\bar{\mathbf{X}} = \mathbf{X} + \mathbf{P}\mathbf{Q}_X, \tag{5.18}$$

$$\bar{\mathbf{Y}} = \mathbf{Y} + \mathbf{P}\mathbf{Q}_Y \tag{5.19}$$

Now the matrix $\mathbf{P}$ is the shared signal matrix between $\mathbf{X}$ and $\mathbf{Y}$. It has dimensions $T \times m$. The matrices $\mathbf{Q}_X$ and $\mathbf{Q}_Y$ are the projection of the signal matrix $\mathbf{P}$ into the spaces of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and they have dimensions $m \times N_X$ and $m \times N_Y$. Thus m latent features get randomly sampled T times (matrix $\mathbf{P}$), and each of the N_X and N_Y measured variables in X and Y, respectively, is a quenched random linear combination of the latent features ($\mathbf{Q}_X$ and $\mathbf{Q}_Y$). We assume $m \leq T, N_X, N_Y$ throughout this work, so that the rank of the signal matrices $\mathbf{P}\mathbf{Q}_X$ and $\mathbf{P}\mathbf{Q}_Y$ is equal to m, and the features can be estimated from the samples.

The entries of $\mathbf{P}$, $\mathbf{Q}_X$ and $\mathbf{Q}_Y$ are Gaussian random variables with zero mean and variances σ_P^2, $\sigma_{Q_X}^2$, and $\sigma_{Q_Y}^2$, respectively:

$$P_{t\mu} \sim \mathcal{N}(0, \sigma_P^2), \tag{5.20}$$

$$t = 1, \ldots, T,\ \mu = 1, \ldots, m. \tag{5.21}$$

and

$$Q_{X\,\mu n_1} \sim \mathcal{N}(0, \sigma_{Q_X}^2)\,, \quad Q_{Y\,\mu n_2} \sim \mathcal{N}(0, \sigma_{Q_Y}^2)\,, \tag{5.22}$$

$$\mu = 1, \ldots, m,\ n_1 = 1, \ldots, N_X,\ n_2 = 1, \ldots, N_Y. \tag{5.23}$$

We make this Gaussian choice for analytic tractability; in applications to real data, the means and variances of the entries may need to be matched to those of the measured variables. There may be instances where a Gaussian distribution may not be the best approximation for specific data. Finally, the elements of the noise matrices $\mathbf{X}$ and $\mathbf{Y}$ are also i.i.d. Gaussian random variables with variance σ_X and σ_Y, respectively. So every observation in X and Y has a variance of σ_X and σ_Y respectively.

We would like to calculate the spectrum of this latent feature model using RMT-based methods. These, however, require a square matrix. Thus, to be able to use these methods, we would first need to square the cross-covariance matrix, resulting in the object of interest:

$$\mathrm{H} = \frac{1}{T^2\sigma_X^2\sigma_Y^2}\bar{\mathbf{X}}^T\bar{\mathbf{Y}}\bar{\mathbf{Y}}^T\bar{\mathbf{X}} \tag{5.24}$$

It is easy to see that H has contributions from 16 terms. Of these, 14 are cross terms with all their associated challenges. We calculated the Stieltjes transform and the spectrum of the first of these terms, $\frac{1}{T^2\sigma_X^2\sigma_Y^2}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$, in [86]. While not reported here, I have also calculated the spectrum and various transforms of second of the sixteen terms, $\frac{1}{T^2\sigma_X^2\sigma_Y^2}\left(\mathbf{P}\mathbf{Q}_X\right)^T\left(\mathbf{P}\mathbf{Q}_Y\right)\left(\mathbf{P}\mathbf{Q}_Y\right)^T\left(\mathbf{P}\mathbf{Q}_X\right)$. Solving for these quantities involved solving a 6^{th} order polynomial equation for the Stieltjes transform. The solution for this equation can give us the lowest order approximation of number of shared latent features that can be identified for a given signal to noise ratio, before the signal cannot no longer be separated from noise bulk. However, this involves ignoring the remaining 14 cross-terms to make the results analytically tractable, and it is unclear when such an approximation would be valid. Exploring this is an interesting future research direction.

5.3 General generative model: shared and private variance

A more general linear latent feature model would involve, along with shared latent features between $\mathbf{X}$ and $\mathbf{Y}$, additional low-dimensional structure in $\mathbf{X}$ and $\mathbf{Y}$ that is not shared between the variables (is private). This model was first introduced and analyzed numerically in Ref. [95]. For this model, the data matrices are given by:

$$\bar{\mathbf{X}} = \mathbf{X} + \mathbf{U_X V_X} + \mathbf{P Q_X}, \tag{5.25}$$

$$\bar{\mathbf{Y}} = \mathbf{X} + \mathbf{U_Y V_Y} + \mathbf{P Q_Y}. \tag{5.26}$$

Evaluating the spectrum of the cross-covariance in this model would include contributions from 81 terms. Of these, 78 are cross-terms, each non-symmetric, which makes the calculations—or the understanding of when they can be safely ignored—especially challenging. If just one of the 78 terms cannot be ignored, traditional RMT methods would no longer be applicable.

An intriguing possibility is that, when the number of contributions is so large, some form of self-averaging among the terms may start happening. Numerical simulations can answer if such complicated expressions combining many random matrices again result in some simple, possibly universal, spectra.

5.4 Deep learning and random matrices

Another area of broad interest is analyzing deep learning using tools from RMT. For example, to understand the dynamics of training of neural networks, one often analyzes the geometric structure of the loss surface — the value of the loss (the optimization objective of the network), averaged over the data as a function of the network weights. Then training NNs involves finding minima of the loss surface in the

weight space. The loss suface is typically investigated through the second order Taylor expansion of the loss, called the Hessian. Assuming independent and identically distributed Gaussian inputs and network path independence, a multi-layer ReLU neural network's loss was shown to be equivalent to that of a spin-glass model [117], and its Hessian spectrum was shown to be given by a Gaussian Orthogonal Ensemble [118] (GOE). GOE is a standard model in RMT, which gives the celebrated Wigner's semi circle law for the eigenvalue spectrum. Because of this correspondence, RMT-based methods have been used to understand the learning dynamics in deep neural networks by studying the Hessian [119]. The calculations assume that dimensionality of each subsequent layer in a deep network is strictly decreasing. This assumption helps because, under this condition, a spike in the input layer is preserved through successive rounds of projections. However, in real-world problems, the subsequent layers of deep learning aren't necessarily decreasing. So evaluating the general case for the effect of random projection (where the projection space can be of a larger dimensionality than the original spike), and understanding if the spike survives sequential projection is an important, unsolved problem. Understanding this would allows one to follow how data with a single spike (that is, data with an embedded signal) changes as it goes from one layer of the neural network to the next. The critical challenge in these calculations will again be dealing with the cross terms. This is because such analysis will be equivalent to solving the latent feature cross covariance model, Eq. (5.13), with the a or b signifying the latent feature strength in subsequent layers.

Other attempts to use RMT methods to understand deep learning involve using the Marachenko-Pastur distributions to explain the structure of features in Large Language Models (LLM) and to understand which layers in them are important for learning [120]. It is known that outliers from the pure noise in the spectra of weights for deep networks are a crucial feature of well-trained models [121]. Thus, it may be interesting to analyze the distribution of trained individual key (K), query (Q)

and value (V) matrices in transformer neural network architectures and search for deviation from a random initialization for the trained models. The layers where the Key, Query or Value matrices have a singular value spectra matching the scaling of the Marchenko-Pastur distribution can be considered as not changing much with training, and thus not storing any learned features. One can also calculate and check how random the attention matrix QK^T is by comparing it with the spectrum of XY^T, for which we now have an exact solution. The exact calculation for cross-covariance allows us to extend similar analysis to more complex cases like latent attention in more modern models, such as Deepseek [122]. In such latent attention, the attention matrix is written as $K_1K_2Q_1Q_2$. We have the exact calculation for spectra of terms of this form, allowing us to understand how nonrandom attention is after initialization and training.

Bibliography

[1] Nobel Prize Outreach. The nobel prize in physics 2024, 2025. Accessed 11 Mar 2025.

[2] Nobel Prize Outreach. The nobel prize in chemeistry 2024, 2025. Accessed 11 Mar 2025.

[3] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91:045002, Dec 2019.

[4] Licheng Jiao, Xue Song, Chao You, Xu Liu, Lingling Li, Puhua Chen, Xu Tang, Zhixi Feng, Fang Liu, Yuwei Guo, Shuyuan Yang, Yangyang Li, Xiangrong Zhang, Wenping Ma, Shuang Wang, Jing Bai, and Biao Hou. Ai meets physics: a comprehensive survey. *Artificial Intelligence Review*, 57(9):256, Aug 2024.

[5] Juan Carrasquilla and Roger G. Melko. Machine learning phases of matter. *Nature Physics*, 13(5):431–434, May 2017.

[6] Wenjian Hu, Rajiv R. P. Singh, and Richard T. Scalettar. Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination. *Phys. Rev. E*, 95:062122, Jun 2017.

[7] Evert P. L. van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber. Learning phase transitions by confusion. *Nature Physics*, 13(5):435–439, May 2017.

[8] Lei Wang. Discovering phase transitions with unsupervised learning. *Phys. Rev. B*, 94:195105, Nov 2016.

[9] Rui Wang, Yu-Gang Ma, R. Wada, Lie-Wen Chen, Wan-Bing He, Huan-Ling Liu, and Kai-Jia Sun. Nuclear liquid-gas phase transition with machine learning. *Phys. Rev. Research*, 2:043202, Nov 2020.

[10] X.L. Zhao and L.B. Fu. Machine learning phase transition: An iterative proposal. *Annals of Physics*, 410:167938, 2019.

[11] Cinzia Giannetti, Biagio Lucini, and Davide Vadacchino. Machine learning as a universal tool for quantitative investigations of phase transitions. *Nuclear Physics B*, 944:114639, 2019.

[12] Jonas Greitemann, Ke Liu, Ludovic D. C. Jaubert, Han Yan, Nic Shannon, and Lode Pollet. Identification of emergent constraints and hidden order in frustrated magnets using tensorial kernel methods of machine learning. *Phys. Rev. B*, 100:174408, Nov 2019.

[13] S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras, and A. J. Liu. A structural approach to relaxation in glassy liquids. *Nature Physics*, 12(5):469–471, May 2016.

[14] Samuel S. Schoenholz, Ekin D. Cubuk, Efthimios Kaxiras, and Andrea J. Liu. Relationship between local structure and relaxation in out-of-equilibrium glassy systems. *Proceedings of the National Academy of Sciences*, 114(2):263–267, 2017.

[15] Daniel M. Sussman, Samuel S. Schoenholz, Ekin D. Cubuk, and Andrea J. Liu. Disconnecting structure and dynamics in glassy thin films. *Proceedings of the National Academy of Sciences*, 114(40):10601–10605, 2017.

[16] E. D. Cubuk, R. J. S. Ivancic, S. S. Schoenholz, D. J. Strickland, A. Basu, Z. S. Davidson, J. Fontaine, J. L. Hor, Y.-R. Huang, Y. Jiang, N. C. Keim, K. D. Koshigan, J. A. Lefever, T. Liu, X.-G. Ma, D. J. Magagnosc, E. Morrow, C. P. Ortiz, J. M. Rieser, A. Shavit, T. Still, Y. Xu, Y. Zhang, K. N. Nordstrom, P. E. Arratia, R. W. Carpick, D. J. Durian, Z. Fakhraai, D. J. Jerolmack, Daeyeon Lee, Ju Li, R. Riggleman, K. T. Turner, A. G. Yodh, D. S. Gianola, and Andrea J. Liu. Structure-property relationships from universal signatures of plasticity in disordered solids. *Science*, 358(6366):1033–1037, 2017.

[17] Giulio Biroli. Machine learning glasses. *Nature Physics*, 16(4):373–374, Apr 2020.

[18] E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras, and A. J. Liu. Identifying structural flow defects in disordered solids using machine-learning methods. *Phys. Rev. Lett.*, 114:108001, Mar 2015.

[19] Ekin D. Cubuk, Andrea J. Liu, Efthimios Kaxiras, and Samuel S. Schoenholz. Unifying framework for strong and fragile liquids via machine learning: a study of liquid silica, 2020.

[20] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E. D. Cubuk, S. S. Schoenholz, A. Obika, A. W. R. Nelson, T. Back, D. Hassabis, and P. Kohli. Unveiling the predictive power of static structure in glassy systems. *Nature Physics*, 16(4):448–454, Apr 2020.

[21] Emanuele Boattini, Frank Smallenburg, and Laura Filion. Averaging local structure to predict the dynamic propensity in supercooled liquids. *Phys. Rev. Lett.*, 127:088007, Aug 2021.

[22] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard,

Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

[23] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, Apr 1980.

[24] Ludovic Berthier and Giulio Biroli. *Glasses and Aging, A Statistical Mechanics Perspective on*, pages 4209–4240. Springer New York, New York, NY, 2009.

[25] John C. Mauro. Grand challenges in glass science. *Frontiers in Materials*, 1, 2014.

[26] Giulio Biroli and Juan P. Garrahan. Perspective: The glass transition. *The Journal of Chemical Physics*, 138(12):12A301, 03 2013.

[27] C. A. Angell. Formation of glasses from liquids and biopolymers. *Science*, 267(5206):1924–1935, 1995.

[28] Sean A. Ridout and Andrea J. Liu. The dynamics of machine-learned "softness" in supercooled liquids describe dynamical heterogeneity, 2024.

[29] Riitta Hari and Lauri Parkkonen. The brain timewise: how timing shapes and supports brain function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668):20140170, 2015.

[30] Francesco Fumarola, Bettina Hein, and Kenneth D. Miller. Mechanisms for spontaneous symmetry breaking in developing visual cortex. *Phys. Rev. X*, 12:031024, Aug 2022.

[31] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933.

[32] VA Marchenko and LA Pastur. Распределение собственных значений в некоторых ансамблях случайных матриц [Distribution of eigenvalues for some sets of random matrices]. *Mat. Sb*, 72:507–536, 1967. in Russian.

[33] Trevor J. Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[34] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295 – 327, 2001.

[35] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643 – 1697, 2005.

[36] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin and. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[37] Haipeng Shen and Jianhua Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

[38] William F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

[39] Juliana Adeola Adisa, Samuel Olusegun Ojo, Pius Adewale Owolawi, and Agnieta Beatrijs Pretorius. Financial distress prediction: Principle component analysis and artificial neural networks. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–6, 2019.

[40] Robert A Eisenbeis. Pitfalls in the application of discriminant analysis in business, finance, and economics. *The journal of finance*, 32(3):875–900, 1977.

[41] J. F. Muzy, J. Delour, and E. Bacry. Modelling fluctuations of financial time series: from cascade process to stochastic volatility model. *The European Physical Journal B - Condensed Matter and Complex Systems*, 17(3):537–548, Oct 2000.

[42] Jianqing Fan and Yazhen Wang. Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102(480):1349–1362, 2007.

[43] Rongyan Liu, Lingyun He, Yufei Xia, Yating Fu, and Ling Chen. Research on the time-varying effects among green finance markets in china: A fresh evidence from multi-frequency scale perspective. *The North American Journal of Economics and Finance*, 66:101914, 2023.

[44] Jozef Barunik, Tomaso Aste, T. Di Matteo, and Ruipeng Liu. Understanding the source of multifractality in financial markets. *Physica A: Statistical Mechanics and its Applications*, 391(17):4234–4251, Sep 2012.

[45] Oliver D Bunn and Robert J Shiller. Changing times, changing values: A historical analysis of sectors within the us stock market 1872-2013. Working Paper 20370, National Bureau of Economic Research, August 2014.

[46] Owain Ap Gwilym and Mike Buckle. Volatility forecasting in the framework of the option expiry cycle. *The European Journal of Finance*, 5(1):73–94, 1999.

[47] Sophie Xiaoyan Ni, Neil D. Pearson, and Allen M. Poteshman. Stock price clustering on option expiration dates. *Journal of Financial Economics*, 78(1):49–87, 2005.

[48] Ashadun Nobi and Jae Woo Lee. State and group dynamics of world stock market by principal component analysis. *Physica A: Statistical Mechanics and its Applications*, 450:85–94, 2016.

[49] Monica Billio, Mila Getmansky, Andrew W. Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559, 2012. Market Institutions, Financial Market Risks and Financial Crisis.

[50] Nick James, Max Menzies, and Georg A. Gottwald. On financial market correlation structures and diversification benefits across and within equity sectors. *Physica A: Statistical Mechanics and its Applications*, 604:127682, 2022.

[51] Mingjie Wang, Siyuan Wang, Jianxiong Guo, and Weijia Jia. Improving stock trend prediction with pretrain multi-granularity denoising contrastive learning. *Knowledge and Information Systems*, 66(4):2439–2466, Apr 2024.

[52] Daniel J. Fenn, Mason A. Porter, Stacy Williams, Mark McDonald, Neil F. Johnson, and Nick S. Jones. Temporal evolution of financial-market correlations. *Phys. Rev. E*, 84:026109, Aug 2011.

[53] Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

[54] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[55] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[56] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

[57] Herman Wold. Causal flows with latent variables: Partings of the ways in the light of nipals modelling. *European Economic Review*, 5(1):67–86, 1974.

[58] Arabind Swain, Sean Alexander Ridout, and Ilya Nemenman. Machine learning that predicts well may not learn the correct physical descriptions of glassy systems. *Phys. Rev. Res.*, 6:033091, Jul 2024.

[59] Tomilola M Obadiya and Daniel M Sussman. Using fluid structures to encode predictions of glassy dynamics. *Physical Review Research*, 5(4):043112, 2023.

[60] Ning Sun, Jinmin Yi, Pengfei Zhang, Huitao Shen, and Hui Zhai. Deep learning topological invariants of band insulators. *Phys. Rev. B*, 98:085402, Aug 2018.

[61] M. D. Ediger. Spatially Heterogeneous Dynamics in Supercooled Liquids. *Annual Review of Physical Chemistry*, 51(1):99–128, 2000.

[62] Walter Kob, Claudio Donati, Steven J. Plimpton, Peter H. Poole, and Sharon C. Glotzer. Dynamical Heterogeneities in a Supercooled Lennard-Jones Liquid. *Physical Review Letters*, 79(15):2827–2830, 1997.

[63] Indrajit Tah and Smarajit Karmakar. Signature of dynamical heterogeneity in spatial correlations of particle displacement and its temporal evolution in supercooled liquids. *Phys. Rev. Res.*, 2:022067, Jun 2020.

[64] Indrajit Tah, Sean A. Ridout, and Andrea J. Liu. Fragility in glassy liquids: A structural approach based on machine learning. *The Journal of Chemical Physics*, 157(12):124501, 09 2022.

[65] Gerhard Jung, Giulio Biroli, and Ludovic Berthier. Predicting dynamic heterogeneity in glass-forming liquids by physics-inspired machine learning. *Phys. Rev. Lett.*, 130:238202, Jun 2023.

[66] Gerhard Jung, Giulio Biroli, and Ludovic Berthier. Dynamic heterogeneity at the experimental glass transition predicted by transferable machine learning. *Phys. Rev. B*, 109:064205, Feb 2024.

[67] Francesco Saverio Pezzicoli, Guillaume Charpiat, and François P. Landes. Rotation-equivariant graph neural networks for learning glassy liquids representations, 2023.

[68] Xiao Jiang, Zean Tian, Kenli Li, and Wangyu Hu. A geometry-enhanced graph neural network for learning the smoothness of glassy dynamics from static structure. *The Journal of Chemical Physics*, 159(14):144504, 10 2023.

[69] Rinske M. Alkemade, Emanuele Boattini, Laura Filion, and Frank Smallenburg. Comparing machine learning techniques for predicting glassy dynamics. *The Journal of Chemical Physics*, 156(20):204503, 2022.

[70] Rinske M. Alkemade, Frank Smallenburg, and Laura Filion. Improving the prediction of glassy dynamics by pinpointing the local cage. *The Journal of Chemical Physics*, 158(13):134512, 04 2023.

[71] Ge Zhang, Hongyi Xiao, Entao Yang, Robert J. S. Ivancic, Sean A. Ridout, Robert A. Riggleman, Douglas J. Durian, and Andrea J. Liu. Structuro-elastoplasticity model for large deformation of disordered solids. *Phys. Rev. Res.*, 4:043026, Oct 2022.

[72] Hongyi Xiao, Ge Zhang, Entao Yang, Robert Ivancic, Sean Ridout, Robert Riggleman, Douglas J. Durian, and Andrea J. Liu. Identifying microscopic factors that influence ductility in disordered solids. *Proceedings of the National Academy of Sciences*, 120(42):e2307552120, 2023.

[73] S. A. Ridout, I. Tah, and A. J. Liu. Building a "trap model" of glassy dynam-

ics from a local structural predictor of rearrangements. *Europhysics Letters*, 144(4):47001, dec 2023.

[74] Vladimir Vapnik, Isabel Guyon, and Trevor Hastie. Support vector machines. *Mach. Learn*, 20(3):273–297, 1995.

[75] B. Scholkopf, Kah-Kay Sung, C.J.C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765, 1997.

[76] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, Jan 2002.

[77] Matt Harrington, Andrea J. Liu, and Douglas J. Durian. Machine learning characterization of structural defects in amorphous packings of dimers and ellipses. *Phys. Rev. E*, 99:022903, Feb 2019.

[78] Robert J. S. Ivancic and Robert A. Riggleman. Identifying structural signatures of shear banding in model polymer nanopillars. *Soft Matter*, 15:4548–4561, 2019.

[79] N. Heckert, James Filliben, C Croarkin, B Hembree, William Guthrie, P Tobias, and J Prinz. Handbook 151: Nist/sematech e-handbook of statistical methods, 2002-11-01 00:11:00 2002.

[80] Darren George and Paul Mallery. Spss for windows step by step: A simple guide and reference. 1998.

[81] J.F. Hair, W.C. Black, and B.J. Babin. *Multivariate Data Analysis: A Global Perspective*. Global Edition. Pearson Education, 2010.

[82] Thomas Burdenski. Evaluating univariate, bivariate, and multivariate normality using graphical procedures. *Mult. Lin. Regression Viewpoints*, 26, 02 2002.

[83] Barbara M. Byrne. *Structural Equation Modeling with Mplus, Basic Concepts, Applications, and Programming.* 2013.

[84] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[85] Daniel A Roberts, Sho Yaida, and Boris Hanin. *The principles of deep learning theory.* Cambridge University Press Cambridge, MA, USA, 2022.

[86] Arabind Swain, Sean Alexander Ridout, and Ilya Nemenman. Distribution of singular values in large sample cross-covariance matrices, 2025.

[87] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists.* Cambridge University Press, 2020.

[88] J.-P. Bouchaud, L. Laloux, M. A. Miceli, and M. Potters. Large dimension forecasting models and random singular value spectra. *The European Physical Journal B*, 55(2):201–207, Jan 2007.

[89] Florent Benaych-Georges, Jean-Philippe Bouchaud, and Marc Potters. Optimal cleaning for singular values of cross-covariance matrices. *The Annals of Applied Probability*, 33(2):1295 – 1326, 2023.

[90] Nikan Firoozye, Vincent Tan, and Stefan Zohren. Canonical portfolios: Optimal asset and signal combination. *Journal of Banking & Finance*, 154:106952, 2023.

[91] Philipp Fleig and Ilya Nemenman. Statistical properties of large data sets with linear latent features. *Phys. Rev. E*, 106:014102, Jul 2022.

[92] Jason W. Rocks and Pankaj Mehta. Bias-variance decomposition of overparameterized regression with random linear features. *Phys. Rev. E*, 106:025304, Aug 2022.

[93] Z. Burda, A. Jarosz, G. Livan, M. A. Nowak, and A. Swiech. Eigenvalues and singular values of products of rectangular gaussian random matrices. *Phys. Rev. E*, 82:061114, Dec 2010.

[94] Craig A. Tracy and Harold Widom. Level-spacing distributions and the airy kernel. *Physics Letters B*, 305(1):115–118, 1993.

[95] Eslam Abdelaleem, Ahmed Roman, K. Michael Martini, and Ilya Nemenman. Simultaneous dimensionality reduction: A data efficient approach for multimodal representations learning. *Transactions on Machine Learning Research*, 2024.

[96] Anne E. Urai, Brent Doiron, Andrew M. Leifer, and Anne K. Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 25(1):11–19, Jan 2022.

[97] Angelique C. Paulk, Yoav Kfir, Arjun R. Khanna, Martina L. Mustroph, Eric M. Trautmann, Dan J. Soper, Sergey D. Stavisky, Marleen Welkenhuysen, Barundeb Dutta, Krishna V. Shenoy, Leigh R. Hochberg, R. Mark Richardson, Ziv M. Williams, and Sydney S. Cash. Large-scale neural recordings with single neuron resolution using neuropixels probes in human cortex. *Nature Neuroscience*, 25(2):252–263, Feb 2022.

[98] Greg J Stephens, Bethany Johnson-Kerner, William Bialek, and William S Ryu. Dimensionality and dynamics in the behavior of c. elegans. *PLoS Comput Biol*, 4(4):e1000028, 2008.

[99] Gordon J. Berman, Daniel M. Choi, William Bialek, and Joshua W. Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672, 2014.

[100] Jie Huang, Xinming Liang, Yuankai Xuan, Chunyu Geng, Yuxiang Li, Haorong Lu, Shoufang Qu, Xianglin Mei, Hongbo Chen, Ting Yu, Nan Sun, Junhua Rao, Jiahao Wang, Wenwei Zhang, Ying Chen, Sha Liao, Hui Jiang, Xin Liu, Zhaopeng Yang, Feng Mu, and Shangxian Gao. A reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience*, 6(5):gix024, 04 2017.

[101] Chen Meng, Bernhard Kuster, Aedín C. Culhane, and Amin Moghaddas Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15(1):162, May 2014.

[102] Michael Sinhuber, Kasper Van Der Vaart, Rui Ni, James G Puckett, Douglas H Kelley, and Nicholas T Ouellette. Three-dimensional time-resolved trajectories from laboratory insect swarms. *Scientific Data*, 6(1):1–8, 2019.

[103] Anthony I. Dell, John A. Bender, Kristin Branson, Iain D. Couzin, Gonzalo G. de Polavieja, Lucas P.J.J. Noldus, Alfonso Pérez-Escudero, Pietro Perona, Andrew D. Straw, Martin Wikelski, and Ulrich Brose. Automated image-based tracking and its application in ecology. *Trends in Ecology & Evolution*, 29(7):417–428, 2014.

[104] Anirvan M Sengupta and Partha P Mitra. Distributions of singular values for some random matrices. *Physical Review E*, 60(3):3389, 1999.

[105] Philippe Loubaton and Pascal Vallet. Almost Sure Localization of the Eigenvalues in a Gaussian Information Plus Noise Model. Application to the Spiked Models. *Electronic Journal of Probability*, 16(none):1934 – 1959, 2011.

[106] M Capitaine and C Donati-Martin. Spectrum of deformed random matrices and free probability, 2016.

[107] Itamar D. Landau, Gabriel C. Mel, and Surya Ganguli. Singular vectors of sums of rectangular random matrices and optimal estimation of high-rank signals: The extensive spike model. *Phys. Rev. E*, 108:054129, Nov 2023.

[108] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024 – 1060, 2012.

[109] Olivier Ledoit and Sandrine P'ech'e. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151:233–264, 2009.

[110] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

[111] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[112] Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin. On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1):459–552, 2016.

[113] Xiucai Ding and Hong Chang Ji. Spiked multiplicative random matrices and principal components. *Stochastic Processes and their Applications*, 163:25–60, 2023.

[114] Xiucai Ding and Fan Yang. Spiked separable covariance matrices and principal components. *The Annals of Statistics*, 49(2):1113 – 1138, 2021.

[115] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[116] Farzad Pourkamali and Nicolas Macris. Rectangular rotational invariant estimator for high-rank matrix estimation, 2024.

[117] Anna Choromanska, MIkael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. The Loss Surfaces of Multilayer Networks. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 192–204, San Diego, California, USA, 09–12 May 2015. PMLR.

[118] Antonio Auffinger, Gérard Ben Arous, and Jiří Černý. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

[119] Yatin Dandi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The computational advantage of depth: Learning high-dimensional hierarchical functions with gradient descent, 2025.

[120] Max Staats, Matthias Thamm, and Bernd Rosenow. Small singular values matter: A random matrix analysis of transformer models, 2025.

[121] Charles H. Martin, Tongsu (Serena) Peng, and Michael W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):4122, Jul 2021.

[122] DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang,

Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.