Testing the effectiveness of principal components in adjusting for relatedness in genetic association studies

Yiqi Yao[1], Alejandro Ochoa[1,2,*]

[1] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

[2] Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC 27705, USA

[*] Corresponding author: alejandro.ochoa@duke.edu

Abstract

Modern genetic association studies require modeling population structure and family relatedness in order to calculate correct statistics. Principal Components Analysis (PCA) is one of the most common approaches for modeling this population structure, but nowadays the Linear Mixed-Effects Model (LMM) is believed by many to be a superior model. Remarkably, previous comparisons have been limited by testing PCA without varying the number of principal components (PCs), by simulating unrealistically simple population structures, and by not always measuring both type-I error control and predictive power. In this work, we thoroughly evaluate both PCA and LMMs with varying number of PCs in various realistic scenarios, including admixture together with family structure, measuring both null p-value uniformity and the area under the precision-recall curves. We find that LMM without PCs performs best overall, although PCA performs as well as LMM when enough PCs are used and there are no close relatives. We also find a remarkable robustness to extreme numbers of PCs. Altogether, we recommend LMMs in general for association studies, but find that PCA performs as well as LMMs in many common scenarios.

1 Introduction

The goal of a genetic association study is to identify loci whose genotypes are correlated significantly with a certain trait. An important assumption made by classical association tests is that genotypes are unstructured: drawn independently from a common allele frequency. However, this assumption does not hold for structured populations, which includes multiethnic cohorts and admixed individuals, and for family data. When naive approaches are incorrectly applied to structured populations and/or family data, association statistics (such as χ^2) become inflated relative to the null expectation, resulting in greater numbers of false positives than expected and loss of power (Devlin and Roeder, 1999; Voight and Pritchard, 2005; Astle and Balding, 2009).

The most popular approaches for conducting genetic association studies with structured populations involve modeling the population structure via covariates. Such covariates may be inferred ancestry proportions (Pritchard et al., 2000b) or transformations of these. Principal components analysis (PCA) represents the most common of these variants, in which the top eigenvectors of the kinship matrix are used to model the population structure (Zhang et al., 2003; Price et al., 2006; Bouaziz et al., 2011). These top eigenvectors are commonly referred to as Principal Components (PCs) in the genetics literature (the convention we adopt here; Patterson et al., 2006), but it is worth noting that in other fields the PCs would instead denote the projections of the data onto the eigenvectors (Jolliffe, 2002). Various works have found that PCs map to ancestry (*e.g.*, Zhou et al., 2016), and PCs work as well as ancestry in association studies and can be inferred more quickly (Patterson et al., 2006; Zhao et al., 2007; Bouaziz et al., 2011). More recent work has focused on speeding up the calculation of PCs rather than on evaluating its performance in association studies (Lee et al., 2012; Abraham and Inouye, 2014; Galinsky et al., 2016; Abraham et al., 2017). PCA remains a popular and powerful approach for association studies (Wojcik et al., 2019).

The other dominant approach for genetic association studies under population structure is the Linear Mixed-effect Model (LMM), in which population structure is a random effect drawn from a covariance model parametrized by the kinship matrix. LMM and PCA share deep connections that suggest that both models ought to perform similarly (Astle and Balding, 2009; Janss et al., 2012; Hoffman, 2013). However, many previous studies have found that LMM outperforms PCA,

although many evaluations have been limited (Zhao et al., 2007; Astle and Balding, 2009; Kang et al., 2010; Song et al., 2015). Other studies find that PCA can outperform LMM in certain settings (Price et al., 2010; Wu et al., 2011; Wang et al., 2013), although these are believed to be unusual (Sul and Eskin, 2013). Moreover, various explanations for if and why LMM outperforms PCA are vague and have not been tested directly (Price et al., 2010; Sul and Eskin, 2013; Price et al., 2013; Hoffman, 2013). Since LMMs tend to be considerably slower than the PCA approach, it is important to understand when the difference in performance between these two approaches is outweighed by their difference in runtime.

PCA has been evaluated in numerous previous works in the context of association studies. However, all of these studies have important limitations, for the most part due to PCA being treated as a competitor rather than a method worthy of exploring more fully. For example, although there are methods for selecting the numbers of PCs (Patterson et al., 2006), most evaluations either admit to selecting 10 because it has long been the default and it performs well enough, regardless of the dataset in question (Epstein et al., 2007; Li and Yu, 2008; Astle and Balding, 2009; Li et al., 2010; Wu et al., 2011), or otherwise test only one number of PCs without justification (Zhang et al., 2003; Kimmel et al., 2007; Zhao et al., 2007; Zhang et al., 2008; Price et al., 2010; Bouaziz et al., 2011; Hoffman, 2013; Wang et al., 2013; Tucker et al., 2014; Yang et al., 2014; Song et al., 2015; Sul et al., 2018). Conversely, only a few studies consider a (small) set of numbers of PCs, where they show remarkable robustness to this choice (Price et al., 2006; Kang et al., 2010; Wojcik et al., 2019). Moreover, most of these evaluations considered simulated data with only $K = 2$ independent subpopulations or admixture from only two subpopulations (exceptions are Astle and Balding (2009) with $K = 3$, and Wang et al. (2013) with $K = 4$), although worldwide human population structure is expected to have a larger dimensionality of at least $K = 9$ (Wojcik et al., 2019). Similarly, only one evaluation simulated data from a family pedigree: Price et al. (2010) included sibling pairs. Some studies include evaluations involving real data that featured known or cryptic relatedness, but these analyses did not measure type-I error rates or power calculations, most of which settled for measuring test statistic inflation. Lastly, many of the earlier evaluations employed case-control simulations exclusively (as opposed to quantitative traits as we do here), were

based on very small real or simulated datasets relative to today's standards, did not include any LMMs in their evaluations, and often did not measure both type-I error rates and power (or one of their proxies). Here we aim to systematically evaluate the robustness of both the PCA (fixed effects only) and LMM (mixed effects) approaches to the choice of number of PCs, especially in cases where the model is grossly misspecified, in more realistic simulations relevant to today's genetics research.

In this work, we study the performance of the PCA and LMM methods in genetic association studies, characterizing their behavior under various numbers of PCs (for LMM as well as PCA) and varying sample sizes, under a reasonable admixture model with $K = 10$ source subpopulations and also a model with admixture and family structure. Our evaluation is more thorough than previous ones, directly measuring the uniformity of null p-values (as required for accurate type-I error control and FDR control via q-values; Storey, 2003; Storey and Tibshirani, 2003) and predictive power by calculating the area under precision-recall curves. Across all tests we find that LMM without PCs performs best, but PCA matches that optimal performance when enough PCs are used and there are no close relatives in the study. Remarkably, PCA is robust even when the number of PCs far exceeds the optimal number for reasonably large studies. However, for smaller studies (100 individuals) there is a more pronounced loss of power for PCA when the number of PCs exceeds the optimal number. LMMs significantly outperforms PCA in the presence of family structure, which is a well-known case where PCA fails (Patterson et al., 2006; Price et al., 2010). We also found that LMM without PCs always performs as well or better than LMMs that use any number of PCs. All together, our simulation studies indicate that LMMs are preferable, and provide clear criteria under which use of PCA results in acceptable performance compared to LMMs.

2 Models and Methods

2.1 Models for genetic association studies

In this subsection we describe the complex trait model and kinship model that motivates both the PCA and LMM models for genetic association studies, followed by further details regarding the PCA and LMM approaches. The derivations of the PCA and LMM models from the general

quantitative trait model are similar to previous presentations (Astle and Balding, 2009; Janss et al., 2012; Hoffman, 2013), but we emphasize the kinship model for random genotypes as being crucial for these connections, and make a clear distinction between the true kinship matrix and its most common estimator, which is biased (Ochoa and Storey, 2016b; Ochoa and Storey, 2018).

2.1.1 The complex trait model and PCA approximation

Let $x_{ij} \in \{0, 1, 2\}$ be the genotype at locus i for individual j, which counts the number of reference alleles. Suppose there are n individuals and m loci, $\mathbf{X} = (x_{ij})$ is their $m \times n$ genotype matrix, and $\mathbf{y}$ is the length-n (column) vector which represents trait value for each individual. The approaches we consider are based on the following additive linear model for a quantitative (continuous) trait:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}^\intercal\beta + \epsilon, \tag{1}$$

where $\mathbf{1}$ is a length-n vector of ones, α is the scalar intercept coefficient, β is the length-m vector of locus effect sizes, and ϵ is a length-n vector of residuals. The residuals are assumed to follow a normal distribution: $\epsilon_j \sim \text{Normal}(0, \sigma^2)$ independently for each individual j, for some residual variance parameter σ^2.

Typically the number of loci m is in the order of millions while the number of individuals n is in the thousands. Hence, the full model above cannot be fit in this typical $n \ll m$ case, as there are only n datapoints to fit (the trait vector) but there are $m + 1$ parameters to fit (α and the β vector). The PCA model with r PCs approximates the full model fit at a single locus i:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_r\gamma_r + \epsilon, \tag{2}$$

where $\mathbf{x}_i$ is the length-n vector of genotypes at locus i only, β_i is the effect size coefficient for that locus, $\mathbf{U}_r$ is an $n \times r$ matrix of PCs, and γ_r is the length-r vector of coefficients for the PCs. This approximation is explained by first noticing that the genotype matrix has the following singular value decomposition: $\mathbf{X}^\intercal = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$, where assuming $n < m$ we have that $\mathbf{U}$ is an $n \times n$ matrix of the left singular vectors of $\mathbf{X}$, $\mathbf{V}$ is an $m \times n$ matrix of its right singular vectors, and $\mathbf{D}$ is an $n \times n$

diagonal matrix of its singular values. Thus, in the full model we have $\mathbf{X}^\intercal\beta = \mathbf{U}\gamma$, where $\gamma = \mathbf{D}\mathbf{V}^\intercal\beta$ is a length-n vector. The approximation consists solely of replacing $\mathbf{U}\gamma$ (the full set of n left singular vectors and their coefficients) with $\mathbf{U}_r\gamma_r$ (the top r singular vectors only, which constitutes the best approximation of rank r). Thus, the extra terms in the PCA approach approximate the polygenic effect of the whole genome, and assumes that the locus i being tested does not contribute greatly to this signal.

The statistical significance of a given association test is performed as follows. The null hypothesis is $\beta_j = 0$ (no association). The null and alternative models are each fit (fitting the coefficients of the multiple regression, where β_j is excluded under the null while it is fit under the alternative). The resulting regression residuals are compared to each other using the t-test, which yields a two-sided p-value. Note that many common PCA implementations trade the more precise t-test for a χ^2 test, which is only asymptotically accurate (it requires the overall degrees of freedom of the model to be much smaller than the number of individuals).

Genetic association studies are a multiple hypothesis testing problem, since there are a large number of loci (m) tested for association. In human genetic studies it has long been customary to set a p-value threshold of 5×10^{-8} [TODO: citation], which may be thought of as controlling the total number of false discoveries, or equivalently, the family-wise error rate [citations]. The recommended solution to this problem in other fields is to control the FDR rather than setting a fixed p-value threshold. We recommend estimating q-values and setting a threshold of $q < 0.05$ so that the FDR is controlled at the 5% level (Storey, 2003; Storey and Tibshirani, 2003). However, in this present work we do not calculate q-values; instead we focus on testing for the correctness of null p-value distributions (a prerequisite for acccurate q-value estimation) and on predictive power (for which ranking of loci by p-values suffice; q-values do not alter these rankings).

2.1.2 Kinship model for genotypes

In order to better motivate the most common estimation procedure of PCs for genotype data, and to connect PCA to LMMs, we shall review the kinship model for genotypes. The model states that

genotypes are random variables with a mean and covariance structure given by

$$\mathrm{E}[x_{ij}] = 2p_i, \qquad \mathrm{Cov}(x_{ij}, x_{ik}) = 4p_i(1-p_i)\varphi_{jk},$$

where p_i is the ancestral allele frequency at locus i and φ_{jk} is the kinship coefficient between individuals j and k (Malécot, 1948; Wright, 1951; Jacquard, 1970). Thus, if we standardize the genotype matrix as

$$\mathbf{X}_S = \left(\frac{x_{ij} - 2p_i}{\sqrt{4p_i\,(1-p_i)}} \right),$$

then this results in a straightforward kinship matrix estimator:

$$\mathrm{E}\left[\frac{1}{m} \mathbf{X}_S^\intercal \mathbf{X}_S \right] = \mathbf{\Phi},$$

where $\mathbf{\Phi} = (\varphi_{jk})$ is the $n \times n$ kinship matrix. Note that replacing the raw genotype matrix $\mathbf{X}$ with the standardized matrix $\mathbf{X}_S$ in the trait model of Eq. (1) results in an equivalent model, as this covariate differs only by a linear transformation. Thus, under the standardized genotype model, the PCs of interest are equal in expectation to the top eigenvectors of the kinship matrix.

2.1.3 Estimation of principal components from genotype data

In practice, the matrix of principal components $\mathbf{U}_r$ in Eq. (2) is determined from an estimate of the earlier standardized genotype matrix $\mathbf{X}_S$, namely

$$\hat{\mathbf{X}}_S = \left(\frac{x_{ij} - 2\hat{p}_i}{\sqrt{4\hat{p}_i\,(1-\hat{p}_i)}} \right),$$

where the true ancestral allele frequency p_i is replaced by the estimate $\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}$, and results in the kinship estimate $\hat{\mathbf{\Phi}} = \frac{1}{m}\hat{\mathbf{X}}_S^\intercal \hat{\mathbf{X}}_S$. This kinship estimate and minor variants are also employed in LMMs (Yang et al., 2011). This estimator of the kinship matrix is biased, and this bias is different for every individual pair (Ochoa and Storey, 2016b; Ochoa and Storey, 2018). However, in the present context of PCA regression in genetic association studies, the existing approach performs as well as when the above estimate is replaced by the true kinship matrix (not shown). Thus, it

appears that in combination with the intercept term ($\mathbf{1}\alpha$ in Eq. (2)), the rowspace of this kinship matrix estimate approximately equals that of the true kinship matrix.

2.1.4 Connection between PCs and ancestry proportions

Genetic association using ancestry proportions can be shown to be equivalent to using PCs, as follows. Here we shall assume the individual-specific admixture model commonly assumed when inferring ancestry proportions (Pritchard et al., 2000a; Falush et al., 2003; Alexander et al., 2009; Gopalan et al., 2016; Cabreros and Storey, 2019). There are K ancestral subpopulations and every individual j draws a proportion q_{ju} of its alleles from subpopulation S_u. These ancestry proportions must be non-negative and sum to one for every individual j ($\sum_{u=1}^{K} q_{ju} = 1$ for every j). Each ancestral subpopulation S_u has an allele frequency $p_i^{S_u}$ at locus i, and thus the individual-specific allele frequency π_{ij} of individual j at locus i is given by the weighted average of the ancestral subpopulation allele frequencies, where the ancestry proportions are the weights:

$$\pi_{ij} = \sum_{u=1}^{K} q_{ju} p_i^{S_u}. \tag{3}$$

Genotypes are constructed by drawing each allele independently from this frequency, or $x_{ij} \sim \text{Binomial}(2, \pi_{ij})$. Thus, the rowspace of the genotype matrix is, in expectation, the same as the rowspace of the individual-specific allele frequency matrix, which by Eq. (3) above is the same as the rowspace of the $n \times K$ admixture proportions matrix $\mathbf{Q} = (q_{ju})$. Therefore, the top K principal components suffice to fully model the rowspace of the genotypes, which only have dimension K. As an intercept term is always included in genetic association studies ($\mathbf{1}\alpha$ in Eq. (2)), and the sum of rows of $\mathbf{Q}$ sums to one, then the rowspace of the combined model has dimension K as well, so only $K - 1$ PCs (plus intercept) are needed to span the rowspace of this admixture model.

2.1.5 Linear mixed-effects model

The LMM is another approximation to the complex trait model in Eq. (1). Excluding additional covariates first, the LMM is

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \epsilon, \tag{4}$$

which is like the PCA model in Eq. (2) except that the PC terms $\mathbf{U}_r\gamma_r$ are replaced by the random effect $\mathbf{s}$, which is a length-n vector drawn from

$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma_s^2\mathbf{\Phi}\right),$$

where $\mathbf{\Phi}$ is the kinship matrix and σ_s^2 is a trait-specific variance scaling factor. This model is derived from treating the standardized genotype matrix $\mathbf{X}_S$ as random rather than fixed, so that the standardized genetic effect $\mathbf{X}_S^\intercal\beta_S$ in Eq. (1) has mean zero and a covariance matrix of

$$\text{Cov}\left(\mathbf{X}_S^\intercal\beta_S\right) = ||\beta_S||^2\mathbf{\Phi}.$$

The above random effect $\mathbf{s}$ satisfies those equations, where the variance scale equals $\sigma_s^2 = ||\beta_S||^2$. Thus, the PCA approach is the fixed model equivalent of the LMM under the additional approximation that only the top r eigenvectors are used in PCA whereas the LMM uses all eigenvectors.

A key advantage of LMM over PCA is that it has fewer parameters to fit: ignoring the shared terms in Eq. (2) and Eq. (4), PCA has r parameters to fit (each PC coefficient in the γ vector), whereas LMMs only fit one additional parameter, namely σ_s^2. Therefore, PCA is expected to overfit more substantially than LMM—and thus lose power—when r is very large, and especially when the sample size (the number of individuals n) is very small. Statistical significance in LMMs is calculated via a likelihood ratio test, whose test statistic has a asymptotic χ^2 distribution under the null hypothesis.

2.1.6 LMM with PCs

An LMM variant we focus on testing in this work incorporates PCs as fixed covariates. Since PCs are the top eigenvectors of the same kinship matrix estimate used to draw the random effects, then the population structure is essentially modeled twice, which can lead to loss of power when the number of PCs is very large. However, some previous work has found the apparent redundancy of an LMM with PCs beneficial (Zhao et al., 2007; Price et al., 2010). It is worth noting that earlier LMM approaches estimated kinship matrices from pedigrees (thus excluding population structure),

and population structure was modeled via admixture proportions rather than PCA (Yu et al., 2006; Zhao et al., 2007), so in those cases there was no modeling redundancy.

2.2 Simulations

2.2.1 Genotype simulation from the admixture model

We consider three simulation scenarios, refered to as (1) large sample size, (2) small sample size, and (3) family structure. All cases are based on the admixture model described previously (Ochoa and Storey, 2016a; Ochoa and Storey, 2016b), and which is implemented in the R package `bnpsd` available on GitHub and the Comprehensive R Archive Network (CRAN).

Here we consider scenarios where the number of individuals n varies: the large sample size and family structure scenarios have $n = 1,000$ whereas small sample size has $n = 100$. The number of loci in all cases is $m = 100,000$. Individuals are admixed from $K = 10$ intermediate subpopulations, where K is also the rank of the population structure; thus, after taking into account the intercept's rank-1 contribution, the population structure can be fit with $r = K-1$ PCs. Each subpopulation S_u ($u \in \{1, ..., K\}$) has an inbreeding coefficient $f_{S_u} = u\tau$, individual-specific admixture proportions q_{ju} for individual j and intermediate subpopulation S_u arise from a random walk model for the intermediate subpopulations on a 1-dimensional geography with spread σ, where the free parameters τ and σ are fit to result in $F_{\mathrm{ST}} = 0.1$ for the admixed individuals and a bias coefficient of $s = 0.5$, exactly as before (Ochoa and Storey, 2016b).

Random genotypes are drawn from this model, as follows. First, uniform ancestral allele frequencies p_i are drawn. The allele frequency $p_i^{S_u}$ at locus i of each intermediate subpopulation S_u is drawn from the Beta distribution with mean p_i and variance $p_i(1 - p_i)f_{S_u}$ (Balding and Nichols, 1995). The individual-specific allele frequency of individual j and locus i is given by $\pi_{ij} = \sum_{u=1}^{K} q_{ju} p_i^{S_u}$. Lastly, genotypes are drawn from $x_{ij} \sim \mathrm{Binomial}(2, \pi_{ij})$. Loci that are fixed (where for some i we had $x_{ij} = 0$ for all j, or $x_{ij} = 2$ for all j) are drawn again from the model, starting from p_i, iterating until no loci are fixed.

2.2.2 Genotype simulation from the family model

Here we describe a simulation of a family structure with admixture that aims to be realistic by: (1) pairing all individuals in every generation, resulting in two children per couple; (2) strictly avoiding close relatives when pairing individuals; (3) strongly favoring pairs that are nearby in their 1-dimensional geography, which helps preserve the population structure across the generations by preferentially pairing individuals with more similar admixture proportions (a form of assortative mating); and (4) iterating for many generations so that a broad distribution of close and distant relatives is present in the data.

Generation 1 has individuals with genotypes drawn from the large sample size scenario described earlier, which features admixture. In subsequent generations, every individual is paired as follows. The local kinship matrix of individuals is stored and updated after every generation, which records the pedigree relatedness; in the first generation, everybody is locally unrelated. Also, individuals are ordered, initially by the 1-dimensional geography, and in subsequent generations paired individuals are grouped and reordered by their average coordinate, preserving the original order when there are ties. For every remaining unpaired individual, one is drawn randomly from the population, and it is paired with the nearest individual that is not a second cousin or closer relative (local kinship must be $< 1/4^3$). Note that every individual is initially genderless, and after pairing one individual in the pair may be set to male and the other to female without giving rise to contradictions. If there are individuals that could not be paired (occurs if unpaired individuals are all close relatives), then the process of pairing individuals randomly is repeated entirely for this generation. If after 100 iterations no solution could be found randomly (there were always unpaired individuals), then the simulation restarts from the very first generation; this may occur for very small populations, but was not observed when $n = 1000$. Once individuals are paired, two children per pair have their genotypes drawn independently of each other. In particular, at every locus, one allele is drawn randomly from one of the parents and the other allele from the other parent. Loci are constructed independently of the rest (no linkage disequilibrium). The simulation continues for 20 generations. As this simulation is very computationally expensive, it was run only once (genotypes did not change as new random traits were constructed as described next).

2.2.3 Trait Simulation

For a given genotype matrix (simulated or real), a simulated complex trait that follows the additive quantitative trait model in Eq. (1) is constructed as follows. In all cases we set the heritability of the trait to be $h^2 = 0.8$. We varied the number of causal loci (m_1) together with the number of individuals (n) so power would remain balanced: for the $n = 1,000$ cases we set $m_1 = 100$, whereas the $n = 100$ simulation had $m_1 = 10$.

Each simulation replicate consists of different causal loci with different effect sizes, as follows. The non-genetic effects are drawn from $\epsilon_j \sim \text{Normal}(0, 1 - h^2)$ independently for each individual j. A subset of size m_1 of loci was selected at random from the genotype matrix to be causal loci. The effect size β_i at each causal locus i is drawn initially from a Standard Normal distribution. At non-causal loci i we have $\beta_i = 0$. Under the kinship model, the resulting genetic variance component is given by

$$\sigma_0^2 = \sum_{i=1}^{m} 2p_i(1 - p_i)\beta_i^2,$$

where p_i is the true ancestral allele frequency at locus i, which is known in our simulations. The desired genetic variance of h^2 is therefore obtained by multiplying every β_i by $\frac{h}{\sigma_0}$. Lastly, the intercept coefficient in Eq. (1) is set to $\alpha = -\sum_{i=1}^{m} 2p_i\beta_i$, so the trait expectation is zero. This trait simulation procedure is implemented in the `simtrait` R package, available at `https://github.com/OchoaLab/simtrait`.

[TODO: describe trait simulation for real genotype datasets.]

2.2.4 Real human datasets

TODO: As before (Ochoa and Storey, 2018), except with an additional filter so loci are in approximate linkage equillibrium. This is required to keep our evaluations simple (so loci that are not causal are not correlated due to LD to causal loci). Used plink2 with parameters "–indep-pairwise 1000kb 0.3" (translate into words). The final dataset sizes are: ...

2.3 Evaluation of performance

All of the approaches considered here are evaluated in two orthogonal dimensions. The first one—the SRMSD_p statistic below—quantifies the extent to which null p-values are uniform, which is a prerequisite for accurate control of the type-I error and successful FDR control via q-values. The second measure—the area under the precision-recall curve—quantifies the predictive power of each method, which makes it possible to qualitatively compare the statistical power of each method without having to select a single threshold, and most importantly, overcoming the problem of comparing methods that may not have accurate p-values (Bouaziz et al., 2011).

2.3.1 SRMSD_p: a measure of p-value uniformity

From their definition, correct p-values (for continuous test statistics) have a uniform distribution when the null hypothesis holds. This fact is crucial for accurate control of the type-I error, and is a prerequisite for the most common approaches that control the FDR, such as q-values (Storey, 2003; Storey and Tibshirani, 2003). We use the Signed Root Mean Square Deviation (SRMSD) to measure the disagreement between the observed p-value quantiles and the expected uniform quantiles:

$$\text{SRMSD}_p = \text{sgn}(u_{\text{median}} - p_{\text{median}}) \sqrt{\frac{1}{m_0} \sum_{i=1}^{m_0} \left(u_i - p_{(i)}\right)^2},$$

where $m_0 = m - m_1$ is the number of null loci ($\beta_i = 0$ cases only), here i indexes null loci only, $p_{(i)}$ is the ith ordered null p-value, $u_i = (i - 0.5)/m_0$ is its expectation, p_{median} is the median observed null p-value, $u_{\text{median}} = \frac{1}{2}$ is the median expected null p-value, and sgn is the sign function (in this case 1 if $u_{\text{median}} \geq p_{\text{median}}$, -1 otherwise). Thus, $\text{SRMSD}_p = 0$ corresponds to the best performance (well-calibrated p-values), large positive SRMSD_p values indicate anti-conservative p-values, and negative values are overly conservative p-values.

One scenario that achieves the maximum SRMSD_p (or worst performance) is when all estimated p-values approach zero, which is what happens to anti-conservative approaches. In that case all of the observed quantiles approach $p_{(i)} = 0$, and then, in the limit as the number of loci goes to infinity,

the statistic in this worst-case scenario approaches

$$\text{SRMSD}_p \to \sqrt{\int_0^1 u^2 du} = \frac{1}{\sqrt{3}} \approx 0.577.$$

The same worst-case value is achieved if all p-values approach 1 instead of 0, except for the change in sign.

2.3.2 The inflation factor λ

In previous evaluations, test statistic inflation has been used to measure the success of corrections for population structure (Astle and Balding, 2009; Price et al., 2010). The inflation factor λ is defined as the median χ^2 association statistic divided by theoretical median under the null hypothesis (Devlin and Roeder, 1999). The inflation factor can be calculated from the median p-value (in this case across all p-values, not just the null ones) using

$$\lambda = \frac{F^{-1}\left(1 - p_{\text{median}}\right)}{F^{-1}\left(1 - u_{\text{median}}\right)},$$

where p_{median} is the median observed p-value, $u_{\text{median}} = \frac{1}{2}$ is the median expected null p-value, and F is the cumulative density function of the χ^2 distribution (F^{-1} is the quantile function). This equation is useful to compare p-values from statistics that have non-χ^2 distributions (fixed-effect associations can be tested more precisely using the t-test).

To compare the properties of λ and SRMSD_p directly, we shall assume that all p-values are from the null distribution (good agreement is expected when a small proportion of p-values are from the alternative distribution, as is common in association studies). Hence, when null test statistics have their expected distribution, we get $\lambda = 1$ and $\text{SRMSD}_p = 0$. However, any other null test statistic distribution with the same median results in $\lambda = 1$ as well, but $\text{SRMSD}_p \neq 0$ unless the entire test statistic distribution is as expected, thus revealing the flaw of λ that SRMSD_p overcomes. The $\lambda > 1$ case always gives $\text{SRMSD}_p > 0$, and corresponds to inflated test statistics (but results in smaller than expected, or anti-conservative, p-values), which occurs when residual population structure is present. On the other hand, $\lambda < 1$ always gives $\text{SRMSD}_p < 0$, and arises if p-values are larger

than expected, or conservative. Thus, $\lambda \neq 1$ always implies $\text{SRMSD}_p \neq 0$ (but not the other way around), and SRMSD_p has the same sign as $\lambda - 1$. Overall, the weakness of λ is that it depends only on the median of the distribution, whereas the SRMSD_p makes use of the complete p-value distribution to evaluate its uniformity, which is stricter. The drawback is that SRMSD_p requires knowing which loci are null, so unlike λ, it is only applicable for simulated traits.

2.3.3 The area under the precision-recall curve

Precision and recall are two common measures for evaluating binary classifiers. Let c_i be the the true classification of locus i, where $c_i = 1$ for truly causal loci (if the true $\beta_i \neq 0$, where the alternative hypothesis holds), and $c_i = 0$ otherwise (null cases). For a given method and some threshold t on its per-locus test statistics, the method predicts a classification $\hat{c}_i(t)$ (for example, if t_i is the test statistic, the prediction could be $\hat{c}_i(t) = 1$ if $t_i \geq t$, and $\hat{c}_i(t) = 0$ otherwise). Across all loci, the number of true positives (TP), false positives (FP) and false negatives (FN) at the given threshold t is given by

$$\begin{aligned}
\text{TP}(t) &= \sum_{i=1}^{m} c_i \hat{c}_i(t), \\
\text{FP}(t) &= \sum_{i=1}^{m} (1 - c_i) \hat{c}_i(t), \\
\text{FN}(t) &= \sum_{i=1}^{m} c_i \left(1 - \hat{c}_i(t)\right).
\end{aligned}$$

Precision and recall at this threshold are given by

$$\begin{aligned}
\text{Precision}(t) &= \frac{\text{TP}(t)}{\text{TP}(t) + \text{FP}(t)} = \frac{\sum_{i=1}^{m} c_i \hat{c}_i(t)}{\sum_{i=1}^{m} \hat{c}_i(t)}, \\
\text{Recall}(t) &= \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)} = \frac{\sum_{i=1}^{m} c_i \hat{c}_i(t)}{\sum_{i=1}^{m} c_i}.
\end{aligned}$$

The precision-recall curve results from calculating the above two values at every threshold t, tracing a curve as recall goes from zero (everything is classified as null) to one (everything is classified as alternative), and the area under this curve is our final measure AUC_{PR}. A method obtains

the maximum $\text{AUC}_{\text{PR}} = 1$ if there is some threshold that classifies all loci perfectly. In contrast, a method that classifies at random (for example, $\hat{c}_i(t) \sim \text{Bernoulli}(p)$ for any p) has an expected precision (= AUC_{PR}) approximately equal to the overall proportion of alternative cases: $\pi_1 = \frac{m_1}{m} = \frac{1}{m}\sum_{i=1}^{m} c_i$. The AUC_{PR} was calculated using the R package `PRROC`, which computes the area by integrating the correct non-linear piecewise function when interpolating between points (Grau et al., 2015).

2.4 Software

The fixed effects (PCA) genetic association was performed using plink 2 (Chang et al., 2015). The statistical evaluation is a standard linear regression with covariates, which employs the t-test. PCs for the PCA-only version were calculated in R using the `popkinsuppl` package, function `kinship_std`, which calculated the ratio-of-means version of the standard kinship estimator, which is more stable than the more common mean-of-ratios version (Ochoa and Storey, 2016b). [TODO: show equation? maybe easier to use plink's built-in PCs instead, which are MOR with MAF filter.]

The mixed effect (LMM+PCA) genetic association was performed using GCTA (Yang et al., 2011). GCTA uses nearly the same biased kinship matrix estimator $\hat{\mathbf{\Phi}}$ as standard PCA approaches (only the diagonal estimates differ (Yang et al., 2011)). Both the kinship estimates and PCs used with GCTA were estimated using GCTA, for consistency. When running GCTA with large numbers of PCs in the small sample size particular simulation, we often encountered errors such as "the information matrix is not invertible", "analysis stopped because more than half of the variance components are constrained", and "Log-likelihood not converged (stop after 100 iteractions)" (sic), which make sense as we are pushing this complex model hard to be underconstrained, in which cases SRMSD_p and AUC_{PR} were treated as missing. These errors were not observed in our other simulations.

3 Results

We simulate genotype matrices and traits to go with the genotypes, in order to control important features of the population structure and to test all methods in an ideal setting where the true causal

loci are known. Our simulations permit exact identification of true positives, false positives, and false negatives, ultimately yielding two measures of interest: SRMSD_p measure null p-value uniformity and relates to the accuracy of type-I error control (closer to zero is better), while AUC_{PR} measures predictive power (higher is better) and serves as a proxy for statistical power when $\text{SRMSD}_p \approx 0$. However, the simulation of genotypes followed by simulation of the trait leads to a considerable amount of variance in the final measured SRMSD_p and AUC_{PR}, which are random variables. For that reason, every evaluation was replicated 50 times, resulting in a distribution of SRMSD_p and AUC_{PR} values per method. Each replicate consisted of a new genotype matrix drawn from the same structure model of the scenario, followed by a new simulated trait based on this genotype matrix, which included selecting new causal loci with new effect sizes.

All scenarios are based on an admixture simulation from $K = 10$ subpopulations and a resulting generalized $F_{\text{ST}} = 0.1$, which establishes the population structure. We vary the sample size (number of individuals) in order to test the extent to which each of PCA and LMM overfits the population structure as the number of PCs increases ($r \in \{0, ..., 90\}$). The theoretically ideal choice for the number of PCs in this simulation, for fixed effects only (PCA), is $r = K - 1 = 9$ (the rank of the population structure minus the rank of the intercept). Lastly, to push all methods to their limits, we evaluate them in a scenario with both admixture and a complex family structure.

3.1 Large sample size simulation

First we evaluate all methods in the large sample size scenario without close relatives, which has a reasonable number of individuals ($n = 1,000$) typical for genetic association studies.

We being by discussing the performance of the fixed effects model (PCA) as the number of PCs is varied. In this scenario we find a clear transition around $r = 4$ number of PCs for PCA, below of which performance is poor and above of which performance is satisfactory (Fig. 1). In particular, when $r < 3$ we find the largest SRMSD_p values, which indicate that p-values are highly non-uniform and would therefore result in inaccurate type-I error control. The smallest AUC_{PR} values also occur for $r < 3$, showing that not enough PCs results in loss of predictive power as well. Interestingly, performance is optimal when the number of PCs is lower than the ideal value of $r = 9$

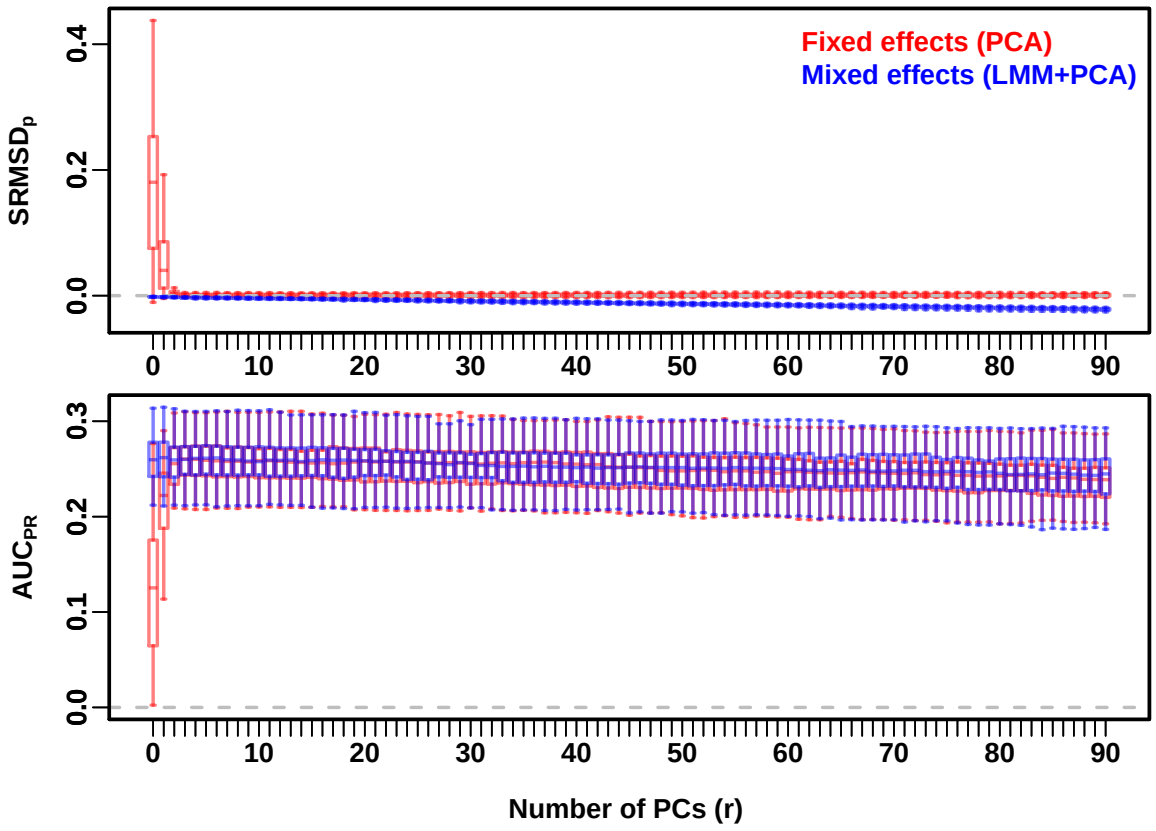

Figure 1: **Evaluation in large sample size admixture scenario.** This simulation has $n = 1,000$ individuals. The PCA and LMM+PCA approaches are tested under varying number of PCs ($r \in \{0, ..., 90\}$ on x-axis), with the distributions (y-axis) of SRMSD$_p$ (top panel) and AUC$_{\text{PR}}$ (bottom panel) for 50 replicates. The best performance is zero SRMSD$_p$ and large AUC$_{\text{PR}}$. For mixed effects (LMM+PCA), the optimal number of PCs is $r = 0$ since SRMSD$_p$ increases and AUC$_{\text{PR}}$ decreases with increasing r. For fixed effects only (PCA), the ideal number of PCs is $r = 9$ in theory, but optimal performance is achieved with as few as $r = 4$ PCs, which results in near zero SRMSD$_p$ and peak AUC$_{\text{PR}}$, and performs as well as the LMM without PCs. PCA with $r < 3$ has incorrect p-values (SRMSD$_p \gg 0$ cases) and lowest predictive power (small AUC$_{\text{PR}}$). Remarkably, PCA is robust in extreme $r \gg 9$ cases, with SRMSD$_p$ near zero up to $r = 90$ and minimal loss of power as r increases to 90.

according to theory (in other words, it is not necessary to model all of the 10 ancestries, modeling the most divergent 5 dimensions suffices). Remarkably, as r is increased up to $r = 90$, then SRMSD_p stays near zero, and there is only a small decrease in AUC_{PR} compared to the optimal $r = 4$ case. Accurate null p-values under excessive numbers of PCs makes sense since the t-test is being used, which is accurate under arbitrary numbers of covariates.

Now we turn to the mixed effects model (LMM+PCA), which models structure as a random effect and is tested here in combination with varying numbers of PCs used as fixed effect covariates. The LMM without PCs ($r = 0$) performs as well as PCA with $r = 4$, with SRMSD_p values near zero and a slightly larger AUC_{PR} values than PCA with $r = 4$. The performance of LMM worsens monotonically as r increases. The decreasing AUC_{PR} of LMM as r increases tracks closely with the decreasing AUC_{PR} of the fixed-effects model only, suggesting that overfitting due to including too many covariates is causing the decrease in AUC_{PR}. Interestingly, SRMSD_p becomes negative for LMM as r increases (p-values become increasingly conservative), an effect that was not observed for fixed effects only. As we will cover in greater depth in the discussion, this is due to LMMs evaluating significance using the asymptotic likelihood ratio test, which becomes less accurate as the number of covariates increases, whereas the fixed effects model uses the t-test, which remains accurate for large numbers of covariates.

Overall, in this common scenario of large sample sizes and no family structure, the LMM approach without PCs performs best, and the (fixed effects only) PCA approach performs as well as LMM as long as enough PCs are used.

3.2 Small sample size simulation

In the previous case PCA performed well when the number of PCs was orders of magnitude greater than its ideal value in theory ($r = 90$ vs $r = 9$), which motivated us to find scenarios where this is no longer the case. We expect the PCA approach to overfit more severely as the number of PCs r approaches the sample size n. Increasing r beyond 90 would never be done in practice; instead, we reduced the number of individuals n to 100, which is small for typical association studies, but which may occur in studies of rare diseases, or be due to low budgets or other constraints. To compensate

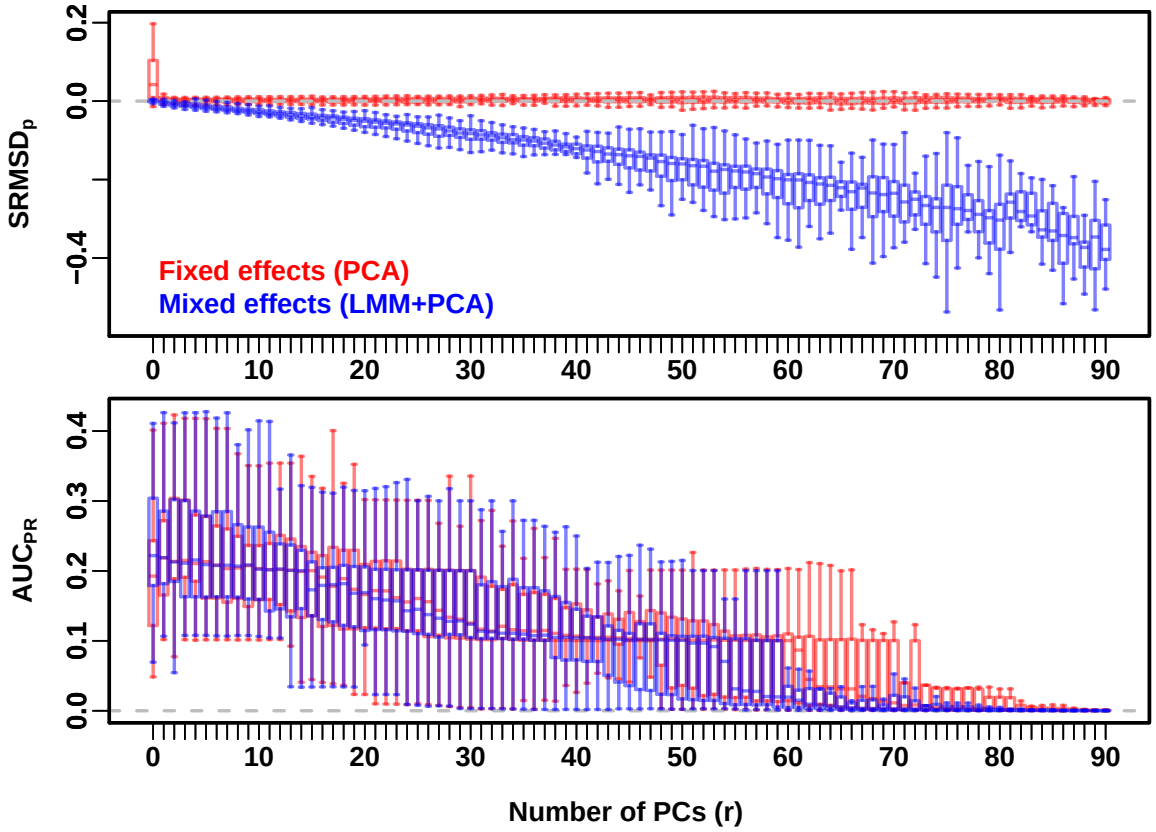

Figure 2: **Evaluation in small sample size admixture scenario.** This simulation has $n = 100$ individuals, otherwise the simulation and figure layout is the same as in Fig. 1. For fixed effects, the pattern for SRMSD_p is similar to the previous figure, while AUC_{PR} drops faster as the number of PCs r increases from $r = 2$ to $r = 90$. For mixed effects, the AUC_{PR} pattern mirrors that of fixed effects, whereas the SRMSD_p becomes negative with r much faster here than in the previous large sample size simulation.

for the loss of power that results from reducing the sample size, we also reduced the number of causal loci from 100 before to $m_1 = 10$, which increases the magnitude of each individual effect size per locus. This reduction in the number of causal loci results in more discreteness in AUC_{PR} values (Fig. 2).

We find for PCA (fixed effects) that the relationship between SRMSD_p and r the same under small and large sample sizes, with ideal near-zero SRMSD_p distributions for $r \geq 2$. On the other hand, we see a more severe overfitting effect here that results in decreased predictive power: AUC_{PR} peaks at $r = 4$, then drops rapidly as r increases, with performance around $r = 25$ that is worse than for $r = 0$, and practically zero AUC_{PR} at $r = 90$ (Fig. 2). Thus, the overfitting effect is quite dramatic in studies with few individuals, where choosing the correct number of PCs is very important.

The behavior of LMM+PCA (mixed effects) with varying numbers of PCs is also a more extreme version than what we observed in the large sample size simulation. The mixed effects model performs increasingly poorly as r increases (SRMSD_p increases and AUC_{PR} decreases). Just as in the first simulation, here the optimal choice is to use LMM without PCs, which successfully avoids overfitting and performs comparably to the best-performing fixed-effects method (with $r = 4$ PCs).

3.3 Family structure admixture simulation

Previous work has shown that PCA performs poorly in the presence of family structure. Here we aim to characterize PCA's behavior in a much more complex structure than before, by simulating a family of admixed founders for 20 generations, so that we may observe numerous siblings, first cousins, etc, while mostly preserving the population structure due to ancestry by biasing pairings for more similar ancestries.

In this case PCA improves in performance (AUC_{PR} and SRMSD_p) until around $r \approx 15$, which is larger since the family structure increases the rank of the genotype matrix compared to the previous simulations without family structure. We find that, although SRMSD_p for PCA decreases monotonically as r increases, this distribution does not go to zero, instead converging to around 0.05 (Fig. 3). In contrast, for LMM+PCA, SRMSD_p is practically zero with $r = 0$ PCs, and

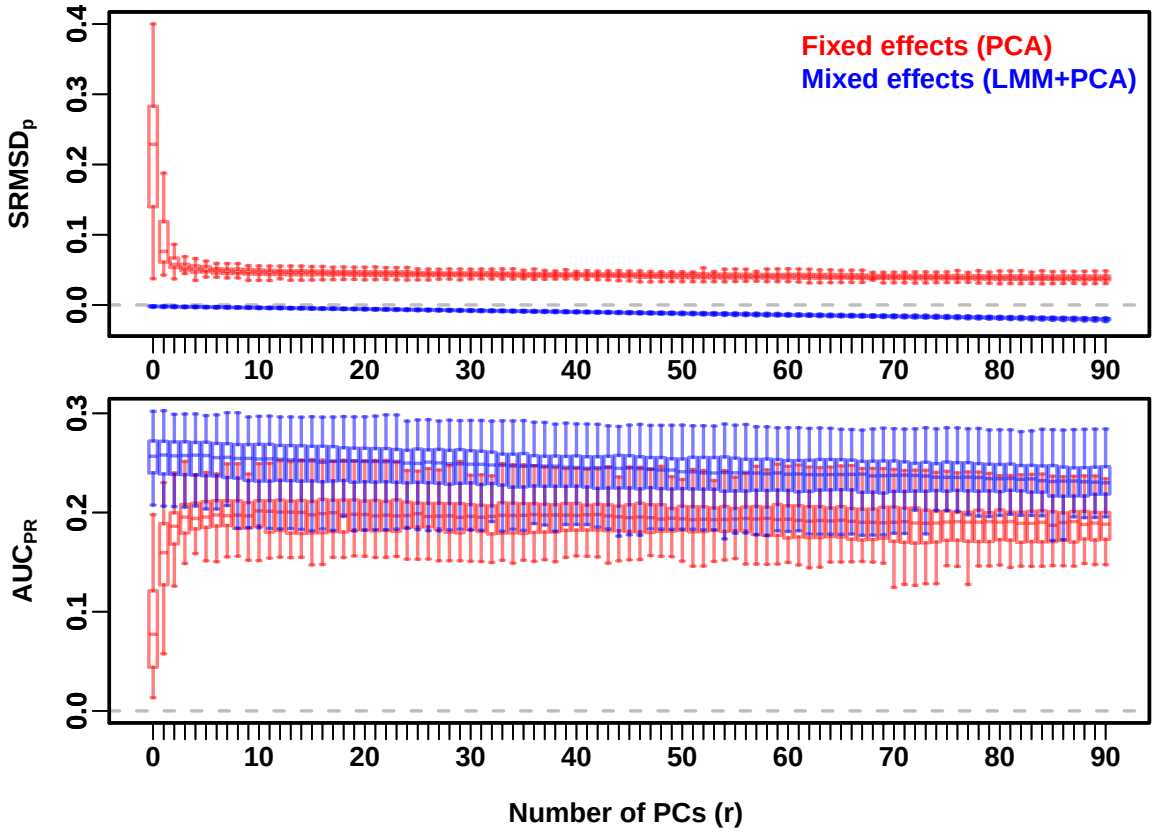

Figure 3: **Evaluation in family structure admixture scenario.** Here there are $n = 1,000$ individuals from a family structure simulation with admixed founders and large numbers of pairs of sibling, first cousins, second cousins, etc, from a realistic random 20-generation pedigree. Unlike previous tests, here PCA never performs as well as LMM, since SRMSD$_p$ (top panel) does not go down to zero as r increases, and there is a significant gap in AUC$_{\text{PR}}$ performance, as expected because of the presence of family structure. The LMM achieves a near-zero SRMSD$_p$ and the highest AUC$_{\text{PR}}$ when no PCs are used ($r = 0$).

SRMSD_p becomes negative slowly with increasing r, overall resembling its performance under no family structure (Fig. 1). There is a sizable gap in AUC_{PR} between LMM with no PCs, which performs best, and the best PCA model. Overall, LMM conclusively outperforms PCA when there is family structure.

4 Discussion

Our main findings are organized into three areas: (1) the robustness of the performance of PCA (fixed effects only) on the number of PCs included; (2) the performance of LMM (mixed effects) when PCs are included; and (3) the overall comparison between LMM and PCA. In each case we will discuss how performance depended on various aspects of each simulation.

4.1 Robustness of PCA (fixed effects only) to the number of PCs

One important conclusion of our evaluation is that the PCA approach for genetic association studies (fixed effects only) is robust to the choice of r (number of PCs), particularly when r is larger, but not smaller, than the optimal choice, and more so when the sample size (number of individuals) is sufficiently large. Thus, while we expect an r that is too small or too large may hurt the performance of PCA (by not modeling enough of the population structure, or by overfitting, respectively), we find that the magnitude of the performance penalty depends very strongly on whether r is too small or too large.

In our simulations that excluded family structure, the optimal choice for PCA (fixed effects only) was $r \approx 4$, which was smaller than the theoretical optimal value of $r = 9$ when we consider the number of ancestries ($K = 10$) and the intercept coefficient. We found that insufficient PCs such as $r = 1$ paid a large penalty in both type-I error control (measured via SRMSD_p) and predictive power (AUC_{PR}; Figs. 1 and 2). In contrast, r can be much larger than its optimal value with no penalty in terms of type-I error control (regardless of sample size), and only a negligible cost in predictive power when sample sizes are large (Fig. 1). Successful type-I error control (*i.e.*, near zero SRMSD_p) for fixed effects and for large enough numbers of PCs (r) follow from employing the t-test for calculating statistical significance, which is more precise than the asymptotic χ^2 test

when there are large numbers of covariates. The loss of predictive power by using excessive PCs is only pronounced when the number of individuals n is much smaller than is common nowadays, *i.e.*, in the hundreds (Fig. 2). This robustness of PCA to the choice of r has long been anecdotal only (Price et al., 2006; Kang et al., 2010). Here we conclusively found that it is far safer to err on the side of larger r, especially for large sample sizes, although testing several r values via simulations is always recommended.

The presence of family structure is the main weakness of the PCA approach. Type-I error is never accurately controlled (SRMSD_p was non-zero) for any number of PCs we tested (up to $r = 90$; Fig. 3). The pattern of AUC_{PR} here resembles that of the simulation without family structure (except optimal performance is achieved around $r = 15$), but best performance is never as good as for LMMs.

4.2 Robustness of LMM (mixed effects) to the number of PCs

We also evaluated use of an LMM (Linear Mixed-effects Model) in combination with PCs, an approach that has performed best in certain previous evaluations (Zhao et al., 2007; Price et al., 2010). In all of our simulations we found that LMM with zero PCs always performed as well or better than using any non-zero number of PCs Figs. 1 to 3). Our findings make sense since the PCs are the eigenvectors of the kinship matrix used to model the random effects, so including both is redundant. It is possible that there are more complex scenarios than considered here where an LMM with PCs performs better than LMM without PCs.

Although our evaluations favor using LMM without PCs, our tests also revealed that LMMs with large numbers of covariates suffer from inaccurate p-values, an LMM-specific problem that we did not observe for fixed effects. The key difference is that p-values for fixed effects are calculated using the t-test, which is robust to large numbers of covariates, while mixed effects necessitate an approximatimation: the likelihood ratio test has test statistics only asymptotically follow the χ^2 distribution used to calculate p-values. These approximations become more unreliable the more covariates are present, as observed in our evaluations. Fortunately, LMM with zero PCs has the best predictive power, and in that case p-values were also accurate in all of our simulations. However,

the same issue of problematic statistics can arise where a large number of (non-PC) covariates are being fit in an association study, particularly when the number of individuals is small (Fig. 2).

4.3 Overall comparison between LMM and PCA

Previous work has mostly ruled in favor of LMM (mixed effects) versus PCA (fixed effects only), or otherwise tend to show comparable performance. Our simulations agree that LMM is always one of the best choices, but the gap in performance of PCA (with the optimal choice of numbers of PCs) varies depending on the simulation. The clearest theoretical advantage of the LMM is its ability to model family relatedness. We confirm this in our family structure simulation, where LMM (without PCs) significantly outperforms the best PCA model in terms of both type-I error control (SRMSD_p) and predictive power (AUC_{PR}; Fig. 3). The other theoretical advantage of LMM over PCA is in having fewer degrees of freedom (Hoffman, 2013); however, we did not see this improving AUC_{PR} in our small sample size simulation where overfitting overall played a large role: the best PCA model performed as well as LMM (Fig. 2). However, PCA can perform poorly when the number of PCs is not chosen carefully, especially under small sample sizes, which may explain early claims that PCA was not effective in preventing test statistic inflation (Epstein et al., 2007; Kimmel et al., 2007; Luca et al., 2008).

Evaluations from others suggest that PCA can outperform LMM when there are loci under selection or otherwise highly differentiated, and rare variants (Price et al., 2010; Wu et al., 2011; Yang et al., 2014). One potential explanation is that the additional degrees of freedom in the additional PCs enables the LMM to better model loci when the pure ($r = 0$) LMM model assumptions break. However, our simulations indicate that LMM is the best choice for most population structures. Since PCA matches the performance of LMM when there is no family structure, this argues against the simple characterization that either fixed or random effects are in principle superior models for association studies (Price et al., 2010; Sul and Eskin, 2013; Price et al., 2013; Sul et al., 2018). The driving principle that explains the PCA and LMM performance is the dimensionality of the genetic structure: when there is no family structure, then structure is low dimensional, so either PCA and LMM are well suited for modeling the population structure. In contrast, the presence

of family relationships increases the dimensionality of the genetic structure, and in that case PCA performs poorly because either not enough PCs are used, or instead so many PCs are needed that overfitting occurs. LMMs naturally model pedigrees via kinship matrices, and in general they fit arbitrary high-dimensional genetic structures, doing so with fewer effective parameters than PCA, preventing overfitting of the data.

In addition to our findings, another advantage of LMMs is that the user need not choose parameters arbitrarily, as is sometimes the case for the number of PCs in the PCA approach. Thus, the key practical advantage of PCA is its speed, as fitting even the most efficient existing mixed-effects models has a higher computational complexity. However, if the number of PCs to use is not known, then expensive simulations need to be performed to determine this number, and this may also eliminate the computational advantage of PCA compared to LMM. For these reasons, we recommend always using LMMs when it is computationally feasible. On the other hand, it is important to understand whether use of PCA in past association studies has resulted in a relative increase of false positives or false negatives. Our simulations suggest that PCA results in accurate association studies as long as (1) there are no close relatives present in the study, and (2) enough PCs were used. Common practice of excluding close relatives and using at least 10 PCs suggests that the majority of previous genetic association studies that used PCA have presented valid results on par with what an LMM would have reported.

References

Abraham, Gad and Michael Inouye (9, 2014). "Fast Principal Component Analysis of Large-Scale Genome-Wide Data". *PLOS ONE* 9(4), e93766.

Abraham, Gad, Yixuan Qiu, and Michael Inouye (1, 2017). "FlashPCA2: principal component analysis of Biobank-scale genotype datasets". *Bioinformatics* 33(17), pp. 2776–2778.

Alexander, David H., John Novembre, and Kenneth Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9), pp. 1655–1664.

Astle, William and David J. Balding (2009). "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

Balding, D. J. and R. A. Nichols (1995). "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity". *Genetica* 96(1), pp. 3–12.

Bouaziz, Matthieu, Christophe Ambroise, and Mickael Guedj (21, 2011). "Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies". *PLOS ONE* 6(12), e28845.

Cabreros, Irineo and John D. Storey (1, 2019). "A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis". *Genetics* 212(4), pp. 1009–1029.

Chang, Christopher C. et al. (25, 2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *GigaScience* 4(1), p. 7.

Devlin, B. and Kathryn Roeder (1, 1999). "Genomic Control for Association Studies". *Biometrics* 55(4), pp. 997–1004.

Epstein, Michael P., Andrew S. Allen, and Glen A. Satten (1, 2007). "A Simple and Improved Correction for Population Stratification in Case-Control Studies". *The American Journal of Human Genetics* 80(5), pp. 921–930.

Falush, Daniel, Matthew Stephens, and Jonathan K. Pritchard (2003). "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". *Genetics* 164(4), pp. 1567–1587.

Galinsky, Kevin J. et al. (3, 2016). "Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia". *The American Journal of Human Genetics* 98(3), pp. 456–472.

Gopalan, Prem et al. (2016). "Scaling probabilistic models of genetic variation to millions of humans". *Nat. Genet.* 48(12), pp. 1587–1590.

Grau, Jan, Ivo Grosse, and Jens Keilwagen (1, 2015). "PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R". *Bioinformatics* 31(15), pp. 2595–2597.

Hoffman, Gabriel E. (2013). "Correcting for population structure and kinship using the linear mixed model: theory and extensions". *PLoS ONE* 8(10), e75707.

Jacquard, Albert (1970). *Structures génétiques des populations.* Paris: Masson et Cie.

Janss, Luc et al. (1, 2012). "Inferences from Genomic Models in Stratified Populations". *Genetics* 192(2), pp. 693–704.

Jolliffe, Ian T. (2002). *Principal Component Analysis.* 2nd ed. New York: Springer-Verlag.

Kang, Hyun Min et al. (2010). "Variance component model to account for sample structure in genome-wide association studies". *Nat. Genet.* 42(4), pp. 348–354.

Kimmel, Gad et al. (1, 2007). "A Randomization Test for Controlling Population Stratification in Whole-Genome Association Studies". *The American Journal of Human Genetics* 81(5), pp. 895–905.

Lee, Seokho et al. (2012). "Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome-Wide Association Studies". *Genetic Epidemiology* 36(4), pp. 293–302.

Li, Mingyao et al. (15, 2010). "Correcting population stratification in genetic association studies using a phylogenetic approach". *Bioinformatics* 26(6), pp. 798–806.

Li, Qizhai and Kai Yu (2008). "Improved correction for population stratification in genome-wide association studies by identifying hidden population structures". *Genetic Epidemiology* 32(3), pp. 215–226.

Luca, Diana et al. (8, 2008). "On the Use of General Control Samples for Genome-wide Association Studies: Genetic Matching Highlights Causal Variants". *The American Journal of Human Genetics* 82(2), pp. 453–463.

Malécot, Gustave (1948). *Mathématiques de l'hérédité.* Masson et Cie.

Ochoa, Alejandro and John D. Storey (2016a). "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". Submitted, preprint at `http://biorxiv.org/content/early/2016/10/27/083915`.

Ochoa, Alejandro and John D. Storey (2016b). "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". Submitted, preprint at `http://biorxiv.org/content/early/2016/10/27/083923`.

— (2018). "New kinship and F_{ST} estimates reveal higher levels of differentiation in the world-wide human population". Submitted, preprint at `http://biorxiv.org/content/early/....`

Patterson, Nick, Alkes L Price, and David Reich (22, 2006). "Population Structure and Eigenanalysis". *PLoS Genet* 2(12), e190.

Price, Alkes L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". *Nat. Genet.* 38(8), pp. 904–909.

Price, Alkes L. et al. (2010). "New approaches to population stratification in genome-wide association studies". *Nature Reviews Genetics* 11(7), pp. 459–463.

— (2013). "Response to Sul and Eskin". *Nature Reviews Genetics* 14(4), p. 300.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000a). "Inference of population structure using multilocus genotype data". *Genetics* 155(2), pp. 945–959.

Pritchard, Jonathan K. et al. (2000b). "Association Mapping in Structured Populations". *The American Journal of Human Genetics* 67(1), pp. 170–181.

Song, Minsun, Wei Hao, and John D. Storey (2015). "Testing for genetic associations in arbitrarily structured populations". *Nat. Genet.* 47(5), pp. 550–554.

Storey, John D. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value". *Ann. Statist.* 31(6). Mathematical Reviews number (MathSciNet): MR2036398; Zentralblatt MATH identifier: 02067675, pp. 2013–2035.

Storey, John D. and Robert Tibshirani (2003). "Statistical significance for genomewide studies". *Proceedings of the National Academy of Sciences of the United States of America* 100(16), pp. 9440–9445.

Sul, Jae Hoon and Eleazar Eskin (2013). "Mixed models can correct for population structure for genomic regions under selection". *Nature Reviews Genetics* 14(4), p. 300.

Sul, Jae Hoon, Lana S. Martin, and Eleazar Eskin (2018). "Population structure in genetic studies: Confounding factors and mixed models". *PLoS Genet.* 14(12), e1007309.

Tucker, George, Alkes L. Price, and Bonnie Berger (1, 2014). "Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select". *Genetics* 197(3), pp. 1045–1049.

Voight, Benjamin F. and Jonathan K. Pritchard (2, 2005). "Confounding from Cryptic Relatedness in Case-Control Association Studies". *PLOS Genetics* 1(3), e32.

Wang, Kai, Xijian Hu, and Yingwei Peng (2013). "An Analytical Comparison of the Principal Component Method and the Mixed Effects Model for Association Studies in the Presence of Cryptic Relatedness and Population Stratification". *HHE* 76(1), pp. 1–9.

Wojcik, Genevieve L. et al. (2019). "Genetic analyses of diverse populations improves discovery for complex traits". *Nature* 570(7762), pp. 514–518.

Wright, S. (1951). "The genetical structure of populations". *Ann Eugen* 15(4), pp. 323–354.

Wu, Chengqing et al. (2011). "A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies". *Annals of Human Genetics* 75(3), pp. 418–427.

Yang, Jian et al. (7, 2011). "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1), pp. 76–82.

Yang, Jian et al. (2014). "Advantages and pitfalls in the application of mixed-model association methods". *Nat Genet* 46(2), pp. 100–106.

Yu, Jianming et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". *Nat. Genet.* 38(2), pp. 203–208.

Zhang, Feng, Yuping Wang, and Hong-Wen Deng (14, 2008). "Comparison of Population-Based Association Study Methods Correcting for Population Stratification". *PLOS ONE* 3(10), e3392.

Zhang, Shuanglin, Xiaofeng Zhu, and Hongyu Zhao (2003). "On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals". *Genetic Epidemiology* 24(1), pp. 44–56.

Zhao, Keyan et al. (19, 2007). "An Arabidopsis Example of Association Mapping in Structured Samples". *PLOS Genetics* 3(1), e4.

Zhou, Quan, Liang Zhao, and Yongtao Guan (2016). "Strong Selection at MHC in Mexicans since Admixture". *PLoS Genet.* 12(2), e1005847.