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:

_____ _____

Zhe Sun Date

Estimating Genetic Effects When Stratification-Score Matching Is Used to Correct for Confounding by Population Stratification in Case-Control Studies

By

Zhe Sun

Master of Science in Public Health

Emory University

Rollins School of Public Health

Department of Biostatistics and Bioinformatics

_____

Thesis Advisor

_____

Reader

Estimating Genetic Effects When Stratification-Score Matching Is Used to Correct for Confounding by Population Stratification in Case-Control Studies

By

Zhe Sun

Bachelor of Science

Fudan University

2012

Advisor: Yijuan Hu, Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2014

Abstract

Estimating Genetic Effects When Stratification-Score Matching Is Used to Correct for Confounding by Population Stratification in Case-Control Studies

By Zhe Sun

Case-control studies are most frequently used to investigate the association between the risk of developing a particular disorder and the genetic variation. This association may be confounded by population stratification, i.e., when genetic variation is correlated with variation in disease risk across latent subpopulations in the case-control sample. Failure to properly account for this confounding can lead to false associations between the genetic markers and disease. An efficient correction proposed by Epstein et al. (2007, 2012) is to infer the ancestry by principal components of the sample correlation matrix of SNP genotypes, and fine-match the case-control samples by the stratification score, which is the probability of disease given genomic variables. However, this approach only provides hypothesis testing of the association but not estimation of the genetic effects. In this thesis, we propose a novel estimation method based on the fine-matched case-control sample. Extensive simulation studies were carried out to evaluate the performance of the proposed method and compare with a few alternative strategies. The simulation results demonstrate little bias of the proposed estimator, even when there is a strong association between the ancestry and the genetic marker under study.

Estimating Genetic Effects When Stratification-Score Matching Is Used to Correct for Confounding by Population Stratification in Case-Control Studies

By

Zhe Sun

Bachelor of Science

Fudan University

2012

Advisor: Yijuan Hu, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2014

Table of Contents

1. Introduction

The association between disease and genetic variation are of main interest in contemporary genetic studies. For example, Genome-wide association studies (GWAS) can be used to identify the association between the risk of developing a particular disorder and single nucleotide polymorphisms (SNPs). There are mainly two types of study designs in examining the genetic effects: prospective cohort studies and retrospective case-control studies. Confounding factors are those related to both the diseases and the genetic variable. They should be accounted for because they can distort the true relationship between the disease and the genetic variable (Knowler et al., 1988). Since genetic studies tend to include samples with heterogeneous ancestral backgrounds, population stratification is the main confounding factor.

For prospective studies, Rosenbaum and Rubin (1983) proposed a propensity-score-matching approach. Stratification on the propensity score, which is the probability of an exposure conditional on confounding variables, can remove confounding when examining the relationship between a binary exposure and disease.

For retrospective case-control studies, many methods have been proposed to correct for population stratification. One of the strategies uses large sets of genetic markers to derive ancestry components, which are often inferred as the principal components of the sample correlation matrix of SNP genotypes (Chen et al., 2003; Patterson et al., 2006; Price et al., 2006). Lee et al. (2009) proposed to use spectral-graph theory to form ancestry components consisting of eigenvectors derived

from the normalized Laplacian matrix of the sample genotype data. We can include all of the ancestry components in the model and directly adjust for these covariates to correct for population stratification. Such a “direct adjustment” method assumes that disease risk is a linear or log-linear function of ancestry components.

In order to relax the assumption of a parametric relationship between the disease and the ancestry components, we can stratify cases and controls based on ancestry components. Conditional logistic regression models or Cochran-Mantel-Haenszel statistic can be used to test the association. The stratification is performed by comparing the dissimilarity of ancestry components between different subjects in a case-control study. Those with similar ancestry components are classified into one stratum. Such a tight matching strategy is advantageous in GWAS if a study recruits controls from large database. It ensures that only controls that have similar ancestry to those of cases are used in the analysis. When there are outliers, the matching strategy can provide a correction for confounding effect according to Luca et al. (2008) and Allen et al. (2010).

In order to determine genetic dissimilarity, Luca et al. (2008) proposed to use Euclidean distance between the principal components of case and control participants (referred to as the GEM approach). Lee et al. (2009) further proposed to replace the principal components with significant ancestry components derived by spectral-graph approach (referred to as the Spectral-GEM Approach). Guan et al. (2009) developed an approach named GSM, which matches subjects based on the average proportion of alleles (weighted by allele frequency) shared identical-by-state (IBS) over tens of

thousands of SNPs. However, without identifying the contribution of ancestry components to confounding, the above methods will lead to unsuitable matching when the uncorrelated ancestry components are included in the model.

Epstein et al. (2007) proposed a novel measurement of dissimilarity based on the estimated odds of disease given the principal components for ancestry. The new measurement is referred to as the stratification score. Epstein et al. (2012) showed that the fine matching based on the stratification score can provide better correction for population stratification compared with the GEM, Spectral-GEM and GSM approaches. The stratification score is univariate while the other approaches attempt to match on multiple quantities simultaneously. The stratification score lowers the contribution of those uncorrelated ancestry components to the population stratification. It can provide a proper matching when there are large numbers of uncorrelated components of ancestry.

However, Epstein et al. (2012) did not provide a solution to the parameter estimates adjusting for ancestry components. We propose a novel estimation method based on the fine-matched case-control sample. In each stratum, we do “direct adjustment” by including ancestry components in the logistic model and estimate the genetic effect by conditional logistic regression. Due to non-collapsibility, the estimation of the genetic effect on disease proposed by Epstein et al. (2012) is problematic when there are multiple ancestry components. This estimation may be different from the estimation of the marginal model or the estimation of the conditional model adjusting for the ancestry components because we estimate the

population stratification based on the stratification score which is univariate.

In the remainder of this thesis, we review the method of conditional logistic regression, stratification score and fine matching approach in Section 2. In Section 3, extensive simulations will be presented to show the performance of our method. We will also compare alternative methods in estimating the genetic effects. In Section 4, we will discuss the advantages and limitations of our method.

2. Methods

2.1 Ancestry Components and Stratification Score

Confounding due to population stratification can lead to biased estimation of the association between the disease and the genetic variable. In order to solve such a problem, a common approach is to utilize genetic markers, typically single-nucleotide polymorphisms (SNPs), to derive influential components of ancestry. Eigenvectors based on the principal components of the sample correlation matrix of SNP genotypes are often used as the inference of ancestry components (Chen et al., 2003; Patterson et al., 2006; Price et al., 2006). We can then analyze the association between the disease and the genotype adjusted for the information of ancestry components.

Epstein et al. (2012) proposed to use the stratification score, a scalar measurement based on potential confounders such as ancestry components, to correct for confounding. Let $\boldsymbol{C_j} = (C_{j,1}, C_{j,2}, \dots, C_{j,q})$ denote the ancestry components of the jth patients, which is composed of q principal components derived by the procedure of Patterson et al. (2006). Suppose $D_j = 1$ if the jth patient had the disease and $D_j = 0$ otherwise. Then the stratification score is defined as

$$\Theta(\boldsymbol{C_j}) = \frac{P(D_j=1|\boldsymbol{C_j})}{P(D_j=0|\boldsymbol{C_j})} = e^{\alpha+\boldsymbol{\gamma}^T\boldsymbol{C_j}} \quad , \tag{1}$$

which is the odds of disease given the ancestry components. We fit model (1) to the case-control data to estimate α and $\boldsymbol{\gamma}$. Let $\hat{\alpha}$ and $\hat{\boldsymbol{\gamma}}$ denote the maximum likelihood estimator of α and $\boldsymbol{\gamma}$. We estimate the stratification score through

$$\hat{\theta}(\boldsymbol{C_j}) = e^{\hat{\alpha} + \hat{\boldsymbol{\gamma}}^{\boldsymbol{T}} \boldsymbol{C_j}} \quad . \tag{2}$$

The estimated stratification score is used to stratify the population and correct the confounding effect.

2.2 Fine Matching

After deriving the estimated stratification score, we perform fine matching of cases and controls using a matching approach proposed by Rosenbaum and Rubin (1985). Let U_{ir} be the dissimilarity measure between a case i $(i = 1, \ldots., N_{case})$ and a control r $(r = 1, \ldots., N_{control})$. It is defined as

$$U_{ir} = \frac{|\log(\hat{\theta}_i) - \log(\hat{\theta}_r)|}{\sqrt{\frac{1}{N-2}\{(N_{case}-1)sd_1[\log(\hat{\theta})] + (N_{control}-1)sd_0[\log(\hat{\theta})]\}}} \quad , \tag{3}$$

where N_{case} and $N_{control}$ are the sample size of cases and controls, respectively, N is the total sample size, $\log(\hat{\theta}_i)$ is the estimated log-transformed stratification score for the ith case, $\log(\hat{\theta}_r)$ is the estimated log-transformed stratification score for the rth control, and $sd_1[\log(\hat{\theta})]$ and $sd_0[\log(\hat{\theta})]$ are the standard deviations of estimated log-transformed stratification score in the cases and controls, respectively. The estimated stratification score is used to minimize T, which is defined as

$$T = \sum_{l=1}^{L} \sum_{i \epsilon [1, N_{case}], r \epsilon [1, N_{control}]} U_{ir} \quad . \tag{4}$$

Rosenbaum (1991) showed that T could be minimized if one case and ≥ 1 controls or one control and ≥ 1 cases are in the same stratum.

2.3 Conditional Logistic Regression Based on Stratification Score

In order to deal with the highly stratified data, Epstein et al. (2012) proposed to perform a Cohran-Mantel Haenszel test to test the association between disease and SNP. They showed that the Cohran-Mantel Haenszel test based on fine matching the stratification score improved the correction for confounding by population stratification, as compared to the GEM and SpectralGEM approaches which directly utilized all the significant ancestry components information.

In this section, we extend the approach of Epstein et al. (2012) and propose a novel estimation method based on the fine-matched case-control sample. Suppose that the genotype of the jth patient is denoted by G_j and this patient is in the stratum k based on the fine matching approach. There are a total of K strata after fine matching. At first, it may seem natural to estimate the magnitude of the association between D and G by fitting the model

$$P(D_j = 1|\{\boldsymbol{S}_k\}, G_j) = \frac{e^{\alpha'_k + \beta G_j}}{1+e^{\alpha'_k + \beta G_j}} \quad , \tag{5a}$$

where $\{\boldsymbol{S}_k\}$ denotes the set of strata. However, the value of β obtained using this model may not correspond to either the value β obtained by fitting the marginal model

$$P(D_j = 1|\{\boldsymbol{S}_k\}, G_j) = \frac{e^{\boldsymbol{\alpha} + \beta G_j}}{1+e^{\alpha + \beta G_j}} \quad , \tag{5b}$$

or the conditional model

$$P(D_j = 1|\boldsymbol{C}'_{\boldsymbol{j}}, G_j) = \frac{e^{\alpha^* + \boldsymbol{\gamma}^{*\boldsymbol{T}}\boldsymbol{C_j} + \beta G_j}}{1+e^{\alpha^* + \boldsymbol{\gamma}^{*\boldsymbol{T}}\boldsymbol{C_j} + \beta G_j}} \ , \tag{5c}$$

due to non-collapsibility (Allen et al. 2010). Here we assume that the parameter β obtained by fitting the conditional model (5c) is of primary interest, but that a matched analysis is desired for control of confounding.

The approach proposed here is to modify (5a) to explicitly include some of the ancestry components in the model. If matching on the stratification score is very tight, then all members of a stratum will have the same value of $\widehat{\boldsymbol{\gamma}}^{\boldsymbol{T}}\boldsymbol{C_j}$. As a result, we cannot add all q ancestry components as they are collinear within each stratum. Thus, we use $\boldsymbol{C_j'} = (C_{j,1}, \dots, C_{j,q-1})$ instead of $\boldsymbol{C_j} = (C_{j,1}, \dots, C_{j,q-1}, C_{j,q})$. If stratification is not very tight, then it may be possible to include all q components of ancestry. Thus, the model we use here is

$$P(D_j = 1|\{\boldsymbol{S}_k\}, \boldsymbol{C_j'}, G_j) = \frac{e^{\alpha_k^* + \boldsymbol{\gamma^{*T}}\boldsymbol{C_j'} + \beta G_j}}{1 + e^{\alpha_k^* + \boldsymbol{\gamma^{*T}}\boldsymbol{C_j'} + \beta G_j}}\ , \qquad (5d)$$

Model (5d) is fitted by the maximum likelihood approach. However, the α_k^* is difficult to estimate if the sample size is small in each stratum. Thus, the conditional likelihood approach (Cox, 1970) is used to eliminate nuisance parameter α_k^* and borrow information from all the strata to estimate the common genotype effect β. Let $\boldsymbol{V}$ denote the $1 \times p$ vector $(\alpha_1, \alpha_2, \dots, \alpha_K, \dots, \beta, \boldsymbol{\gamma^{*T}})$. Suppose that in the kth stratum, there are N_k subjects. The full likelihood for the kth stratum based on our model is

$$L(\boldsymbol{V}|\boldsymbol{D}) = \prod_{j=1}^{N_k} \left\{ \frac{e^{\alpha_k^* + \boldsymbol{\gamma^{*T}}\boldsymbol{C_j'} + \beta G_j}}{1 + e^{\alpha_k^* + \boldsymbol{\gamma^{*T}}\boldsymbol{C_j'} + \beta G_j}} \right\}^{D_j} \left\{ 1 - \frac{e^{\alpha_k^* + \boldsymbol{\gamma^{*T}}\boldsymbol{C_j'} + \beta G_j}}{1 + e^{\alpha_k^* + \boldsymbol{\gamma^{*T}}\boldsymbol{C_j'} + \beta G_j}} \right\}^{1 - D_j}\ , \qquad (6)$$

where $\boldsymbol{D}=\left(D_1,\dots,D_{N_k}\right)$. Suppose that in this stratum, only the jth subject is a case. By Bayes' rule, the conditional probability of the observed results in this stratum is

$$P\left(D_1=0,\dots,D_j=1,\dots,D_{N_k}=0\middle|D_1+D_2+\cdots D_{N_k}=1\right)$$

$$=\frac{P(D_1=0,\dots,D_j=1,\dots,D_{N_k}=0)}{P\left(D_1=1,D_2=0,\dots,D_{N_k}=0\right)+\cdots+P(D_1=0,\dots,D_{N_k}=1)} \quad . \tag{7}$$

Since we assume the mutually independence of the N_k subjects in this stratum, we have

$$P\left(D_1=0,\dots,D_j=1,\dots,D_{N_k}=0\right)=P(D_1=0)\dots P\left(D_j=1\right)\dots P\left(D_{N_k}=0\right) \quad . \tag{8}$$

Thus, the conditional probability in equation (7) can be simplified as

$$P\left(D_1=0,\dots,D_j=1,\dots,D_{N_k}=0\middle|D_1+\cdots+D_{N_k}=1\right)$$

$$=\frac{\exp\left(\boldsymbol{\gamma^T C_j'}+\beta G_j\right)}{\exp(\boldsymbol{\gamma^T C_1'}+\beta G_1)+\cdots+\exp\left(\boldsymbol{\gamma^T C_{N_k}'}+\beta G_{N_k}\right)} \quad . \tag{9}$$

For a stratum m where there is only one control (the j' th subject), when the total number of subjects is N_m, the conditional probability of the observed results in this stratum is

$$P\left(D_1=1,\dots,D_{j'}=0,\dots,D_{N_m}=1\middle|D_1+\cdots+D_{N_m}=N_m-1\right)$$

$$=\frac{\exp(\boldsymbol{\gamma^T C_1'}+\beta G_1)\dots\exp\left(\boldsymbol{\gamma^T C_{j'-1}'}+\beta G_{j'-1}\right)\exp(\boldsymbol{\gamma^T C_{j'+1}'}+\beta G_{j'+1})\dots\exp(\boldsymbol{\gamma^T C_{N_m}'}+\beta G_{N_m})}{\exp(\boldsymbol{\gamma^T C_2'}+\beta G_2)\dots\exp\left(\boldsymbol{\gamma^T C_{N_m}'}+\beta G_{N_m}\right)+\cdots+\exp(\boldsymbol{\gamma^T C_1'}+\beta G_1)\dots\exp\left(\boldsymbol{\gamma^T C_{N_m-1}'}+\beta G_{N_m-1}\right)} \quad . \tag{10}$$

The equation (9) and (10) only depends on $\boldsymbol{\gamma^T}$ and β. The maximum likelihood estimator of β can be derived by maximizing conditional likelihood based on equation (9) and (10).

Our method utilizes the stratification score stratum information. In each stratum, we fit a conditional logistic regression directly adjusting for ancestry components. Compared with the method proposed by Epstein et al. (2012), our method can estimate the genetic effect on disease adjusting for ancestry components. Extensive simulations will be presented in section 3.

2.4 Software Implementation

We implemented our study in R code using existing R packages. The stratification score is derived by glm() function in R package. “optmach” package is used to calculate the stratum information. Note that, because the package has been updated, mdist() is used to calculate the dissimilarity score. All functionality of the pscore.dist() function (used by Epstein et al., 2012) has been moved into to the mdist() function. fullmatch() can be useful in fine matching approach. The main analysis, conditional logistic regression, is implemented by “survival” package. The clogit() function within survival package can be used to fit model (5).

3. Results

3.1 Simulation Studies

To evaluate the performance of the proposed approach, we conducted simulation studies. In each replicate, we enrolled 500 cases and 500 controls to form a case-control study. For simplicity, we assumed the ancestry components to be a two-dimensional vector. Each component of the vector is independent and follows the standard normal distribution.

To generate the genotype at the test locus, we assumed that the genotypes follow Hardy-Weinberg Equilibrium conditional on the ancestry components. The odds of possessing of the minor SNP allele over the major SNP allele $\psi(\boldsymbol{C_j})$ is modeled through

$$log\ \psi(\boldsymbol{C_j}) = \xi + \boldsymbol{\eta^T C_j} \quad , \tag{8}$$

where $\boldsymbol{\eta}$ denotes a two-dimensional vector of coefficients corresponding to the two significant principal components. We set the overall MAF to be 0.2 at this locus so that $\xi = \log(0.2/0.8)$. The disease status is generated through

$$P(D_j = 1) = \frac{e^{\alpha+\boldsymbol{\gamma^{*T}C_j}+\beta G_j}}{1+e^{\alpha+\boldsymbol{\gamma^{*T}C_j}+\beta G_j}} \ . \tag{9}$$

The prevalence of the disease is assumed to be 0.05, so that $\alpha = \log(0.05/0.95)$.

In each simulation, subjects with disease were discarded after we enrolled 500 cases. Subjects without disease were discarded from the simulated dataset after we enrolled 500 controls. The whole process of sampling would not end until all 1000 subjects enrolled in the case-control study.

Five methods (model 1 to 5) have been compared in the following section.

1. Unconditional Logistic Model (UL):

$$logit\ P(D_j = 1) = \alpha + \beta G_j\ ,$$

which is the model not considering the confounding effect.

2. Conditional Logistic Model without Principal Components (CL.NPC):

$$logit\ P(D_j = 1) = \alpha_k + \beta G_j$$

3. Conditional Logistic Model with Two Principal Components (CL.KPC):

$$logit\ P\big(D_j = 1\big) = \alpha_k + \gamma_1 C_{1,j} + \gamma_2 C_{2,j} + \beta G_j$$

4. Conditional Logistic Model with the First Principal Components (CL. PC1):

$$logit\ P\big(D_j = 1\big) = \alpha_k + \gamma_1 C_{1,j} + \beta G_j$$

5. Conditional Logistic Model with the Second Principal Components (CL. PC2):

$$logit\ P\big(D_j = 1\big) = \alpha_k + \gamma_2 C_{2,j} + \beta G_j$$

The parameter of main interest is the genetic effect on disease β. Through 5000 simulations, we evaluated the performance of different models in five aspects: the bias, the sample standard error, the average of estimated standard error, the power and the type I error. We designed four scenarios to evaluate the performance of our method. We defined that $\boldsymbol{\eta}_1 = (0.05, 0.05)^T$ and $\boldsymbol{\gamma}_1 = (0.05, 0.05)^T$; $\boldsymbol{\eta}_2 = (0.1, 0.1)^T$ and $\boldsymbol{\gamma}_2 = (0.1, 0.1)^T$. The first scenario is referred to as S1. S1 is designed to test our method under both weak population stratification and weak association between principal components and genotype (i.e. $\boldsymbol{\eta} = \boldsymbol{\eta}_1$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_1$). Under the second scenario (S2), $\boldsymbol{\eta} = \boldsymbol{\eta}_1$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_2$. The true value of the parameters are changed to $\boldsymbol{\eta} = \boldsymbol{\eta}_2$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_1$ under scenario 3 (S3). The last

scenario (S4) is used to test our method under both stronger population stratification and stronger association between principal components and genotype (i.e., $\boldsymbol{\eta} = \boldsymbol{\eta}_2$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_2$) than those in S1.

3.2 Under the Null Hypothesis

The results of the performance of different methods are presented by Table I. Under S1, the bias of CL.KPC, UL, CL.NPC, CL.PC1 and CL.PC2 are respectively 0.004, 0.010, 0.009, 0.007 and 0.006. Under weak ancestry component effects on disease, the bias of conditional logistic models (model 2-5) is smaller than the bias of UL. The bias of UL increases when there is a stronger ancestry component effect on disease or stronger ancestry component effect on genotype. The bias of UL is respectively 0.014, 0.014 and 0.024 under S2, S3 and S4. In contrast, the bias of conditional logistic models (model 2-5) in these scenarios is smaller, compared with the bias of UL. Thus, the conditional logistic regression models can estimate the genetic effect on disease more precisely than the UL method.

In all scenarios, the bias of CL.NPC is quite close to the bias of UL, which is the limitation mentioned by Epstein et al. (2012). The bias of CL.KPC is the smallest in all scenarios because we fitted the model which we used to generate disease status. Our method (i.e. CL.PC1 and CL.PC2) can estimate a more precise genetic effect on disease, compared with CL.NPC and UL when there is no true genotype effect on disease. Especially under S4, the bias of CL.PC1 and CL.PC2 are both 0.012, which are quite smaller than the bias of UL and CL.NPC (0.24 and 0.021). As shown by

Table I, the SE and SEE are very close for the estimator, which indicates the standard error estimator is accurate.

We evaluated the type I error under different models. The results are shown in Table I. The type I error of conditional logistic models (model 2-5) is similar to each other and close to the nominal significance level 0.05 under all scenarios. The type I error of UL is higher, compared with the conditional logistic models. Under S4, the type I error is 0.063 for UL, which is inflated. In contrast, the type I error are respectively 0.049, 0.051, 0.052 and 0.049 for CL.KPC, CL.NPC, CL.PC1 and CL.PC2. The type I error under CMH test is very close to the type I error of CL.NPC, which means that CL.NPC is equivalent to CMH in terms of hypothesis testing.

3.3 Under the Alternative Hypothesis

In this section, we compared our method with other three models under the alternative hypothesis. The genetic effect on disease β increases from 0.2 to 0.8 in each scenario. The disease status is generated under different values of β by equation (9).

The bias of the CL.NPC method is close to the bias of UL method when the genetic effect on disease increases. Under S4, the bias of CL.NPC is 0.024, 0.028, 0.024 and 0.022, respectively. Meanwhile, CL.KPC, CL.PC1 and CL.PC2 have a much smaller bias when the genetic effect increases, compared with the bias of CL.NPC and UL. The bias of CL.PC1 is 0.015, 0.019, 0.016 and 0.015 respectively. There is a small difference between CL.PC1 and CL.PC2. Our method (i.e. CL.PC1 and CL.PC2) can estimate a more precise genetic effect on disease, compared with

CL.NPC and UL under the alternative hypothesis. CL.KPC still has the smallest bias because the model we fitted is exactly the model we used to generate the disease status. Under other scenarios (S1, S2 and S3), although the difference of bias among different models is smaller than those in S4, CL.PC1 and CL.PC2 can always provide a smaller bias compared with CL.NPC and UL.

Figure I gives us a picture of the bias under S4. The bias of the CL.KPC is the closest to the reference line (i.e. Bias=0). The curves of the bias of UL and CL.NPC model deviate quite far away from the reference line. The curve of the bias is closer to the reference line for CL.PC1 and CL.PC2, compared with UL and CL.NPC. From Table I, under all scenarios, the power of CL.PC1 and CL.PC2 is larger than CL.KPC. Including all ancestry components makes CL.KPC more conservative, compared with our method. Also, under the alternative hypothesis, the power under CMH test is very close to the power of CL.NPC.

4. Discussion

In this thesis, we propose a novel estimation method based on the fine-matched case-control sample. The inclusions of controls with substantial dissimilar ancestry from cases can lead to improper stratification in “direct adjustment” method. The stratification score, proposed by Epstein et al. (2007), was used for fine matching method to correct population confounding effect. The stratification score can upweight the contribution of components of ancestry that are potential confounders to the population stratification, and meanwhile, downweight those components that are not. As a result, it lowers the chance of inaccurate matches for confounding in fine matching approach. After combining these two methods, the bias of parameter estimators is lower under different genetic effects on disease compared with the method proposed by Epstein et al. (2012).

Stratification based on the propensity score will correct confounding effect in prospective studies (Rosenbaum and Rubin, 1983, 1984) when testing the relationship between exposure and disease. The propensity score is defined as the odds of an exposure conditional on confounder variables. Matching on the stratification score in case-control association studies is comparable to matching on the propensity score in prospective studies.

We extended the approach proposed by Epstein et al. (2012) by using conditional logistic regression instead of the Cochran-Mantel-Haenszel (CMH) tests of SNP-disease association. In fact, the CMH test corresponds to the score test for the

association parameter in conditional logistic regression. Meanwhile, after implementing the conditional logistic model, we can estimate the genetic effect on disease adjusting for ancestry components.

The main limitation for our method is that the bias will increase when there is a strong ancestry component effect on disease, leading to residual within-stratum confounding. This residual confounding occurs because the dissimilarity of the stratification score among the subjects in the same stratum is high when the ancestry component effect is strong. As a result, the bias will increase because of the inadequate correction of the confounding. We are working on an alternative matching approach in which we draw a fine-matched control from a general population for each case.

5. Reference

Knowler, W. C., Williams, R. C., Pettitt, D. J., & Steinberg, A. G. (1988). Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American journal of human genetics*, 43(4), 520.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Chen, H. S., Zhu, X., Zhao, H., & Zhang, S. (2003). Qualitative Semi- Parametric Test for Genetic Associations in Case- Control Designs Under Structured Populations. *Annals of human genetics*, *67*(3), 250-264.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, *2*(12), e190.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904-909.

Lee, A. B., Luca, D., Klei, L., Devlin, B., & Roeder, K. (2010). Discovering genetic ancestry using spectral graph theory. *Genetic epidemiology*, *34*(1), 51-59.

Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., et al. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *American Journal of Human Genetics*, *82*(2), 453-463.

Allen, A., Epstein, M. P., & Satten, G. A. (2010). Score-based adjustment for confounding by population stratification in genetic association studies. *Genetic epidemiology*, *34*(5), 383.

Guan, W., Liang, L., Boehnke, M., & Abecasis, G. R. (2009). Genotype- based matching to correct for population stratification in large- scale case- control genetic association studies. *Genetic epidemiology*, *33*(6), 508-517.

Epstein, M. P., Allen, A. S., & Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *American Journal of Human Genetics*, *80*(5), 921-930.

Epstein, M. P., Duncan, R., Broadaway, K. A., He, M., Allen, A. S., & Satten, G. A. (2012). Stratification- Score Matching Improves Correction for Confounding by Population Stratification in Case- Control Association Studies. *Genetic*

epidemiology, *36*(3), 195-205.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 597-610.

Cox, D. R. (1970). Analysis of binary Data. Chapman & Hall.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524.

6. Appendix

TABLE I. Simulation results for studying genetic effect

		CL.KPC				UL				CL.NPC				CL.PC1				CL.PC2				CMH
	β	Bias	SE	SEE	PW	Bias	SE	SEE	PW	Bias	SE	SEE	PW	Bias	SE	SEE	PW	Bias	SE	SEE	PW	PW
S1	0.0	0.004	0.134	0.125	0.058	0.010	0.120	0.112	0.060	0.009	0.133	0.124	0.056	0.007	0.133	0.124	0.058	0.006	0.133	0.124	0.059	0.058
	0.2	0.008	0.129	0.123	0.386	0.011	0.116	0.110	0.495	0.012	0.128	0.122	0.414	0.010	0.129	0.123	0.399	0.010	0.129	0.123	0.397	0.414
	0.4	0.008	0.125	0.122	0.924	0.009	0.114	0.108	0.966	0.012	0.124	0.121	0.930	0.010	0.125	0.122	0.926	0.010	0.125	0.122	0.927	0.930
	0.6	0.009	0.129	0.123	0.999	0.009	0.112	0.108	0.998	0.011	0.128	0.123	0.999	0.010	0.128	0.123	0.999	0.010	0.128	0.123	0.999	0.999
	0.8	0.012	0.133	0.127	1.000	0.009	0.114	0.109	1.000	0.013	0.132	0.126	1.000	0.013	0.132	0.126	1.000	0.013	0.132	0.126	1.000	1.000
S2	0.0	0.003	0.128	0.125	0.052	0.014	0.119	0.112	0.064	0.013	0.126	0.124	0.053	0.008	0.127	0.125	0.050	0.008	0.127	0.125	0.048	0.053
	0.2	0.005	0.130	0.123	0.396	0.015	0.116	0.110	0.504	0.014	0.128	0.122	0.424	0.010	0.129	0.123	0.411	0.009	0.129	0.123	0.408	0.430
	0.4	0.004	0.127	0.123	0.912	0.015	0.113	0.108	0.972	0.012	0.126	0.122	0.923	0.008	0.127	0.122	0.917	0.008	0.126	0.122	0.918	0.924
	0.6	0.004	0.127	0.124	0.999	0.013	0.112	0.108	1.000	0.011	0.126	0.123	0.999	0.008	0.127	0.123	0.999	0.008	0.126	0.123	0.999	0.999
	0.8	0.012	0.133	0.127	1.000	0.012	0.114	0.109	1.000	0.017	0.132	0.126	1.000	0.015	0.133	0.127	1.000	0.014	0.132	0.127	1.000	1.000
S3	0.0	0.003	0.133	0.125	0.062	0.014	0.119	0.112	0.064	0.013	0.132	0.124	0.062	0.009	0.133	0.124	0.061	0.008	0.132	0.124	0.061	0.062
	0.2	0.005	0.127	0.123	0.394	0.015	0.115	0.109	0.515	0.014	0.126	0.122	0.439	0.010	0.127	0.122	0.410	0.009	0.126	0.122	0.408	0.439
	0.4	0.009	0.127	0.122	0.922	0.014	0.112	0.108	0.974	0.017	0.126	0.121	0.933	0.013	0.127	0.122	0.927	0.013	0.126	0.122	0.929	0.933
	0.6	0.007	0.127	0.124	0.998	0.013	0.112	0.108	0.998	0.015	0.126	0.123	0.998	0.011	0.126	0.123	0.998	0.011	0.126	0.123	0.998	0.998
	0.8	0.010	0.130	0.127	1.000	0.013	0.113	0.108	1.000	0.016	0.129	0.126	1.000	0.013	0.130	0.126	1.000	0.013	0.129	0.126	1.000	1.000
S4	0.0	0.002	0.128	0.125	0.049	0.024	0.118	0.112	0.063	0.021	0.126	0.124	0.051	0.012	0.127	0.125	0.052	0.012	0.126	0.125	0.049	0.051
	0.2	0.005	0.130	0.123	0.409	0.024	0.115	0.109	0.541	0.024	0.129	0.122	0.472	0.015	0.130	0.123	0.434	0.014	0.129	0.123	0.442	0.474
	0.4	0.010	0.125	0.123	0.930	0.024	0.112	0.108	0.982	0.028	0.123	0.122	0.952	0.019	0.124	0.122	0.949	0.019	0.124	0.122	0.943	0.954
	0.6	0.007	0.125	0.124	0.999	0.022	0.112	0.108	0.999	0.024	0.124	0.123	0.999	0.016	0.126	0.123	0.999	0.016	0.124	0.123	0.999	0.999
	0.8	0.007	0.129	0.127	1.000	0.022	0.113	0.108	1.000	0.022	0.129	0.126	1.000	0.015	0.129	0.126	1.000	0.015	0.129	0.126	1.000	1.000

S1-S4 denotethe four scenarios mentioned in section 3.2. Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. PW is the type I error/power for the testing zero parameter value at the 0.05 nominal significance level. Each entry is based on 5000 replicates.

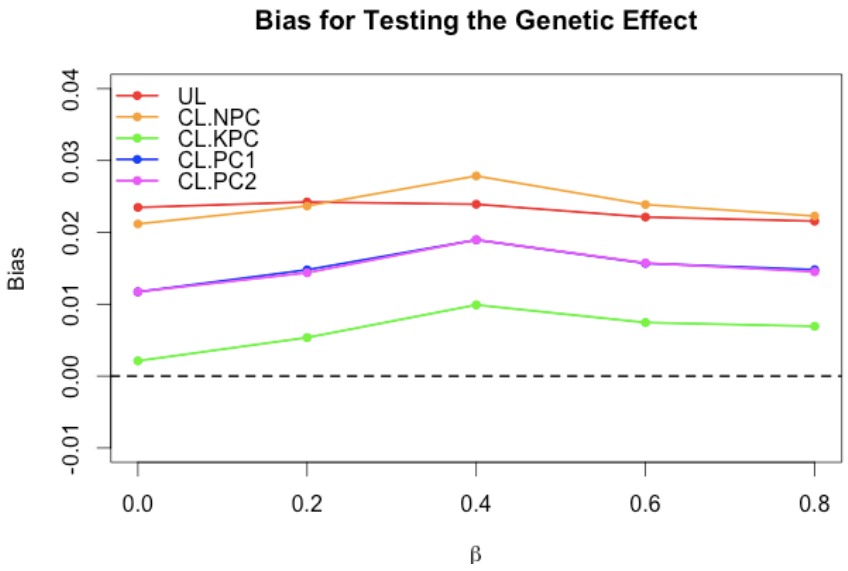

Fig I. Bias for different genetic effect on disease under the condition that $\boldsymbol{\gamma} = (0.1,0.1)^T$ and $\boldsymbol{\eta} = (0.1,0.1)^T$

R Code for Implementing Stratification Score Matching and Fitting Conditional Logistic Regression Models

```
### simulation ###
require(stats)
require(optmatch)
require(survival)
require(coin)

### function prob ###
prob<-function(lp){
      ep<-exp(lp)
      return(ep/(1+ep))
}

### set the number and seed of simulation ###
sim<-5000
sig<-rep(0,sim)
sig_value<-rep(0,sim)

set.seed(123)
seed<-runif(sim)*1000

nPC = 2

effect_ConY<-c(log(0.05/0.95),0.1,0.1)
effect_ConG<-c(log(0.2/0.8),0.1,0.1)
effect_GonY<-0

nCO = 500
nCC = 500

effect_ConY_est = NULL
effect_ConY_sd = NULL

fit.beta<-rep(0,sim)
fit1.beta<-rep(0,sim)
fit2.beta<-rep(0,sim)
fit3.beta<-rep(0,sim)
fit4.beta<-rep(0,sim)

fit.se<-rep(0,sim)
fit1.se<-rep(0,sim)
fit2.se<-rep(0,sim)
```

```
fit3.se<-rep(0,sim)
fit4.se<-rep(0,sim)

fit.p<-rep(0,sim)
fit1.p<-rep(0,sim)
fit2.p<-rep(0,sim)
fit3.p<-rep(0,sim)
fit4.p<-rep(0,sim)

### start the loop ###
for (z in 1:sim){
    set.seed(seed[z])

    print(z)

    pc<-NULL
    g<-NULL
    dis<-NULL

    nCO_cum = 0
    nCC_cum = 0

    alltable<-NULL

    while (nCO_cum+nCC_cum < nCO+nCC){

        ### generate the ancestry components ###
        cn<-rnorm(nPC)

        ### generate the genotype at the test locus ###

        g.lp<-sum(c(1,cn)*effect_ConG)
        sim.g<-prob(g.lp)
        minor_allele<-runif(2)
        gn<-(minor_allele[1]<sim.g)+(minor_allele[2]<sim.g)

        ### generate the disease outcome ###
        d.lp<-sum(c(1,cn)*effect_ConY) + gn*effect_GonY
        sim.p<-prob(d.lp)
        pt<-runif(1)
        disn<-(pt<sim.p)*1

        if ( (nCC_cum < nCC) & (disn == 1) ) {
            nCC_cum = nCC_cum+1;
```

```
            g=c(g,gn)
            dis=c(dis,disn)
            pc=rbind(pc,cn)
        }
        if ( (nCO_cum < nCO) & (disn == 0) ) {
            nCO_cum = nCO_cum+1;
            g=c(g,gn)
            dis=c(dis,disn)
            pc=rbind(pc,cn)
        }
    }#while


    ### stratification-score matching ###
    rownames(pc)<-1:nrow(pc)
    datag<-cbind(dis,g,pc)

    head(datag)
    datag<-data.frame(datag)

    ### construct the stratification score using the significant eigenvectors ###
    sscore_pc<-glm(dis~pc,family=binomial(),data=datag)
    effect_ConY_est = rbind(effect_ConY_est, summary(sscore_pc)$coef[,1])
    effect_ConY_sd  = rbind(effect_ConY_sd,  summary(sscore_pc)$coef[,2])

    ### calculate the dissimilarity measure based on the stratification score
###
    ssd_pc<-mdist(sscore_pc)

    ### Step 3: perform full matching of cases and controls ###
    fmatch_ssd_pc1<-fullmatch(ssd_pc)
    fmatch_ssd_pc<-as.numeric(fmatch_ssd_pc1)

    ### form a new dataframe combining the disease, test-SNP genotype, and
matched-stratum indicator ###
    full_match_dat_pc<-data.frame(cbind(dis,g,fmatch_ssd_pc))

    ### perform CMH test of SNP-disease association:
    ### must first remove strata with fewer than two observations
    ### (possible if SNP vector g contains missing data)
    orig_table<-table(full_match_dat_pc)
    orig_table_strat<-max(full_match_dat_pc[,3])
    miss_strat<-0
```

```
for(i in 1:orig_table_strat){
      if(sum(orig_table[,,i])<=1)
      miss_strat<-c(miss_strat,i)
}

if(length(miss_strat)>1){
      miss_strat<-miss_strat[-1]
      final_table<-orig_table[,,-c(miss_strat)]
      nstrat<-orig_table_strat-length(miss_strat)
}
else if(length(miss_strat)==1){
      final_table<-orig_table
      nstrat<-orig_table_strat
}

ng<-length(table(full_match_dat_pc[,2]))
gscore<-seq(0,(ng-1))

### logistic regression ###

fit<-glm(dis~g,data=datag,family=binomial())
fit1<-clogit(dis~g+strata(fmatch_ssd_pc),data=full_match_dat_pc)
fit2<-clogit(dis~g+pc+strata(fmatch_ssd_pc),data=full_match_dat_pc)
fit3<-clogit(dis~g+pc[,1]+strata(fmatch_ssd_pc),data=full_match_dat_pc)
fit4<-clogit(dis~g+pc[,2]+strata(fmatch_ssd_pc),data=full_match_dat_pc)

fit.beta[z]<-fit$coefficients[2]
fit1.beta[z]<-fit1$coefficients[1]
fit2.beta[z]<-fit2$coefficients[1]
fit3.beta[z]<-fit3$coefficients[1]
fit4.beta[z]<-fit4$coefficients[1]

fit.se[z]<-sqrt(vcov(fit)[2,2])
fit1.se[z]<-sqrt(fit1[2]$var)
fit2.se[z]<-sqrt(fit2[2]$var[1,1])
fit3.se[z]<-sqrt(fit3[2]$var[1,1])
fit4.se[z]<-sqrt(fit4[2]$var[1,1])


### implement CMH test ###
cmh_analysis<-cmh_test(as.table(final_table),scores=list(dis=0:1,g=gscore))
```

```
    fit.p[z]<-summary(fit)$coefficients[2,4]
    fit1.p[z]<-summary(fit1)$coefficients[1,5]
    fit2.p[z]<-summary(fit2)$coefficients[1,5]
    fit3.p[z]<-summary(fit3)$coefficients[1,5]
    fit4.p[z]<-summary(fit4)$coefficients[1,5]

    ### obtain p-value from CMH test ###
    pvalue_cmh<-pvalue(cmh_analysis)
    sig[z]<-(pvalue_cmh<=0.05)
    sig_value[z]<-pvalue_cmh

}

### Bias, SE and SEE ###
fit.bias<-sum(fit.beta-effect_GonY)/sim
fit1.bias<-sum(fit1.beta-effect_GonY)/sim
fit2.bias<-sum(fit2.beta-effect_GonY)/sim
fit3.bias<-sum(fit3.beta-effect_GonY)/sim
fit4.bias<-sum(fit4.beta-effect_GonY)/sim


fit.se<-sqrt(sum((fit.beta-mean(fit.beta))^2)/sim)
fit1.se<-sqrt(sum((fit1.beta-mean(fit1.beta))^2)/sim)
fit2.se<-sqrt(sum((fit2.beta-mean(fit2.beta))^2)/sim)
fit3.se<-sqrt(sum((fit3.beta-mean(fit3.beta))^2)/sim)
fit4.se<-sqrt(sum((fit4.beta-mean(fit4.beta))^2)/sim)

fit.see<-mean(fit.se)
fit1.see<-mean(fit1.se)
fit2.see<-mean(fit2.se)
fit3.see<-mean(fit3.se)
fit4.see<-mean(fit4.se)
```