Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 114
DOI: 10.11159/icsta22.114

Understanding the Population Structure Correction Regression

The Tien Mai[(1)] , **Pierre Alquier**[(2)]
[(1)] Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway
[(2)] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan.
the.t.mai@ntnu.no

Abstract - Although genome-wide association studies (GWAS) on complex traits have achieved great successes, the current leading GWAS approaches simply perform to test each genotype-phenotype association separately for each genetic variant. Curiously, the statistical property for using these approaches is not known when a joint model for the whole genetic variants is considered. Here we advance in GWAS in understanding the statistical properties of the "population structure correction" (PSC) approach, a standard univariate approach in GWAS. We further propose and analyse a correction to the PSC approach, termed as "corrected population correction" (CPC). Together with the theoretical results, numerical simulations show that CPC is always comparable or better than PSC, with a dramatic improvement in some special cases.

***Keywords*:** GWAS, population structure correction, linear regression, bias, variance.

1. Introduction

In high dimensional data analysis where the number of covariates p is larger than the number of samples n, penalized regression, such as the Lasso, is one of the most popular approaches [1-4]. However, interpreting the results of the Lasso in terms of hypothesis testing or uncertainty quantification is difficult.

Motivated by genome-wide association studies (GWAS), we focus on the following question: given a response vector y of n samples and a matrix of (genetic) covariates $X_{n\times p}$ formed by p (genetic) covariates of n samples, we want to determine which covariates associate with the response. Although the variable selection problem is a classical problem in statistics, in this context it is still a big challenge as the number of covariates is huge compared to the sample size, which prohibits the use of the classical methods. Moreover, in many practical situations, the genomic data are huge and cannot even be examined on a personal laptop. For example, in human genomics, the number of covariates (SNPs, single-nucleotide polymorphism) as well as the number of samples are often at hundreds of thousands [5]; or it can be at order tens of millions of covariates when using k-mers (an alignment-free biomarker type) as in bacterial genomics [6] with thousands of samples.

Besides the computational reason, most of the theoretical result on the Lasso are on l_2 estimation of the parameter and not on variable selection. These two objectives are known to be incompatible in general, see [7, 8] (and [9] illustrated this in the case of the Lasso). Some results for variable selection were derived for thresholded versions of the Lasso, for example in [10], but are valid under strong assumptions that are usually not satisfied in GWAS. Some procedure leading to significance tests and confidence intervals are proposed e.g in [10-12] but they are not easy to handle and costly to compute when using penalized regression such as Lasso.

Because of this, it is common for users to use a univariate model for testing the association of a trait and a covariate (say $X_{\cdot 1}$) to estimate its effect and test its significant. However, the omitted variables have an effect, which we will model by a multivariate linear model. The effect of omitted variables depends strongly on the dependence between $X_{\cdot 1}$ and the other covariances. In GWAS data, the covariates (often SNPs) are in some dependent structures which is called *linkage disequilibrium* [13, 14]. If one uses a univariate regression and ignores the effect of the other covariates, then they can be effectively modelled as part of the error. However, the covariates are correlated due to the population structure, this leads to a correlation between the tested covariate, say $X_{\cdot 1}$, and the noise term together with the noise of the samples. These correlations can cause in inflated type-1 error rates [15]. To handle this problem, the so-called 'population structure correction' approach had been introduced and successfully applied in practice, see for example [13, 14, 16] among others.

In principle, population structure correction is an alternative way to implicitly model the other covariates that are not being tested at the time. This can be done through the latent subspace of these variables. A natural way is to use principal component analysis to extract some features that contain most information of the other covariates $X_{\cdot -1}$ and use these features

as representatives added in the univariate regression of $X_{\cdot 1}$. In this way, it can be seen as a dimension-reduction approach. Another way that is also being the standard approach in GWAS is to use 'linear mixed model' framework in which the covariates that are not directly being tested are treated as random. However, several works had shown that inclusion of $X_{\cdot 1}$ in calculating the principal components can lead to loss in power [17, 18]. This motivates and leads to the popular usage of leave-one-chromosome-out method [16-18].

Although univariate regression approach with population structure correction has become the state-of-the-art approach in GWAS, there are several numerical works have showed that fitting a penalized multivariate regression can exceed it, e.g [19-25]. This can be explained by the bias of population structure correction. Moreover, population structure correction very much depends on the added latent features of untested covariates.

In this paper, we study the statistical properties of the population structure correction when assuming the true underlying model is a multivariate linear regression. More specifically, we derive explicitly the bias and the variance of the population structure correction method. Moreover, we also propose and study a simple version of the leave-one-chromosome-out method, termed as 'corrected population correction'. We show theoretically and empirically that 'corrected population correction' approach reduces the variance when compared to the population structure method.

2. Model And Methods

2.1. Model

Given a response vector y of n samples and p covariates X_j with $n << p$, we assume that the response vector relates to the covariates by the following linear model

$$Y_i = \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i, i = 1, \ldots, n \tag{1}$$

or, summarized by $Y = X\beta + \varepsilon$ in matrix form. We assume that $\varepsilon_i \sim N(0,\sigma^2)$.

The problem is to estimate the coefficient β_j of a specified covariate X_j. Up to a re-ordering of the variables, say that the coefficient is β_1. For simplicity, there is no intercept: we assume that the variables are already centered, and normalized.

Notations: For any matrix A, $A_{(-j)}$ denotes the matrix A without its j-th column, and $A_{s:j}$ is the submatrix with only columns $s,s+1,\ldots,j$. We use the same convention for column vectors: $\beta_{s:j}$ means that we extract entries from s to j.

2.2. Population Structure Correction (PSC)

The idea is to perform a principal component analysis (PCA) on X and then use some principal components corresponding to the top leading eigenvalues. In other words, $X^TX = \bar{W}^T diag(\overline{\lambda_1}, \ldots, \overline{\lambda_p})\bar{W}$ and $\overline{\lambda_1} \geq \cdots \geq \overline{\lambda_p} \geq 0$. The matrix X^TX is also known as the 'kinship' matrix in genomic research. Here it will be more convenience to think of PCA as an SVD, that is $X = \bar{U} diag(\overline{\sigma_1}, \ldots, \overline{\sigma_r})\bar{V}^T$ and $\overline{\sigma_1} \geq \cdots \geq \overline{\sigma_r} \geq 0$ where $\bar{r} = \text{rank}(X)$ and $\bar{\sigma}_1 > \cdots > \bar{\sigma}_r > 0$. It is easy to see that $\bar{V} = \bar{W}$ as soon as rank(X) = n.

Then the idea is simply to estimate the model for $Y_i = \bar{\alpha} X_{1,i} + \sum_{j=1}^{k} \bar{\gamma}_j \bar{U}_{i,j} + \bar{e}_i^{(k)}$, some k, or

$$Y = X_{\cdot 1}\bar{\alpha} + \bar{U}_{(1:k)}\bar{\gamma} + \bar{e}^{(k)}, \tag{2}$$

hoping that $\bar{\alpha}$ is a good proxy for β_1 in model (1).

2.3. Corrected Population Correction (CPC)

We propose a modified procedure: first, perform a PCA on $X_{(-1)}$ (X without its first column $X_{\cdot 1}$ is denoted by $X_{(-1)}$), in other words: $X_{(-1)} = U\Sigma V^\top$ where $r = \text{rank}(X_{(-1)})$ and $\sigma_1 > \cdots > \sigma_r > 0$. Similar to PSC, it is simply to estimate the model, for some k, $Y_i = \alpha X_{1,i} + \sum_{j=1}^{k} \gamma_j U_{i,j} + e_i^{(k)}$, or

$$Y = X_{\cdot 1}\alpha + U_{(1:k)}\gamma + e^{(k)}, \tag{3}$$

and hoping that α is a good proxy for β_1 in model (1).

We would like to note that this method is a simplified version of the so-called leave-one-chromosome-out method, that is popular in GWAS [18].

2.4. Why does PSC need to be corrected?

In general, the CPC model is not "correct" in the sense that $E(Y) \neq X_{\cdot 1}\alpha + U_{1:k}\gamma$. In other words, in (3) we don't have $E(e^{(k)}) = 0$, in general. However, assume that k = rank($X_{\cdot 1}$), then we have $Y = X_{\cdot 1}\alpha + U\gamma + e$ and note that $X_{(-1)} = U\Sigma V^T$ leads to $X_{(-1)}V(V^\top V)^{-1}\Sigma^{-1} = U$. Thus, this model is equivalent to (1) with $\alpha = \beta_1$, by identification:

$$X_{\cdot 1}\alpha + X_{(-1)}(V(V^\top V)^{-1}\Sigma^{-1}\gamma) = X\beta$$

(and so ε = e in this case). Therefore, for a well-chosen k, the model is actually exact.

For this reason, we can reformulate the problem as: with the true model

$$Y = X_{\cdot 1}\alpha + U\gamma + \varepsilon, \tag{4}$$

$\gamma \in \mathbb{R}^r$ where r = rank($X_{(-1)}$), what is the effect on α to estimate instead, for some k,

$$Y = X_{\cdot 1}\alpha + U_{(1:k)}\gamma_{1:k} + e_{(k)} \tag{5}$$

where we actually have $e^{(k)} = U_{(k+1):r}\gamma_{(k+1):r} + \varepsilon$. This is simply a problem of omission of variables: what is the effect of the omission of $U_{(k+1):r}$?

On the other hand, for k = rank(X), we have $Y = X\beta + \varepsilon = \bar{U}\bar{\Sigma}\bar{V}^\top\beta + \varepsilon = \bar{U}(\bar{\Sigma}\bar{V}^\top\beta) + \varepsilon = \bar{U}\bar{\gamma} + \varepsilon$ and so the PSC model

$$Y = \bar{\alpha}X_1 + \bar{U}\bar{\gamma} + \varepsilon \tag{6}$$

is simply not identifiable (the variable $X_{\cdot 1}$ is twice in the model). When $k <$ rank(X), the model might be identifiable, but the fact that $X_{\cdot 1}$ is in the first term, and "partly" in the second, will lead to a greater bias than in CPC. This drawback has been figured out in the field of genetic research [18]. For a formal statement see the analysis below.

3. Statistical Analysis

In the following we explicitly derive the bias and the variance for each considered method above. These results bring insights on understanding how the population structure correction is working practically. All technical proofs are given in the online Supplement at [27].

We first provide some statistical properties for the CPC method.

Theorem 1. Assume that model (1) or equivalently (4) holds. Then with CPC method we have

$$\text{bias}(\hat{\alpha}) = \frac{X_{\cdot 1}^\top U_{(k+1):r}\gamma_{(k+1):r}}{X_{\cdot 1}^\top X_{\cdot 1} - \left\|X_{\cdot 1}^\top U_{1:k}\right\|^2}, \quad \text{Var}(\hat{\alpha}) = \frac{\sigma^2}{X_{\cdot 1}^\top X_{\cdot 1} - \left\|X_{\cdot 1}^\top U_{1:k}\right\|^2}. \tag{7}$$

It can be seen that $X_{\cdot 1}^\top U_{(k+1):r}\gamma_{(k+1):r}$ measures the correlation between $X_{\cdot 1}$ and the other part of X (as in the true model (4)) which was not included in the wrong model (5). Obviously, if the wrong model is actually not "too wrong" in the sense that $||U_{(k+1):r}\gamma_{(k+1):r}|| \approx 0$ then the bias would be small. But when this is not the case, the term is problematic only if $X_{\cdot 1}$ is correlated with this quantity.

The $X_{\cdot 1}^\top X_{\cdot 1} - \left\|X_{\cdot 1}^\top U_{1:k}\right\|^2$ denominator is just an identifiability term: if $X_{\cdot 1}$ is *too* correlated with the other variables used in model (5), then the variance of $\hat{\alpha}$ will increase (as usual) but also the bias due to misspecification.

Remark 1. *As γ is unknown in practice, $c_j := X_{\cdot 1}^\top U_{(j)}$ but the are observed. So we can give a result under an assumption that depends only on γ. For example if we assume that $\|\gamma\|_1 \leq B$ (as in the Lasso) then*

$$|\text{bias}(\hat{\alpha})| \le \frac{\|\gamma\|_1 \sum_{j=k+1}^{r} |c_j|}{X_{\cdot 1}^\top X_{\cdot 1} - \sum_{j=1}^{k} c_j^2} \le \frac{B \sum_{j=k+1}^{r} |c_j|}{X_{\cdot 1}^\top X_{\cdot 1} - \sum_{j=1}^{k} c_j^2}.$$

Statistical properties of the PSC method are given in the following theorem.

Theorem 2. *Assume that model* (1) *holds. For the model* (6)*, with PSC method, we have*

$$\text{bias}(\hat{\hat{\alpha}}) = \frac{X_{\cdot 1}^\top (\bar{U}_{(k+1):\bar{r}} \bar{\gamma}_{(k+1):\bar{r}} - X_{\cdot 1}\beta_1) - \beta_1 \left\| X_{\cdot 1}^\top \bar{U}_{1:k} \right\|^2}{X_{\cdot 1}^\top X_{\cdot 1} - \left\| X_{\cdot 1}^\top \bar{U}_{1:k} \right\|^2}; \quad \text{Var}(\hat{\hat{\alpha}}) = \frac{\sigma^2}{X_{\cdot 1}^\top X_{\cdot 1} - \left\| X_{\cdot 1}^\top \bar{U}_{1:k} \right\|^2}. \tag{8}$$

In the following theorem, we derive the relationship between the variances of these two methods.

Theorem 3. *The corrected population correction method reduces the variance of the original population structure correction.*

Remark 2. *From Theorem 3, it states that the corrected population correction (CPC) always returns estimate with smaller variance compared to the structured population correction (SPC). From simulations, we conjecture that the biasness of CPC method is also smaller than those from SPC method, however this is not easy to show from our analysis.*

4. Numerical Simulations

In this section, we investigate basic properties of the PSC and CPC methods studied above. We fix p = 100, n = 1000 for low dimension setting and p = 1000, n = 600 for high dimension setting. The noise variance is fixed at $\sigma 2 = 1$. We generate the parameter $\beta \in R^p$ such that its first component $\beta 1$ is fixed to 1, and other non-zero components were sampled uniformly at random from $\{\pm 1\}$. The sparsity of β will be changed in each setting corresponding to $\|\beta\|0 = 20, 100$. The response Y is simulated as in linear model (1).

For each setting, we simulated 100 independent datasets and report the average results together with their standard deviations. The number of principal components *k* added in models (3) and (2) are varied from 1 to 30.

Example: worst case scenario for PSC

Here we show cases that PSC does not work well while CPC performs superior results. We consider the structured X such that its first two columns $X_{\cdot 1}$ and $X_{\cdot 2}$ are corresponding to its first two leading principal components. A brief summary of the data can be found in the Figure 1.

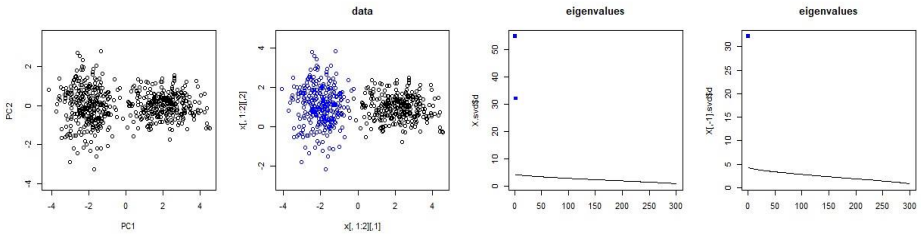

Figure 1: *Summary the structured X: When $X_{\cdot -1}$ is removed, a principal component is also removed.*

In this case, it is clear to see that PSC can actually be very biased whereas CPC is very stable and accurate, see Figure 2. This example demonstrates that including X_1 (the covariate being tested) in the calculation of the principal components can be very harmful.

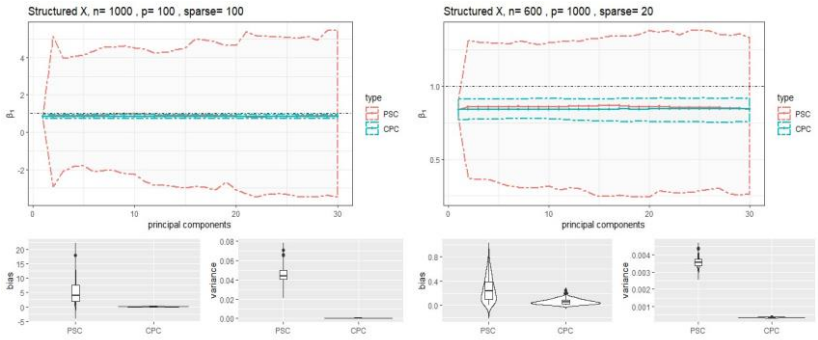

Figure 2: *Structured X. Estimates of β_1 with different number of Principal components (PCs)*

Behaviour in other cases

We further consider the following settings for the design matrix:

- Independent X: In this setting, $X_{ij} \sim N(0,1)$.
- Dependent X: We consider $X_{i\cdot} \sim N(0,\Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$.
- Binary X: the X_{ij} is simulated from the set $\{\pm 1\}$ with equal probability.

Results from simulations, Figures 4 and 5 confirm our theoretical results above. In general, the CPC method performs similarly to PSC method. However, CPC return the results with less variation than the PSC method. Moreover, the PSC method is very much depending on k, the number of principal components added in the model.

Real data assessment in a wheat GWAS data

We apply two methods to a real wheat GWAS data which is available in the R package 'BGLR' [25]. The data consists of 599 wheat lines: lines (responses) were evaluated for grain yield and each line has been genotyped with 1279 markers. We run CPC and PSC across 1279 covariates and report the absolute errors $|\hat{\beta}_j^{CPS} - \hat{\beta}_j^{PSC}|$ and the relative errors

$$\frac{|\hat{\beta}_j^{CPC} - \hat{\beta}_j^{PSC}|}{|\hat{\beta}_j^{PSC}|}.$$

These results are given in Figure 3. Regarding the histogram in Figure 3, the conclusion is clear: for most coefficients, PSC and CPC lead to similar estimation, but for some of them, the deviation is extremely high. There are in total 55 covariates such that their relative errors are greater than 0.5 (and there are 33 covariates such that their relative errors are greater than 1). Therefore, including the covariate being tested in the calculation of the principal components could create a huge difference.

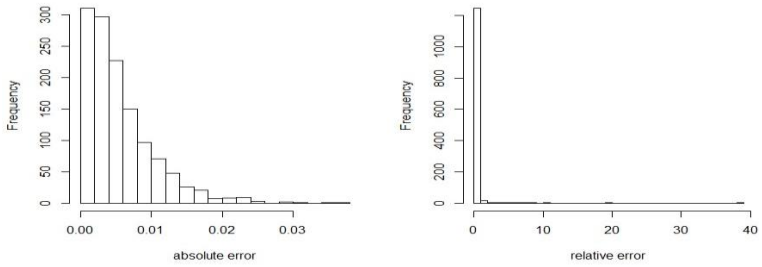

Figure 3: *Histogram of the absolute errors and relative errors of 1279 covariates in wheat data with 10 principal components.*

5. Conclusion

In this paper, we have discussed the statistical properties of the widely used method, structure population correction method, in genome-wide association studies. We have also proposed and studied a simple version of the 'leave-one-chromosome-out' in GWAS, termed as Corrected population correction method. Our theoretical analysis and simulations show that the structure population correction method (although efficient computationally) should be used with more careful as it comes with higher variance due to model-misspecification. The corrected population correction method, which requires higher computational cost, returns better results as it avoids model-misspecification.

Acknowledgments

T.T.M would like to thank Jukka Corander and John A Lees for useful discussion on GWAS. The research of T.T.M was supported by the Norwegian Research Council grant number 309960 through the Centre for Geophysical Forecasting at NTNU. The Rcodes and data used in the numerical experiments are available at: https://github.com/tienmt/understand_SPC . A longer version of this work, with the proofs, can be found in [27].

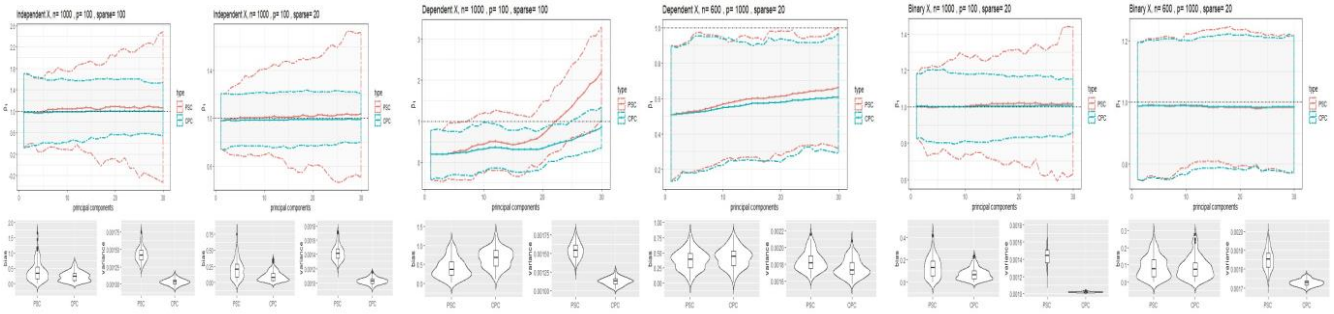

Figure 4: *Independent X (left) and Dependent X (middle) and Binary X (right). Estimates of β_1 with different number of Principal components (PCs)*

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[2] P. Buhrmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[3] C. Giraud, *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

[4] B. Efron and T. Hastie, *Computer age statistical inference*, vol. 5. Cambridge University Press, 2016.

[5] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell and A. Cortes, "The UK biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, p. 203, 2018.

[6] J. A. Lees, M. Vehkala, N. Valimaki, S. R. Harris, C. Chewapreecha, N. J. Croucher, P. Marttinen, M. R. Davies, A. C. Steer, S. Y. Tong and A. Honkela, "Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes," *Nature communications*, vol. 7, p. 12797, 2016.

[7] Y. Yang, "Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation," *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[8] H. Leeb and B. M. Potsherd, "Sparse estimators and the oracle property, or the return of hodges' estimator," *Journal of Econometrics*, vol. 142, no. 1, pp. 201–211, 2008.

[9] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[10] K. Lounici, "Sup-norm convergence rate and sign concentration property of lasso and Dantzig estimators," *Electronic Journal of statistics*, vol. 2, pp. 90–102, 2008.

[11] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high dimensional regression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869– 2909, 2014.

[12] S. Van de Geer, P. Buhlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, vol. 42, no. 3, pp. 1166–1202, 2014.

[13] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 217–242, 2014.

[14] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, p. 904, 2006.

[15] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7, p. 459, 2010.

[16] E. Derks, A. Zwinderman, and E. Gamazon, "The relation between inflation in type-i and type-ii error rate and population divergence in genome-wide association analysis of multiethnic populations," *Behavior genetics*, vol. 47, no. 3, pp. 360–368, 2017.

[17] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies," *Nature methods*, vol. 8, no. 10, p. 833, 2011.

[18] J. Listgarten, C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman, "Improved linear mixed models for genome-wide association studies," *Nature methods*, vol. 9, no. 6, p. 525, 2012.

[19] J. Yang, N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, "Advantages and pitfalls in the application of mixed-model association methods," *Nature genetics*, vol. 46, no. 2, p. 100, 2014.

[20] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[21] T. T. Mai, P. Turner, and J. Corander, "Boosting heritability: estimating the genetic component of phenotypic variation with multiple sample splitting," *BMC bioinformatics*, vol. 22, no. 164, pp. 1–16, 2021.

[22] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of gwas discovery: biology, function, and translation," *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.

[23] L. Buzdugan, M. Kalisch, A. Navarro, D. Schunk, E. Fehr, and P. Buhlmann, "Assessing statistical significance in multivariable genome wide association analysis," *Bioinformatics*, vol. 32, no. 13, pp. 1990–2000, 2016.

[24] D. Brzyski, C. B. Peterson, P. Sobczyk, E. J. Candes, M. Bogdan, and C. Sabatti, "Controlling the rate of gwas false discoveries," *Genetics*, vol. 205, no. 1, pp. 61–75, 2017.

[25] J. A. Lees, T. T. Mai, M. Galardini, N. E. Wheeler, S. T. Horsfield, J. Parkhill, and J. Corander, “Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions,” *MBio*, vol. 11, no. 4, pp. e01344–20, 2020.
[26] P. Perez and G. de Los Campos, “Genome-wide regression and prediction with the bglr statistical package,” *Genetics*, vol. 198, no. 2, pp. 483–495, 2014.
[27] Mai, T. T., & Alquier, P. . Understanding the population structure correction regression. *arXiv preprint arXiv:2108.05655* (2021).