Fast Probabilistic Whitening Transformation for Ultra-High Dimensional Genetic Data

Gabriel E. Hoffman* and Panos Roussos

Center for Disease Neurogenomics,
Department of Psychiatry,
Department of Genetics and Genomic Sciences,
Icahn School of Medicine at Mount Sinai,
New York, NY, USA

Center for Precision Medicine and Translational Therapeutics,
Mental Illness Research, Education and Clinical Center VISN2,
James J. Peters VA Medical Center, Bronx, NY, USA.

*Contact: gabriel.hoffman@mssm.edu

Statistical methods often make assumptions about independence between the samples or features of a dataset. Yet correlation structure is ubiquitous in real data, so these assumptions are often not met in practice. Whitening transformations are widely applied to remove this correlation structure. Existing approaches to whitening are based on standard linear algebra, rather than a probabilistic model, and application to high dimensional datasets with n samples and p features is problematic as p approaches or exceeds n. Moreover, the computational time becomes prohibitive since the naive transform is cubic in p. Here we propose a probabilistic model for data whitening and examine its properties based on first principles as p increases. We demonstrate the statistical properties of the probabilistic model and derive a remarkably efficient algorithm that is linear instead of cubic time in the number of features. We examine the out-of-sample performance of the probabilistic whitening model on simulated data, as well as real gene expression and genotype data. In an application to impute z-statistics from unobserved genetic variants from a genome-wide association study of schizophrenia, the probabilistic whitening transformation, implemented in our open source R package `decorrelate`, had the lowest mean square error while being up to an order of magnitude faster than other methods.

1 Introduction

Correlation structure is ubiquitous in real world datasets. Yet statistical models often make assumptions about independence between a dataset's rows or columns that are not satisfied practice. Whitening transformations are a widely used preprocessing step to remove correlation structure in the data before performing other analyses [1]. In general, for a data matrix $Y_{n\times p}$ with n samples and p features, the procedure computes the sample covariance matrix between features, $\Sigma_{p\times p}$, and transforms the observed data by post-multiplying by the non-unique whitening matrix $\Sigma^{-\frac{1}{2}}$ defined using standard linear algebra [1]. Transforming rows instead follows a complementary procedure. Covariance and their corresponding whitening matrices, along with correlation and their corresponding 'decorrelating' matrices, play a central role in many standard analysis methods such as multivariate regression, canonical correlation analysis, linear and quadratic discriminate analysis, Gaussian mixture models, Maholanobis distance, multivariate outlier detection, and independent component analysis [2–4], linear mixed models [5, 6], as well as deep learning applications [7–9].

For low dimensional data where $n \gg p$ seen in classical statistical settings, whitening transforms the data to have an identity sample covariance matrix. This procedure assumes that the true covariance matrix is known, or is well estimated by the sample covariance matrix. Yet the use of the sample covariance matrix for this transformation can be problematic for two distinct reasons. Computationally, evaluating the covariance matrix and creating the whitening matrix requires $\mathcal{O}(p^2)$ memory and $\mathcal{O}(p^3)$ time, which becomes infeasible for large p [1]. Statistically, the transformation must be adapted for the high dimensional case since the sample covariance matrix is no longer full rank or invertible. In modern high dimensional datasets, these computational and statistical limitations raise issues for the application of whitening transforms either as a preprocessing step, or a central part of widely used statistical methods.

Beyond general applications in statistical analysis, the field of statistical genetics makes extensive use of whitening transformations in order to account for the correlation between nearby genetic variants. This correlation structure is termed 'linkage disequlibirum' or 'LD' and is a major complication in downstream analysis of regression summary statistics (i.e. coefficient estimates or z-statistics) from genome-wide association studies [10]. It is common practice to estimate the correlation matrices from an external 'reference panel' of samples not included in the study of interest [10]. These 'out-of-sample' estimates are then used either as a preprocessing step to transform the summary statistics to be approximately uncorrelated, or internally in a statistical model. This approach has become standard in the field for phenotype prediction [11, 12], statistical fine-mapping [13–15], imputing z-statistics [16], detecting anomalous z-statistics [17], multiple testing correction [18], identifying genome annotations enriched for risk variants [19], and simulating knockoff datasets with similar correlation structure [20]. While the literature on whitening transformations has focused on estimating and applying the transform on the same data [1], an understanding of the performance of these out-of-sample applications has been lacking [21].

Direct estimation of covariance and precision matrices is a huge field of research [22–24]. Many

approaches apply regularization to ensure the estimated covariance or precision matrix has specific properties. Widely used methods apply regularization so that the estimated covariance matrix is positive definite by shrinking the sample eigen-values [25, 26], or ensure sparsity of the covariance matrix and its inverse by shrinking regression coefficients corresponding to partial correlation coefficients [27]. Others regularize the condition number [28]. There have been many methods formulated as a convex combination of two matrices. Some methods use a convex combination between a non-sparse and sparse components [29, 30]. Stein-type estimators use a convex combination between the sample covariance and a target matrix [25, 26, 31]. A probabilistic approach modeling the observed data as a multivariate Gaussian with an inverse Wishart prior on the covariance gives a posterior mean that is a convex combination between the sample covariance matrix and the prior scale matrix [32–36].

Yet whitening transformations used in a preprocessing step or internally in a statistical model do not require *direct* estimates of the covariance matrix. Instead, covariance matrices are merely nuisance parameters: essential to the model, but not of direct interest. In data whitening, only the transformed data, rather the covariance matrix itself, is of interest. Here we seek to account for the covariance, rather than estimate it directly, and we propose an implicit covariance approach. This important distinction motivates the probabilistic framework, statistical theory and computationally efficient algorithms we use here to perform whitening transformations without directly estimating the covariance or whitening matrices.

Bayesian or, more broadly, probabilistic formulations of existing techniques can lead to deeper understanding of model behavior, generalization of the model, regularization to incorporate prior knowledge, or incorporation into a hierarchical model [3, 4]. Widely used Gaussian mixture models and extensions follow from a probabilistic formulation of the k-means problem, and probabilistic formulations of principal components analysis have long been influential in statistical machine learning [37, 38]. Yet we are not aware of such a rigorous probabilistic formulation for the widely used whitening transformation.

Here we propose a probabilistic formulation of the whitening transformation under Gaussian noise, and incorporate regularization in the form of an inverse-Wishart prior distribution on the covariance matrix. This produces a well-known estimator with a single hyperparameter that can be efficiently estimated using an empirical Bayes approach. Using first principles of the probabilistic model, we give a direct interpretation of the hyperparameter and examine the statistical behavior of this whitening transformation. We demonstrate that the probabilistic whitening transformation does not require direct estimation of the covariance or whitening matrices, and derive a remarkably efficient approach to apply the transformation to ultra-high dimensional data.

We examine the empirical performance of the probabilistic whitening transformation on simulated data, gene expression data from the Genotype Tissue Expression (GTEx) project [39], and genetic data from the 1000 Genomes Project [40]. Finally we apply the whitening transformation to a genome-wide association study of schizophrenia [41], and impute z-statistics of unobserved variants using an external reference panel.

2 Results

2.1 Introduction to whitening transformations

Consider observed data $Y_{n\times p}$ with n samples and p features drawn from a distribution with mean vector $\vec{\mu}$, identity covariance between rows, and covariance between features given by $\Sigma_{p\times p}$. Let

$$Y \sim [\vec{\mu}, I_n, \Sigma] \tag{1}$$

indicate these properties without specifying particular distributional assumptions. Transforming the data by the non-unique whitening matrix $\Sigma^{-\frac{1}{2}}$ produces mean and covariances according to

$$Y\Sigma^{-\frac{1}{2}} \sim \left[\vec{\mu}\,\Sigma^{-\frac{1}{2}}, I_n, I_p\right]. \tag{2}$$

where Σ is known. This follows from the definition of covariance and the standard property that $\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} = I$. While it is common to assume a multivariate Gaussian distribution for the observed data, this transformation applies to the first two moments of any multivariate distribution. Applying this transformation scales and rotates the data to produce an identity covariance matrix. The whitening matrix is not uniquely defined and different transformation matrices result in different rotations all with identity covariance [1].

The sample covariance is $\hat{\Sigma} = n^{-1}Y^TY$, following mean centering of the columns. Let the singular value decomposition of the scaled observed data be

$$Y/\sqrt{n} \;=\; U_{n\times p}D_{p\times p}V^T_{p\times p}, \tag{3}$$

with singular values $\vec{d}$, so that the eigen decomposition of the sample covariance is $\hat{\Sigma} = VD^2V^T$ and the whitening matrix $\hat{\Sigma}^{-\frac{1}{2}} = VD^{-1}V^T$ transforms the data to have identity covariance (**SI 5.1**).

This definition of the whitening matrix minimizes the total squared distance between original and transformed data, and is termed the 'zero-phase components analysis' or ZCA whitening [1]. This three step transform, one for each matrix multiplication, can be interpreted as rotating, scaling, then unrotating the observed data to give an identity sample covariance (**Figure 1**). This whitening transform is the most intuitive since it returns the data to its original axes, and it is the foundation for the transformations we describe below.

In practice, the true value of Σ is not known and it must be estimated from the observed data Y. An exact whitening transformation that produces an identity covariance matrix is only possible when $n > p$ so that that sample covariance matrix, $\hat{\Sigma}$, is invertible. Even in this case, the transform can be unstable when the trailing singular-values are very small, since evaluating $D^{-1} = \mathrm{diag}(1/\vec{d})$ becomes very unstable for small values of d_i and are sensitive to precision limitations in numerical linear algebra. In high dimensional data with $n < p$, the trailing singular values are exactly zero, so some form of regularization is needed. An exact whitening transformation no longer exists in these cases, we want to learn a transformation that that transformed data has a covariance close to identify by some metric (**SI 5.2**).

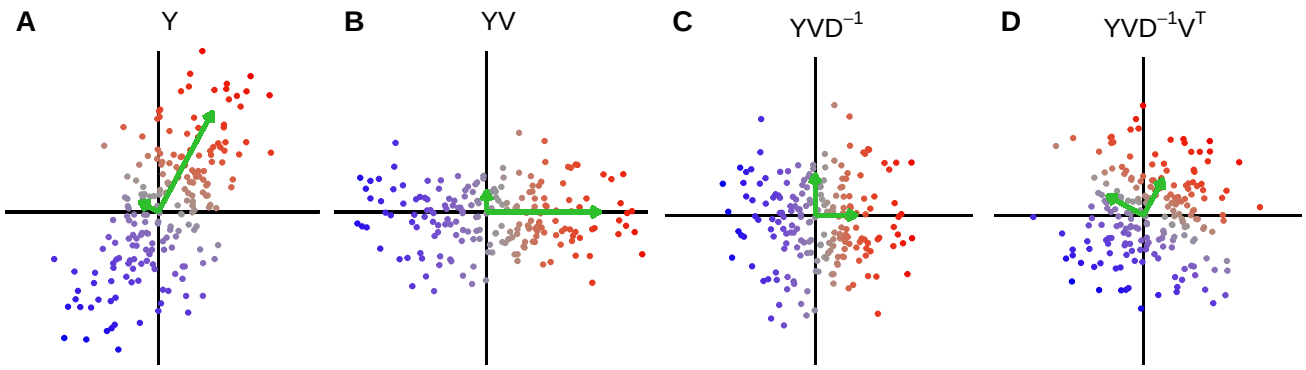

Figure 1: Intuition for data whitening transformation. A) Original data, **B)** Data rotated along principal components, **C)** Data rotated and scaled, **D)** Data rotated, scaled and rotated back to original axes. Green arrows indicate principal axes and lengths indicate eigen-values.

2.2 A probabilistic model for whitening transformations

Here we propose a whitening transform based on an probabilistic model of the observed data and demonstrate superior statistical performance compared to standard data whitening transformations [1], even in the low-dimension case where both are applicable. In proposing a probabilistic model, we consider the observed data to have a matrix normal distribution[1] with unknown mean, identity covariance between rows, and unknown covariance between columns. We then place an inverse Wishart (IW) prior on the covariance matrix between columns. The normal assumption is of course natural for unbounded continuous data, and the IW is natural because it is the conjugate prior so that the posterior distribution of the covariance is also IW [42]. Formally,

$$Y \sim \mathcal{MN}(\vec{\mu}, I_n, \Sigma) \tag{4}$$

$$\Sigma|T \sim \mathcal{IW}\left((\delta - p - 1)T, \delta\right) \tag{5}$$

where T is the target matrix of the prior and $\delta > p-1$. Based on this model, the posterior expectation of the covariance is

$$E\left[\Sigma|Y\right] = (1-\lambda)\hat{\Sigma}_{MLE} + \lambda T \tag{6}$$

$$\lambda = (\delta - p - 1)/(\delta + n - p - 1) \tag{7}$$

where $\hat{\Sigma}_{MLE} = n^{-1}Y^TY$ is the sample covariance matrix and $\lambda \in [0, 1]$ by definition [32–35] The posterior expectation is therefore a convex combination between the sample covariance, $\hat{\Sigma}_{MLE}$, and the prior target matrix, T.

Setting the target matrix to be symmetric positive definite (SPD) ensures that the posterior expected covariance is also SPD [33]. Restricting T to be a minimally informative prior target corresponds to

[1] We use the matrix normal here in order to model covariance between columns while having independent rows, but the results remain unchanged when transposing the observed data and using the multivariate normal notation instead.

setting all prior covariances to zero. We focus here on the important case where the IW prior is centered around a scaled identity matrix, and we demonstrate that the estimate has exceptionally efficient computational scaling and produces estimates that are tractable to study analytically. Following previous work [25, 26, 33–35, 43] we set the target matrix to be $T = \nu I$ where $\nu = \text{tr}(\hat{\Sigma}_{MLE})/p$ is the average variance across all p features. This gives the important special case we consider here with

$$\hat{\Sigma}_\lambda = (1-\lambda)\hat{\Sigma}_{MLE} + \lambda\nu I \tag{8}$$

where $\hat{\Sigma}_\lambda$ is defined as the posterior expected covariance from the probabilistic model.

In the important case of out-of-sample whitening where the whitening transform is learned from a training set before being applied to the observed data, the standard whitening transform motived by linear algebra assumes the correlation structure of the training and observed data are the same. This is a strong assumption that is rarely satisfied in practice. The finite size of the training set means that correlation structure will differ from the observed data, so the standard whitening transform will overfit the training data. Instead, the probabilistic model substantially relaxes this central assumption, and assumes only that the training and observed data are drawn from the same distribution.

In practice, the shrinkage can be applied to the correlation matrix while retaining the diagonal variance terms (**SI 5.3**).

2.3 Estimating the shrinkage parameter

Covariance estimators that are convex combinations of the sample covariance and target matrices have received much attention. The estimator for λ depends on assumptions about the data and the specified loss function. Ledoit and Wolf [25] and Schäfer and Strimmer [26] propose minimizing the expected squared Frobenius loss between the true and estimated covariance matrices, then use a consistent estimate of λ computed from the observed data. Further work extended this approach to improve estimation of λ in the high dimensional setting [31, 43, 44]. Others use a probabilistic model [32–35], cross-validation [45, 46] or the out-of-sample likelihood [47, 48]. Yet these methods are designed for direct estimation of covariance or precision matrices, rather than data whitening.

Following the probabilistic formulation, we take a Bayesian approach to estimating λ, and we refer to this as the Gaussian Inverse Wishart empirical Bayes (GIW-EB) model. We integrate out the covariance matrix and maximize the likelihood of the observed data as a function of λ. Due to the analytic properties of the Gaussian and IW distributions, the covariance matrix can be treated as a nuisance parameter and marginalized out according to

$$p(Y|\delta,\nu) = \int p(Y|\Sigma,\delta,\nu)p(\Sigma|\delta,\nu)d\Sigma, \tag{9}$$

so the marginal distribution is multivariate Student-t [42]. The hyperparameter δ, and therefore λ, can be estimated directly from data using an empirical Bayes approach by maximizing $p(Y|\delta,\nu)$ (**SI 5.4**). Given the special form of the target matrix, νI, we demonstrate that the data only enters the

log-likelihood through the sample singular values. This has important implications for the probabilistic whitening transformation described below.

2.4 Properties of posterior expected covariance

The behavior of this probabilistic whitening transformation corresponding to the posterior expected covariance in Equation (8) can be understood in terms of the spectral decomposition of the observed data matrix, Y.

Based on the singular value decomposition of the observed data, the eigen decomposition of $\hat{\Sigma}_\lambda$ is

$$\begin{aligned} \hat{\Sigma}_\lambda &= (1-\lambda)\hat{\Sigma}_{MLE} + \lambda\nu I && (10) \\ &= V \mathrm{diag}\left[(1-\lambda)\vec{d}^{\,2} + \lambda\nu\right] V^T. && (11) \end{aligned}$$

Therefore the eigen vectors of $\hat{\Sigma}_\lambda$ are the same as for the sample covariance matrix, but the eigen-values are shrunk so that the jth eigen-value is $(1-\lambda)d_j^2 + \lambda\nu$. Since $\lambda \in [0,1]$, $\nu > 0$, and $d_j^2 \geq 0$, it follows that the eigen-values are always positive, so that $\hat{\Sigma}_\lambda$ is aways positive definite and invertible, even when $\hat{\Sigma}_{MLE}$ is not.

2.5 Statistical properties of probabilistic whitening transformation

The probabilistic whitening transformation can be understood in similar terms as the posterior expected covariance. Consider a fixed (i.e. non-random) latent data matrix $X_{n\times p}$ with uncorrelated columns, and let the observed data $Y_{n\times p}$ be generated by applying a rotation based on a symmetric positive definite covariance matrix $\Sigma_{p\times p}$ so that

$$Y \sim \left[X\Sigma^{\frac{1}{2}}, I_n, \Sigma\right]. \tag{12}$$

Given the observed data, Y, estimate the covariance matrix as $\hat{\Sigma}_\lambda$ and apply the whitening transformation so that

$$Y\hat{\Sigma}_\lambda^{-\frac{1}{2}} \sim \left[X\Sigma^{\frac{1}{2}}\hat{\Sigma}_\lambda^{-\frac{1}{2}}, I_n, \hat{\Sigma}_\lambda^{-\frac{1}{2}}\Sigma\hat{\Sigma}_\lambda^{-\frac{1}{2}}\right]. \tag{13}$$

Since $\hat{\Sigma}_\lambda$ converges to Σ as $n \to \infty$ for a fixed p [33], the whitening transform recovers the original signal according to

$$\lim_{n\to\infty} Y\hat{\Sigma}_\lambda^{-\frac{1}{2}} \xrightarrow{d} [X, I_n, I_p]. \tag{14}$$

The whitening transformation can then be expressed in terms of singular vectors and modified singular values according to

$$\boxed{Y\hat{\Sigma}_\lambda^{-\frac{1}{2}} = U \mathrm{diag}\left(\sqrt{\frac{\vec{d}^{\,2}}{(1-\lambda)\vec{d}^{\,2} + \lambda\nu}}\right) V^T} \tag{15}$$

The central statement in Equation (15) of the whitening transformation in terms of the original singular vectors and modified singular values has both important statistical and computational implications we discuss below.

The whitening transformation retains the original singular vectors and acts only by modifying the sample singular values. From here we can examine the behavior of the transformation as a function of λ (**SI 5.5**), and interpret the effect of λ on the effective number of parameters used in covariance estimation (**SI 5.6**).

2.6 Scalar summaries of the correlation or covariance matrix

Common scalar summary statistics of covariance matrices can be computed efficiently based on the form of the spectral decomposition of $\hat{\Sigma}_\lambda$ (**Table 1**). Standard values like trace, log determinant and condition number can be computed easily from the shrunken eigen-values. In addition the effective degrees of freedom [50], average correlation, average squared correlation [49], effective variance [51], effective number of features [52] are easily computed (**SI 5.7**).

Statistic	Formula	Reference
Trace	$\sum_{i=1}^{p} \tilde{d}_i$	
Log det	$\sum_{i=1}^{p} \log \tilde{d}_i$	
Condition number	$\tilde{d}_1/\tilde{d}_p$	
Average squared correlation	$\sum_{i=1}^{p}(\tilde{d}_i - a)^2/p(p-1)a^2$	[49]
Effective degrees of freedom	$(1-\lambda)p(p-1)/2$	[50]
Average correlation	$(\vec{\mathbf{1}}^T\Sigma_\lambda\vec{\mathbf{1}} - p)/\left(2p(p-1)\right)$	
Effective variance	$\exp\left[\frac{1}{p}\sum_{i=1}^{p}\log\tilde{d}_i\right]$	[51]
Effective number of features	$\vec{\mathbf{1}}^T\Sigma_\lambda^{-1}\vec{\mathbf{1}}$	[52]

Table 1: Computing scalar summary statistics of correlation matrices. Terms are defined in the text except for letting the shrunken eigen-values be $\tilde{d}_i = d_i^2(1-\lambda) + \nu\lambda$, and the mean shrunken eigen-value be $a = \frac{1}{p}\sum_{i=1}^{p}\tilde{d}_i$. Let $\vec{\mathbf{1}}$ be a vector of length p with all entries having value 1.

2.7 Computational methods for data whitening

The computational time for applying the standard whiting transformation, even in the low dimensional setting, can quickly become intractable for moderate p. The time complexity of the standard algorithm is $\mathcal{O}(p^3 + np^2)$ (**SI 5.8**), and becomes very expensive for $p > 100$ and can be intractable for $p > 2000$ features. Yet the entire procedure can be substantially accelerated using properties linear algebra

(**Table 2**). It is not necessary to explicitly compute the sample covariance and whitening matrices and then transform the observed data. Equation (15) rephrases the whitened data in terms of modified singular values, and the original left and right singular vectors. This gives a substantially faster algorithm. In the low dimensional setting of $n > p$, computing the SVD of the observed data is $\mathcal{O}(np^2)$, estimating and applying the ν and λ is negligable, and evaluating the matrix products is $\mathcal{O}(np^2)$. The procedure is thus substantially reduced to $\mathcal{O}(np^2)$ and uses $\mathcal{O}(np)$ memory.

2.7.1 Accelerating computation in the low dimensional setting

Method	**Decomposition**	**Estimate λ**	**Transform**	**Total**	**Memory**
Standard	$p^3 + np^2$	$\geq np$	np^2	$p^3 + np^2$	p^2
GIW-EB ($n > p$)	np^2	n	np^2	np^2	np
GIW-EB ($n < p$)	n^2p	p	n^2p	n^2p	np
GIW-EB, low rank	knp	k	knp	knp	np

Table 2: Computational and memory complexity of whitening methods for n samples and p features using a rank k transformation.

2.7.2 Accelerating computation in the high dimensional setting

The standard formulation of the data whitening transformation does not apply to the high dimensional setting of $n < p$ since the sample covariance matrix is no longer invertible. Some regularization of the singular values is necessary. The probabilistic whitening transformation indeed shrinks the singular values towards ν and addresses the statistical challenge of the high dimensional setting. Yet this computational approach is still quadratic in p, and is infeasible for large p typical in real datasets.

In general, the rank, r, of the sample covariance matrix is bounded above by $\min(n, p)$. In the high dimensional setting, at most only $r \leq n$ singular values are nonzero. Evaluating the SVD is $\mathcal{O}(n^2p)$ in this case only produces the first n singular vectors and values. So there remain $p - n$ uncomputed singular vectors with corresponding singular values of zero. Directly computing these is impractical and, in fact, unnecessary.

Again, the key insight comes from Equation (15). Since all p singular values, including ones with value zero, are modified to be positive it seems that even left and right singular vectors with no variation (i.e. zero singular values) are required to compute the transformation. Yet here we demonstrate that in fact only non-zero singular values and their corresponding singular vectors are required. This insight further reduces the computational complexity of the entire transform.

The substantial improvement in computational complexity for the whitening transformation in the high dimensional settings comes from computing and using only singular vectors corresponding to nonzero singular values. The method is useful beyond this specific context, so here we derive a general

theorem and discuss special cases and applications. The theorem is exact if all nonzero singular values are retained, and is approximate of only a partial set are retained. The derivation generalizes and simplifies the work of Lippert, et al. [6] (**SI 5.9**).

Theorem 1. *Let C be a $p \times p$ symmetric positive semi-definite matrix, with $k \leq p$ positive eigen values. Let the eigen decomposition of C be $V\,diag\,(\vec{s})\,V^T$ where $s_i = 0$ for $i > k$. Let the eigen-vectors be partitioned as $V = [V_1\, V_2]$, and the eigen-values be partitioned as $\vec{s} = [\vec{s}_1\, \vec{s}_2]$ to separate the nonzero and zero components. Let raising a matrix to an exponent α correspond to raising the eigen-values of the matrix to α. Then*

$$A(C + \tau I_p)^\alpha B \;=\; (AV_1)\,diag(s_1^\alpha)V_1^T B + \tau^\alpha \left(AB - (AV_1)V_1^T B\right) \tag{16}$$

where V_1 and $\vec{s}_1$ indicate the first k eigen-vectors and values, respectively, and $A_{a\times p}$ and $B_{p\times b}$ are arbitrary matrices, τ is a positive scalar, and I_p is the p-dimensional identity matrix.

Applying Theorem 1 to the whiting transform, and using the decomposition

$$\hat{\Sigma}_\lambda = V\mathrm{diag}\left[(1-\lambda)\vec{d}^2 + \lambda\nu\right] V^T, \tag{17}$$

restated here from Equation (11), it follows that

$$Y\hat{\Sigma}_\lambda^{-\frac{1}{2}} \;=\; YV\mathrm{diag}\left[(1-\lambda)\vec{d}^2 + \lambda\nu\right]^{-\frac{1}{2}} V^T \tag{18}$$

$$\;=\; (YV_1)\mathrm{diag}\left[(1-\lambda)\vec{d}^2_{1:k} + \lambda\nu\right]^{-\frac{1}{2}} V_1^T + \lambda^{-\frac{1}{2}}\nu^{-\frac{1}{2}}\left(Y - (YV_1)V_1^T\right) \tag{19}$$

where V_1 stores the first k eigen vectors of Σ_λ corresponding to the right singular vectors of $Y/\sqrt{n}$, and $\vec{d}^2_{1:k}$ are the first k eigen values.

Using standard properties of matrix multiplication, the time complexity of evaluating these terms are defined as follows.

Corollary 1.1. *Since the time complexity of multiplying $a \times b$ and $b \times c$ matrices is $\mathcal{O}(abc)$, the time to evaluate the general formulation of Theorem 1 is $\mathcal{O}(abnp)$, once the SVD is already computed. In probabilistic data whitening in Equation (19), $B = I$ and $a = n$ so the time complexity reduces to $\mathcal{O}(knp)$.*

In general, the whitening transformation can be applied in $\mathcal{O}(knp)$ time after the SVD is computed. When the first n singular vectors and values are computed using the standard SVD, then $k = n$, so the entire procedure is $\mathcal{O}(n^2p)$, which is now *linear* in the number of features. This is a dramatic improvement from the naive method that is *cubic* in the number of features (**Table 2**).

2.7.3 Scaling to the ultra-high dimensional setting

While this algorithm is now linear in p, it is still quadratic in the number of samples, n. Yet empirical data can often be well approximated as a low rank matrix [53]. As n and p increase but the dimension of the underlying signal remains fixed, it is not necessary to compute all $\min(n, p)$ singular values and vectors. Instead a rank $k \ll \min(n, p)$ partial SVD can be computed in $\mathcal{O}(knp)$ time to provide a rank k approximation of the observed data [54]. Moreover, the algorithm in Equation (19) can also be evaluated in $\mathcal{O}(knp)$ time given the partial SVD (Table 2). The complexity of this approach is now linear in both n and p, and dependent on the rank k approximation chosen by the analyst.

2.8 Beyond the whitening transformation

While this work focuses on the whitening transformation, applying Theorem 1 with other values for the general exponent α are also useful in practice. In the context of probabilistic whitening transformation with variables defined as in Equation (19), four common operations involving $\hat{\Sigma}_\lambda^\alpha$ can be accelerated:

- $\alpha = 1$ gives the posterior expected covariance matrix.
- $\alpha = -1$ gives the corresponding precision matrix.
- $\alpha = -\frac{1}{2}$ gives the whitening transformation.
- $\alpha = \frac{1}{2}$ gives the 'coloring matrix' so that post-multiplying by $\hat{\Sigma}_\lambda^{\frac{1}{2}}$ produces a correlation of $\hat{\Sigma}_\lambda$ between columns.

3 Results

3.1 Performance on simulated data

The formulation of the probabilistic whitening transformation provides two advances of practical interest in analyses of large-scale datasets. First, the implicit covariance approach reduces computational time for high-dimensional data from the standard cubic time method that creates and then inverts the covariance matrix. For example, in simulations with $n = 5000$ and an increasing number of features the GIW-EB methods runs in < 2 min and the low rank GIW-EB with k=50 runs in < 1 min, when the standard method becomes intractable both in terms of time and memory (**Figure 2A**). Second, the probabilistic model motivates the empirical Bayes estimation of the shrinkage parameter which only uses the sample singular values of the data. Due to its probabilistic formulation, the whitening high-dimensional data using the GIW-EB method can produce lower out-of-sample error than other methods. For example, in a simulation Scenario 3 (described below), both full and low rank GIW-EB give estimated values of λ that approach the optimal value while other methods give λ values that are too small and have higher error (**Figure 2B**).

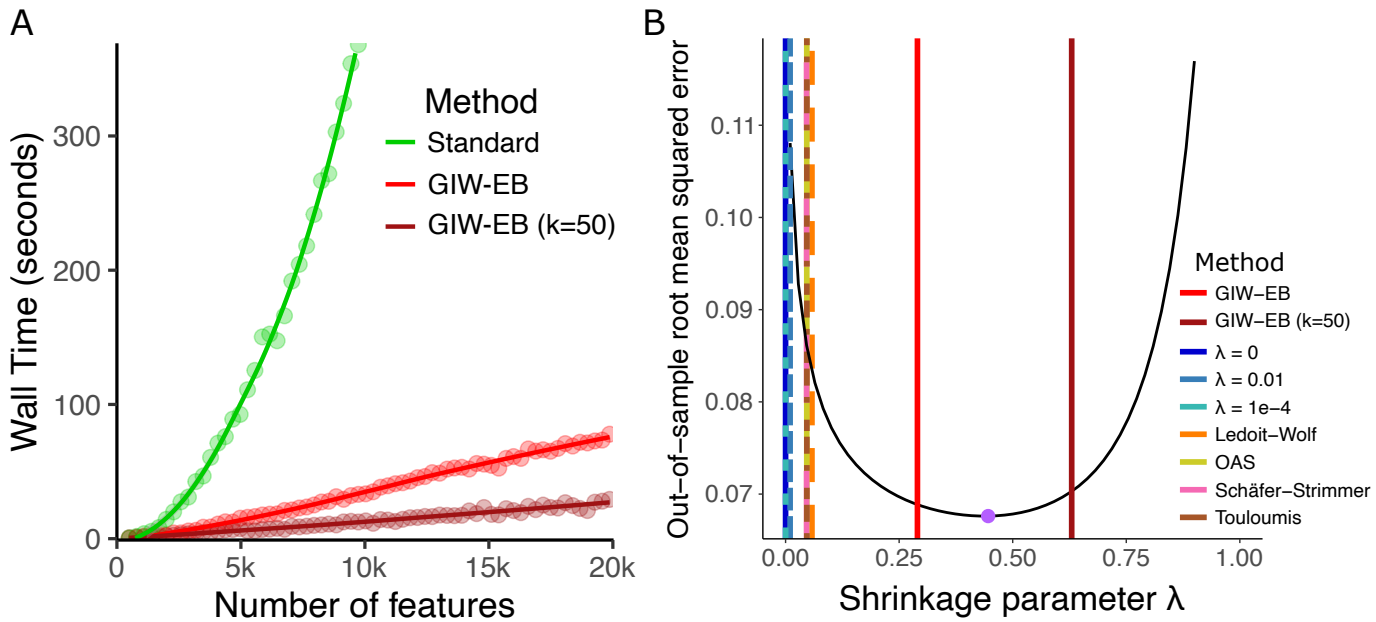

Figure 2: Performance of probabilistic whitening transformation. **A**) Run times are shown for standard (green), GIW-EB (red) and low rank GIW-EB with $k = 50$ (dark red) whitening transformations. Results shown for $n = 5000$ samples and an increasing number of features. Points indicate observed wall time and lines show loess smooth. **B**) The empirical out-of-sample root mean squared error for a simulated dataset across the range of λ values is shown by the black curve. The λ value giving the lowest error is show by the purple point. Values of λ estimated by other methods are shown by vertical lines.

Since an exact whitening transformation produces an identity covariance matrix when applied to the data it was estimated from, the accuracy of an approximate whitening transformation can be evaluated by estimating it from a subset of the data and computing the empirical covariance after applying the transform to the rest of the dataset. This out-of-sample accuracy of the whitening transformation is captured by the root mean squared error of the difference between the empirical covariance of the transformed data and the identity matrix.

We consider simulations with $n = 500$ and the number of features p increasing from 20 to 2000, with correlation between features in each of 4 scenarios:

1) constant correlation of $\rho = 0.6$
2) auto-correlation with $\rho = 0.95$
3) 4 blocks of constant correlation of $\rho = 0.6$
4) 4 blocks of auto-correlation with $\rho = 0.95$

In Scenario 3, most methods show equivalent performance for small values of p, error increases as p approaches n, and then error decreases and converges to the optimal root mean squared error (**Figure 3A**). The 'Oracle' method, shown in black here, transforms the data using the true covariance, so

it gives a lower bound on the error. Using a whitening transformation without shrinkage by setting $\lambda = 0$ performs poorly even for small p and error only increases with p. Using the pseudoinverse improves performance when $p > n$, but does not compete with shrinkage methods. Examining different approaches to estimate the shrinkage parameter, the GIW-EB method gives the smallest error across all values of p and the low rank GIW-EB with k = 50 outperforms other methods for moderate p. Both full and low rank GIW-EB give errors that decrease with p and avoids the increase in error at $p \approx n$ seen with other shrinkage methods.

Overall, the GIW-EB method consistently gives the lowest out-of-sample root mean square error across a range of p values in the 4 simulation scenarios (**Figure S1-4**). The low rank GIW-EB method is competitive in Scenarios 1 and 3, were the true covariance is low rank.

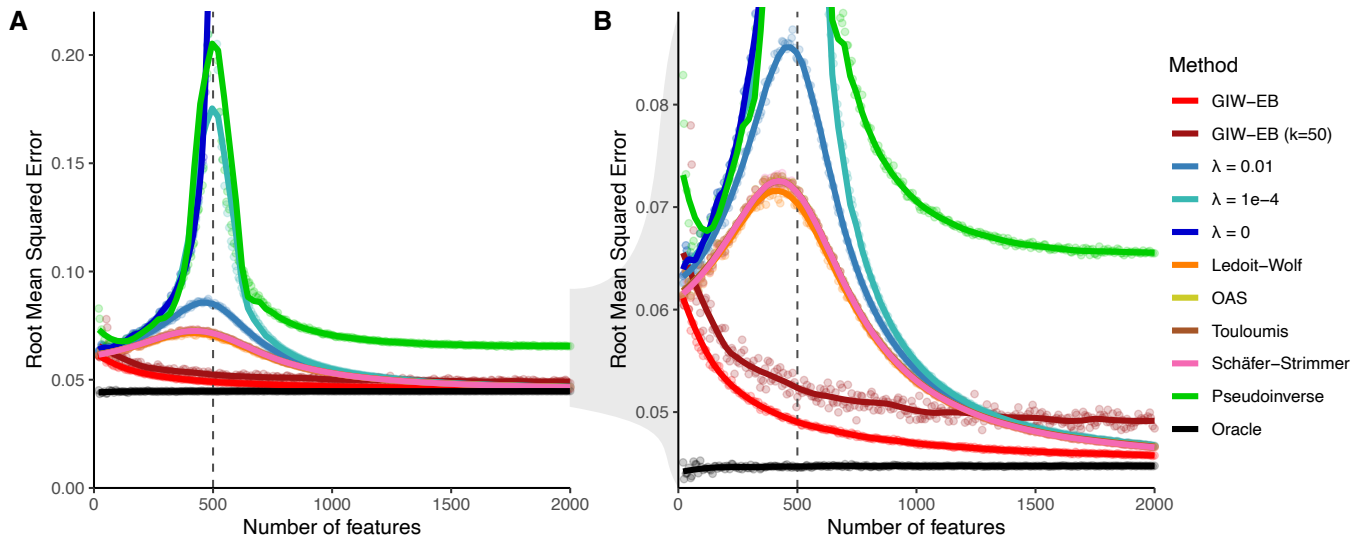

Figure 3: Out-of-sample performance of whitening transformation. Simulated data for $n = 500$ sample are drawn from a multivariate normal distribution with known covariance matrix and 11 whitening transformations are then applied to 500 additional samples from the same distribution. Of these, 9 methods either estimate the shrinkage parameter λ from the data or use a fixed value. Setting $\lambda = 0$ corresponds to no shrinkage, and pseudoinverse does not perform any shrinkage. The oracle method performs a whitening transformation using the true correlation matrix from which the data are simulated, and so gives the best out-of-sample performance. Performance is evaluated by comparing the correlation matrix of the transformed data to the identity matrix, and computing the root mean squared error (rMSE). **A**) rMSE of the whitening transformation for an increasing number of features **B**) Zoom-in on y-axis from (**A**). Points indicate observed values and lines show loess smooth.

3.2 Application to high-dimensional gene expression data

Since covariance estimation and data whitening play a central role in many standard analyses, we evaluate the out-of-sample performance of these whitening transformations. We applied the whitening

transformations to RNA-seq data from up to 800 individuals and ∼20K genes in each of 37 tissues from the Genotype Tissue Expression (GTEx) project using $\log_2$ transcripts per million [39]. The whitening transformation is estimated from 90% of the individuals in each tissue, and performance is evaluated on the remaining individuals. Since the number of individuals and genes varies across each tissue, whitening performance is evaluated by dividing the root mean squared error (rMSE) of the empirical covariance matrix after transformation by the baseline of the rMSE of the covariance matrix without transformation. Most methods performed well in terms of normalized rMSE (**Figure 4A**, **S5**), with the exception of $\lambda = 0$ and pseudoinverse. The GIW-EB methods and methods with fixed λ values use the implicit covariance algorithm, and these give the best computational performance with a wall time of ∼30 seconds compared to 7-30 hours for the other methods (**Figure 4B**).

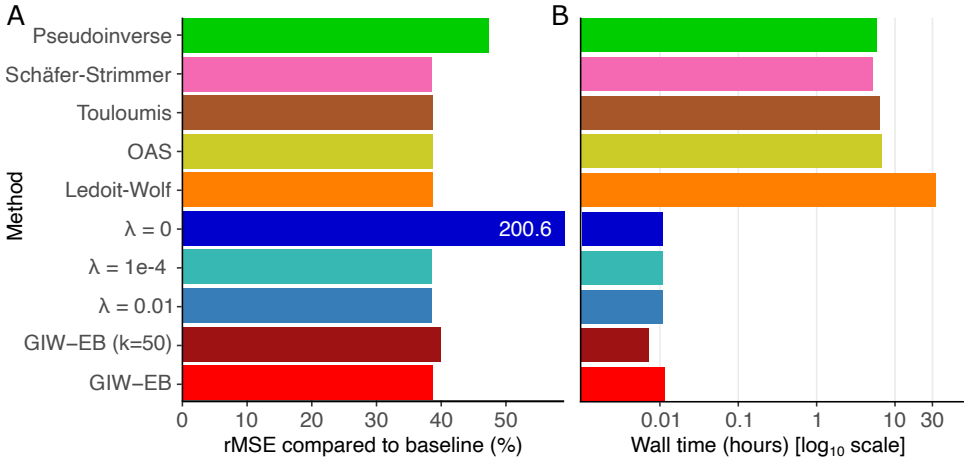

Figure 4: Performance of whitening transformations applied to gene expression data from GTEx. **A**) Root mean squared error (rMSE) compared to baseline for 209 individuals and 17,699 genes from the frontal cortex of the brain. **B**) Compute time required for each method. Methods using GIW-EB or fixed λ values use the implicit covariance algorithm.

3.3 Performance on genetic data from 1000 Genomes Project

Since many analyses of summary statistics from genome-wide association studies apply a whitening transformation estimated from an external reference panel, we evaluated the out-of-sample performance of these whitening transformations on genetic data from the 1000 Genomes Project [40]. Individuals of European, Asian and African ancestry were analyzed separately, and the whitening transformation was performed on independent LD blocks including varying number of genetic variants defined in each population [55]. The whitening transform was trained on half of the samples and evaluated on the other half. Despite the genetic data not being multivariate normal, shrinkage methods performed well across most windows in European individuals (**Figure 5A**). The GIW-EB matches or is competitive with other methods that estimate λ from the data. The low rank GIW-EB method with $k = 50$ is most competitive for windows with $< 1,500$ variants. The GIW-EB methods and methods with fixed λ values use the implicit covariance algorithm, and these give the best computational performance with a wall time of ∼5 min compared to 3-30 hours for the other methods (**Figure 5B**). Error profiles and computational time was similar for individuals of African and Asian ancestry (**Figure S6**).

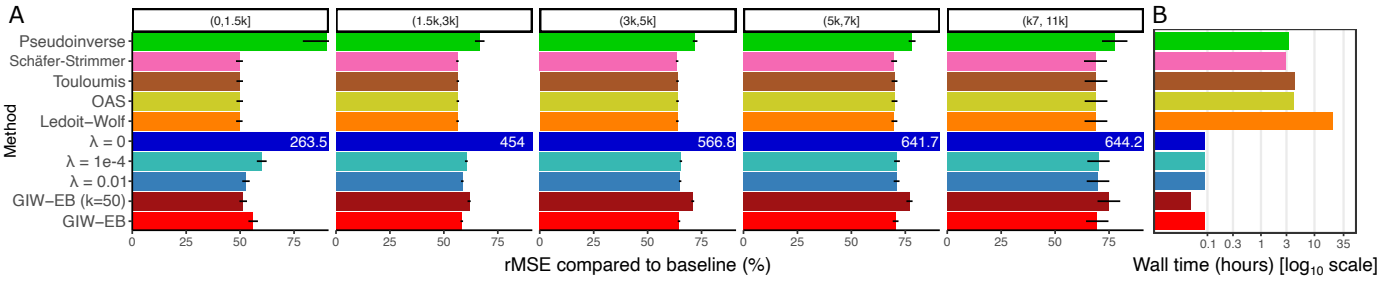

Figure 5: Out-of-sample whitening performance on genetic data. **A**) Root mean squared error (rMSE) compared to baseline for European individuals stratified by the number of genetic variants in the LD window. Values indicate the values averaged across windows, and bars indicate the 95% confidence interval. When normalized rMSE values exceed the axes, the value is shown in white text. **B**) Wall time for each method shown on a $\log_{10}$ scale. Methods using GIW-EB or fixed λ values use the implicit covariance algorithm.

3.4 Application to imputing GWAS summary statistics

Genome-wide associations studies (GWAS) release summary statistics from their data corresponding to z-statistics from the set of genetic variants analyzed in the study. Since the number of genetic variants analyzed in the original study can vary widely, there has been interest in imputing z-statistics of additional variants observed in an external reference panel. Since the correlation between z-statistics of two genetic variants matches the linkage disequilibrium between the two variants, Pasaniuc, et al. [16] proposed a method to impute z-statistics from unobserved variants using the set of observed z-statistics and the correlation structure between variants in a genomic region.

Imputing z-statistics from unobserved variants can be stated in terms of out-of-sample whitening (**SI 5.10**), and we evaluated the performance of different whitening methods on imputation accuracy. We used summary statistics for 5.6M variants from schizophrenia GWAS of 58K cases and 77K controls of European ancestry [41]. Z-statistics for 15M variants were imputed using a reference panel of 407 unrelated individuals of European ancestry from the 1000 Genomes Project [40] using a sliding window of 1 Mb and a flanking region of 250 kb across all autosomes. Accuracy was evaluated by comparing the imputed and observed z-statistic for 460K randomly selected variants with minor allele frequency $\geq 5\%$. GIW-EB gave high imputation accuracy across a broad range of z-statistic values (**Figure 6A**), and give smallest genome-wide root mean squared error (**Figure 6B**). Imputation error increases for small minor allele frequencies (MAF), and GIW-EB gives the smallest squared error across the range of MAF values (**Figure 6C**). Given 407 samples and a mean of $\sim$ 3100 variants in each genome window, the low rank GIW-EB with $k = 50$ was not a good approximation in this case. The GIW-EB methods and methods with fixed λ values use the implicit covariance algorithm, and these give the best computational performance and require < 1 hr for genome-wide analysis for this reference panel compared to 4-126 hours for other methods (**Figure 6D**). Overall, GIW-EB method is the most accurate and fastest approach to impute GWAS summary statistics.

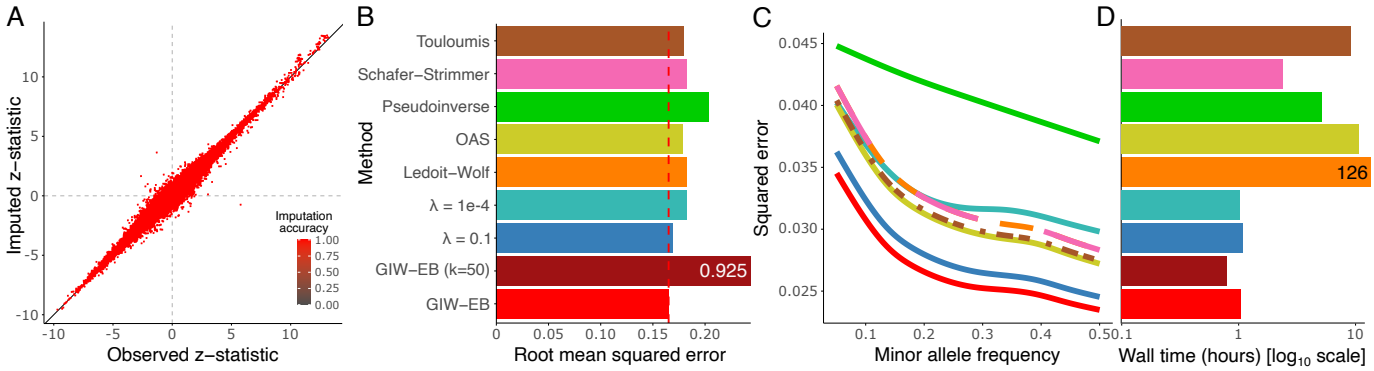

Figure 6: Imputation GWAS summary statistics. **A**) Imputed z-statistics plotted against the observed z-statistics for 461,707 SNPs. **B**) Root mean squared prediction error for each method. Red vertical dashed line indicates the value for GIW-EB. **C**) Square prediction error as a function of minor allele frequency. Smooth curves are fit by a generalized additive model that is the default in `ggplot2::geom_smooth()`. GIW-EB (k=50) was excluded from this plot due to high imputation error. **D**) Compute time required for each method. When values exceed the axes, values are shown in text. Methods using GIW-EB or fixed λ values use the implicit covariance algorithm.

4 Discussion

The whitening transformation is a central, if often unstated, component of many widely used statistical methods. While many standard statistical methods have been reformulated with a probabilistic model, or extended to handle regularization and modern high dimensional datasets, the whitening transform has received less attention. The standard whitening transformation is limited by computational time, memory usage, and gives poor out-of-sample performance on modestly sized datasets. Regularizing the covariance estimate improves statistical performance, but is still very computationally demanding for high dimensional datasets.

We proposed a regularized probabilistic whitening transformation and examined its statistical properties based on the singular value decomposition of the data matrix. The particular form of the probabilistic formulation enables an implicit covariance approach that avoids direct estimation of the covariance matrix and reduces the computational time from cubic to linear in the number of features. This can reduce computational time by a factor of ≥ 100 in practical applications.

In simulations and applications high-dimensional datasets of gene expression from the GTEx project and genotype data from the 1000 Genomes Project, we demonstrated that our GIW-EB method matches or exceeds other methods in out-of-sample whitening benchmarks. In an application to impute z-statistics from a schizophrenia GWAS, GIW-EB gave the best imputation performance. As the scale of genomic datasets continues to increase, the implicit covariance approach of the probabilistic model, enables application of whitening transforms on high dimensional data where existing methods are intractable. Future applications include single cell gene expression and larger genetic references panels.

Analysis code

Code implementing the probabilistic whitening transformation with implicit covariance is available in the R package `decorrelate` from `https://cran.r-project.org/package=decorrelate`. Documentation is available at `https://gabrielhoffman.github.io/decorrelate`. Code implementing the imputation of summary statistics is available in the R package `imputez` available at `https://gabrielhoffman.github.io/imputez`. Code for simulation and analysis is available at `https://github.com/GabrielHoffman/decorrelate_analysis`. Times were evaluated using 12 cores on a 48 core Intel Xeon Platinum 8462Y CPU @ 2.8GHz running R v4.3.3 linked to a parallelized Intel Math Kernel Library for linear algebra operations.

Competing interests: The authors declare no competing interest.

Acknowledgements: We thank Bernie Devlin for valuable feedback.

Funding: This work was supported by NIA and NINDS grant U24AG087563

References

[1] Kessy, A., Lewin, A., and Strimmer, K. (2018). Optimal Whitening and Decorrelation. American Statistician *72*, 309–314.

[2] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. Springer Series in Statistics. (New York, NY: Springer) 2nd edition.

[3] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. (Cambridge, Massachusetts: MIT Press).

[4] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. (New York: Springer Science).

[5] Hoffman, G. E. (2013). Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. PLoS ONE *8*, e75707.

[6] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nature Methods *8*, 833–5.

[7] Krizhevsky, A. and Hinton, G. (2009). Learning Multiple Layers of Features from Tiny Images. Technical Report, University of Toronto.

[8] Wadia, N. S., Duckworth, D., Schoenholz, S. S., Dyer, E., and Sohl-Dickstein, J. (2021). Whitening and second order optimization both make information in the dataset unusable during training, and can reduce or prevent generalization. In Proceedings of the 38th International Conference on Machine Learning pp. 10617–10629.

[9] Zhang, S., Nezhadarya, E., Fashandi, H., Liu, J., Graham, D., and Shah, M. (2021). Stochastic Whitening Batch Normalization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10973–10982.

[10] Pasaniuc, B. and Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics *18*, 117–127.

[11] Vilhjalmsson, B. J., Yang, J., Kenny, E. E., Schierup, M. H., De Jager, P., Patsopoulos, N. A., McCarroll, S., Daly, M., Purcell, S., Chasman, D., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. American Journal of Human Genetics *97*, 576–592.

[12] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature Communications *10*, 1776.

[13] Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the "Sum of Single Effects" model. PLoS Genetics *18*, 1–24.

[14] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.

[15] Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501.

[16] Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics (Oxford, England) *30*, 2906–2914.

[17] Chen, W., Wu, Y., Zheng, Z., Qi, T., Visscher, P. M., Zhu, Z., and Yang, J. (2021). Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. Nature Communications *12*, 1–10.

[18] Davis, J. R., Fresard, L., Knowles, D. A., Pala, M., Bustamante, C. D., Battle, A., and Montgomery, S. B. (2016). An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. American Journal of Human Genetics *98*, 216–224.

[19] Reshef, Y. A., Finucane, H. K., Kelley, D. R., Gusev, A., Kotliar, D., Ulirsch, J. C., Hormozdiari, F., Nasser, J., O'Connor, L., van de Geijn, B., et al. (2018). Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. Nature Genetics *50*, 1483–1493.

[20] He, Z., Liu, L., Belloy, M. E., Le Guen, Y., Sossin, A., Liu, X., Qi, X., Ma, S., Gyawali, P. K., Wyss-Coray, T., et al. (2022). GhostKnockoff inference empowers identification of putative causal variants in genome-wide association studies. Nature Communications *13*, 7209.

[21] Benner, C., Havulinna, A. S., Järvelin, M. R., Salomaa, V., Ripatti, S., and Pirinen, M. (2017). Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. American Journal of Human Genetics *101*, 539–551.

[22] Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. Statistical Science *26*, 369–387.

[23] Pourahmadi, M. (2013). High-Dimensional Covariance Estimation. Wiley Series in Probability and Statistics. (Hoboken, NJ, USA: John Wiley and Sons).

[24] Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. The Econometrics Journal *19*, C1–C32.

[25] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis *88*, 365–411.

[26] Schäfer, J. and Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology *4*.

[27] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics (Oxford, England) *9*, 432–41.

[28] Won, J.-h., Lim, J., Kim, S.-j., and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *75*, 427–450.

[29] Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society. Series B: Statistical Methodology *75*, 603–680.

[30] Huang, N. and Fryzlewicz, P. (2019). NOVELIST estimator of large correlation and covariance matrices and their inverses. Test *28*, 694–727.

[31] Fisher, T. J. and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Computational Statistics and Data Analysis *55*, 1909–1918.

[32] Chen, C.-F. (1979). Bayesian Inference for a Normal Dispersion Matrix and its Application to Stochastic Multiple Regression Analysis. Journal of the Royal Statistical Society: Series B (Methodological) *41*, 235–248.

[33] Leday, G. G. and Richardson, S. (2019). Fast Bayesian inference in large Gaussian graphical models. Biometrics *75*, 1288–1298.

[34] Hannart, A. and Naveau, P. (2014). Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework. Journal of Multivariate Analysis *131*, 149–162.

[35] Van Wieringen, W. N. and Peeters, C. F. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. Computational Statistics and Data Analysis *103*, 284–303.

[36] Kuismin, M. and Sillanpää, M. J. (2016). Use of Wishart Prior and Simple Extensions for Sparse Precision Matrix Estimation. PLOS ONE *11*, e0148171.

[37] Roweis, S. (1998). EM algorithms for PCA and SPCA. Advances in Neural Information Processing Systems pp. 626–632.

[38] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society. Series B: Statistical Methodology *61*, 611–622.

[39] Strober, B. J., Wen, X., Wucher, V., Kwong, A., Lappalainen, T., Li, X., and Liang, Y. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *18*, 1318–1330.

[40] 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

[41] Trubetskoy, V., Pardinas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C. Y., Dennison, C. A., Hall, L. S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature *604*, 502–508.

[42] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2013). Bayesian Data Analysis. (CRC Press).

[43] Touloumis, A. (2015). Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. Computational Statistics and Data Analysis Data Analysis *83*, 251–261.

[44] Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O. (2010). Shrinkage algorithms for MMSE covariance estimation. IEEE Transactions on Signal Processing *58*, 5016–5029.

[45] Boileau, P., Hejazi, N. S., van der Laan, M. J., and Dudoit, S. (2023). Cross-Validated Loss-based Covariance Matrix Estimator Selection in High Dimensions. Journal of Computational and Graphical Statistics *32*, 601–612.

[46] Tong, J., Hu, R., Xi, J., Xiao, Z., Guo, Q., and Yu, Y. (2018). Linear shrinkage estimation of covariance matrices using low-complexity cross-validation. Signal Processing *148*, 223–233.

[47] Theiler, J. (2012). The incredible shrinking covariance estimator. In Proc. SPIE 8391, Automatic Target Recognition XXII pp. 83910P.

[48] Hoffbeck, J. P. and Landgrebe, D. A. (1996). Covariance matrix estimation and classification with limited training data. IEEE Transactions on Pattern Analysis and Machine Intelligence *18*, 763–767.

[49] Watanabe, J. (2022). Statistics of eigenvalue dispersion indices: Quantifying the magnitude of phenotypic integration. Evolution *76*, 4–28.

[50] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association *93*, 120–131.

[51] Peña, D. and Rodríguez, J. (2003). Descriptive measures of multivariate scatter and linear dependence. Journal of Multivariate Analysis *85*, 361–374.

[52] Liu, G. and Liang, K. Y. (1997). Sample size calculations for studies with correlated observations. Biometrics *53*, 937–947.

[53] Udell, M. and Townsend, A. (2019). Why Are Big Data Matrices Approximately Low Rank? SIAM Journal on Mathematics of Data Science *1*, 144–160.

[54] Baglama, J. and Reichel, L. (2005). Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. SIAM Journal on Scientific Computing *27*, 19–42.

[55] Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.

[56] Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. Journal of the American Statistical Association *103*, 340–349.

[57] Donoho, D., Gavish, M., and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. The Annals of Statistics *46*, 1742–1778.

5 Supplementary Information

5.1 Defining the whitening matrix

Consider data matrix $Y_{n\times p}$ covariance between columns given by $\Sigma_{p\times p}$ and each column having zero mean after centering. Defining the singular value decomposition $Y/\sqrt{n}$ as UDV^T, then

$$\begin{aligned} \mathrm{cov}(Y) &= n^{-1}Y^TY && (1)\\ &= VDU^TUDV^T && (2)\\ &= VD^2V^T. && (3) \end{aligned}$$

Setting $\Sigma^{-\frac{1}{2}} = VD^{-1}V^T$, and evaluating the covariance between columns following the whitening transform gives identity covariance according to

$$\begin{aligned} \mathrm{cov}(Y\Sigma^{-\frac{1}{2}}) &= \Sigma^{-\frac{1}{2}}\mathrm{cov}(Y)\Sigma^{-\frac{1}{2}} && (4)\\ &= \left(VD^{-1}V^T\right)\left(VD^2V^T\right)\left(VD^{-1}V^T\right) && (5)\\ &= VD^{-1}D^2D^{-1}V^T && (6)\\ &= VV^T && (7)\\ &= I_p. && (8) \end{aligned}$$

5.2 Evaluating whitening performance

Let $\hat{A}$ be the whitening transformation learned from the observed data so that

$$Y\hat{A} \sim \left[\vec{\mu}\,\hat{A}, I_n, \hat{A}^T\Sigma\hat{A}\right]. \tag{9}$$

When $\hat{A}^T\Sigma\hat{A} = I_p$, as in Equation (2), this transformation produces data $Y\hat{A}$ whose columns are *exactly* uncorrelated when $\hat{A}$ is estimated from Y. Yet this condition can only be satisfied when the sample covariance matrix, $\hat{\Sigma}$, is invertible, and this requires $n > p$. In high dimensional data with $n < p$, so some form of regularization is needed. An exact whitening transformation no longer exists in these cases, so we want to learn $\hat{A}$ from the data so that $\hat{A}^T\Sigma\hat{A}$ is 'close' to the identity matrix by some metric. Formally, we could seek a transform so that $\|\hat{A}^T\Sigma\hat{A} - I_p\|_F$ is small, where $\|.\|_F$ is the Frobenius norm of a matrix. This is closely related to the root mean squared error of the difference between the covariance of the transformed data and the identity matrix.

The Frobenius norm of a matrix $M_{p\times p}$ is defined as

$$\|M\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} M_{i,j}^2} \tag{10}$$

so that scaling by p gives

$$\|M\|_F/p = \sqrt{\left(\sum_{i=1}^{p}\sum_{j=1}^{p} M_{i,j}^2\right)/p^2}, \tag{11}$$

which is the root mean squared error when $M = \hat{A}^T\Sigma\hat{A} - I_p$.

5.3 Regularizing the correlation matrix and retaining sample variances

The approach described so far shrinks all diagonal variance entries towards a single value ν. While this is reasonable when the variances are all equal, most covariance matrices in practice have heterogeneous variances. Instead, we can factor out the variance terms, shrink the correlation matrix and then rescale by the variances. Let $\hat{\Sigma}_{MLE}$ have variance $\hat{\sigma}_i^2$ in the ith entry on the diagonal, let Z be a diagonal matrix with $\hat{\sigma}_i$ in the ith entry and let $\hat{C}_{MLE}$ be the corresponding correlation matrix estimate so that the covariance can be decomposed according to $\hat{\Sigma}_{MLE} = Z\hat{C}_{MLE}Z$. The regularized covariance matrix retaining the sample variances along the diagonal is then

$$\hat{\Sigma}_\lambda = Z\left[(1-\lambda)\hat{C}_{MLE} + \lambda I\right]Z. \tag{12}$$

The approach described in the main text corresponds to setting $Z = I$ and estimating ν from the data. Alternatively, a second approach corresponds to estimating the standard deviation matrix, Z, from the data and setting $\nu = 1$. We omit the variance terms from the notation throughout for the sake of simplicity.

5.4 Estimation of shrinkage parameter

Setting the value of the shrinkage parameter λ is a key component for any convex combination method and many methods have been proposed under different, often nonparametric, assumptions [25, 26, 34, 35, 43, 56]. Following the probabilistic formulation, we take advantage of the Gaussian Inverse Wishart model and use an empirical Bayes approach to estimate the shrinkage parameter after marginalizing out the covariance. The probability of the observed data can be marginalized with respect to the unknown covariance term Σ according to

$$p(Y|\delta,\nu) = \int p(Y|\Sigma,\delta,\nu)p(\Sigma|\delta,\nu)d\Sigma \tag{13}$$

$$= \pi^{-np/2}\frac{\Gamma_p\left(\frac{\delta+n}{2}\right)}{\Gamma_p\left(\frac{\delta}{2}\right)}\frac{|(\delta-p-1)T|^{\frac{\delta}{2}}}{|(\delta-p-1)T+S|^{\frac{\delta+n}{2}}} \tag{14}$$

$$= \pi^{-np/2}\frac{\Gamma_p\left(\frac{\delta+n}{2}\right)}{\Gamma_p\left(\frac{\delta}{2}\right)}\frac{|(\delta-p-1)\nu I|^{\frac{\delta}{2}}}{|(\delta-p-1)\nu I+S|^{\frac{\delta+n}{2}}}. \tag{15}$$

where $S = YY^T$, and we focus on the case where the target matrix is a scaled identity matrix so that $T = \nu I$.

The empirical Bayes estimates of δ and ν can be obtained from the observed data according to

$$\hat{\delta}, \hat{\nu} = \underset{\delta,\nu}{\operatorname{argmax}} \ \ell(\delta, \nu|Y). \tag{16}$$

where

$$\begin{aligned}
\ell(\delta, \nu|Y) &= \log p(Y|\delta, \nu) && (17)\\
&= -\frac{np}{2}\log\pi + \log\Gamma_p\left(\frac{\delta+n}{2}\right) - \log\Gamma_p\left(\frac{\delta}{2}\right) && (18)\\
&\quad + \frac{\delta p}{2}\log\left(\nu(\delta-p-1)\right) && (19)\\
&\quad - \frac{\delta+n}{2}\sum_i \log\left(\nu(\delta-p-1)+d_i^2\right) && (20)
\end{aligned}$$

where d_i^2 are the eigen values of S. Since $\hat{\nu}$ can be estimated in closed form from the data as the mean diagonals of S, and $\hat{\delta}$ can be obtained subsequently using a line search, $\hat{\lambda}$ can then be obtained from $\hat{\delta}$. Importantly, this function has one parameter and the data enters only through the eigen-values d^2 and the prior scale, ν.

5.4.1 Low rank approximation

With the full spectral decomposition, the estimate of the correlation matrix

$$\hat{\Sigma}_\lambda = (1-\lambda)\hat{\Sigma}_{MLE} + \lambda\nu I \tag{21}$$

has diagonal values equal to 1. But when the data is approximated by a truncated singular value decomposition, $\hat{\Sigma}_\lambda$ no longer has diagonals of value 1 when simply setting $\nu = 1$. Instead, we set the value of ν so that the mean of the diagonals of $\hat{\Sigma}_\lambda$ is closest to 1.

Let the target sum of the diagonal values of $\hat{\Sigma}_\lambda$ be p, and let

$$a = \sum_i \left[U_{1:k}\mathrm{diag}(d_{1:k}^2)U_{1:k}^T\right]_{i,i} \tag{22}$$

be the sum of the diagonals of the low rank approximation to the correlation matrix, $\hat{\Sigma}_{MLE}$. Therefore, the sum of the diagonals of $\hat{\Sigma}_\lambda$ is $(1-\lambda)a + \lambda\nu p$.

Setting this value equal to the desired sum of the diagonals, p, and solving for ν gives

$$\hat{\nu} = \left(p - (1-\lambda)a\right)/\lambda p. \tag{23}$$

This value is estimated *within* each evaluation of $\ell(.)$ so that the function has just 1 parameter.

5.5 Properties as a function of λ

Here we can examine the behavior of the transformation as a function of λ. First, all *modified* singular values are positive, even in cases where the corresponding singular value of the observed data is zero. This is intuitive since all eigen-values of the implied covariance matrix $\hat{\Sigma}_\lambda$ must be positive for it to be invertible.

Based on results above, setting $\lambda = 0$ corresponds to the algorithmically defined whitening transformation using the sample covariance matrix where there is no influence from the prior. Indeed, Equation (15) shows that this corresponds to setting all singular values of the transformed data to 1, thus scaling the sample variances in all dimensions to be equal. At the other extreme, setting $\lambda = 1$ divides each singular value by $\sqrt{\nu}$ so the singular values of the transformed data are just scaled versions of the original values, and thus retains the same correlation structure. For intermediate values of λ that are used in practice, they are shrunk using ν as a 'pivot': values greater then ν are shrunk towards the pivot and values less than ν are inflated towards the pivot.

As $\lambda \to 1$ the prior is increasingly concentrated around the target matrix and recovers the target matrix in the limit. As $\lambda \to 0$, this corresponds to an infinitely diffuse prior and recovers the sample covariance matrix in the limit. As n increases for fixed p, a good estimator of λ should converge to 0 so the posterior expectation is an asymptotically consistent estimator of the covariance [33].

Mapping from the observed to the modified singular values depends on λ, which in turn depends on n and p, and $\vec{d}$. Small sample sizes require stronger regularization with larger λ values. In this case, the modified singular values retain some level of variation so the whitening transform is approximate. With increasing sample size, the effect of the prior decreases so that $\lambda \to 0$ as $n \to \infty$ with fixed p, so the modified singular values approach the observed singular values. At the same time, $\hat{\Sigma}_\lambda \to \hat{\Sigma}_{MLE}$ and $\text{cov}\left(Y\hat{\Sigma}_\lambda^{-\frac{1}{2}}\right) \to I$ so that the whiting transform is asymptotically exact, and approximate for finite sample size.

The out-of-sample performance of the whitening transform has received less attention. For small sample size, the whiting transform estimated from the sample covariance can give very poor out-of-sample performance and can in fact increase the correlation between features. Regularization produces better estimates of the true singular values [57] and we demonstrate that it can also substantially improve out-of-sample whitening performance.

5.6 Interpretation of shrinkage parameter

The sample covariance $\hat{\Sigma}_{MLE}$ requires $p(p-1)/2$ off-diagonal covariance parameters. Since the posterior expectation is a convex combination between prior and sample covariance, we can compute the effective number of parameters of the posterior estimate, $\hat{\Sigma}_\lambda$, as a function of the λ shrinkage parameter.

Following Ye [50], the effective number of parameters is determined by the degree to which changes in

the observed data change the estimator. Letting $f_\lambda(\hat{\Sigma}_{MLE}) = (1-\lambda)\hat{\Sigma}_{MLE} + \lambda\nu I$ be the mapping from the sample covariance to the posterior covariance, Ye [50] defines the effective number of parameters as

$$\sum_{i,j} \frac{\partial f_\lambda(\hat{\Sigma}_{MLE})}{\partial \hat{\Sigma}_{MLEi,j}}, \tag{24}$$

which is interpretable as the sum of the degree to which changes in the sample covariance change the posterior estimate. Considering only the off-diagonal covariance terms,

$$\frac{\partial f_\lambda(\hat{\Sigma}_{MLE})}{\partial \hat{\Sigma}_{i\neq j}} = 1 - \lambda, \tag{25}$$

so each covariance term uses $1-\lambda$ effective parameters. Therefore the effective number of parameters used to estimate the covariance terms is $(1-\lambda)p(p-1)/2$. Compared to the sample covariance, the effective number of parameters is thus reduced by a factor of $1-\lambda$.

As the number of features, p, increases with fixed sample size n, $\hat{\lambda} \to 1$ to apply stronger shrinkage [33]. While the effective number of parameters for the sample covariance matrix grows quadratically with p, it grows slower than quadratically for the shrinkage model since λ increases with p for fixed n.

5.7 Scalar summaries of the correlation or covariance matrix

The condition number of $\hat{\Sigma}_\lambda$ is defined as the ratio between the largest and smallest eigen-values:

$$\kappa_\lambda = \frac{(1-\lambda)d_1^2 + \lambda\nu}{(1-\lambda)d_{min}^2 + \lambda\nu}. \tag{26}$$

When the sample covariance is low rank, the smallest sample eigen-value is zero so the condition number has the simple form

$$\kappa_\lambda = \frac{(1-\lambda)}{\lambda\nu} d_1^2 + 1. \tag{27}$$

as long as $\lambda > 0$. Instead of regularizing the condition number using an ad hoc tuning parameter, or estimating it by cross validation, in the probabilistic model λ is a function of n, p and $\vec{d}$. As λ increases, more weight is put on the target matrix, and the condition number approaches its minimum value of 1.

The effective variance is defined as the geometric mean of the eigen-values and satisfies a number of properties for a metric of multivariate variability [51]. This is defined as $|\Sigma_\lambda|^{1/p}$, but can also be computed directly from the eigen-values. The effective number of features is analogous to the effective sample size based on the sum of all elements in the inverse correlation matrix [52] which is defined as $\sum_{i,j}[\Sigma_\lambda^{-1}]_{i,j}$. The average correlation can be computed by evaluating the sum of the off-diagonal entries of the correlation matrix divided by the number of entries. The effective number of features and average correlation can be computed efficiently using the implicit correlation matrix as described below. These values can be computed for the Σ_{MLE} matrix by setting $\lambda = 0$.

5.8 Scaling of standard whitening transform

Constructing $\hat{\Sigma}_{MLE}$ is $\mathcal{O}(np^2)$, requires $\mathcal{O}(p^2)$ memory, and evaluating its eigen decomposition is $\mathcal{O}(p^3)$. Finally, explicitly creating the whitening matrix is $\mathcal{O}(p^2)$, and evaluating the matrix product is $\mathcal{O}(np^2)$. Estimating λ with the probabilistic model is $\mathcal{O}(p)$, but other methods look at the entire dataset so require at least $\mathcal{O}(np)$ time. Estimating ν is negligible in comparison, and applying shrinkage to the standard whiting transformation is also negligible. This allows us to consider the standard whitening transformation as a special case of the probabilistic model with $\lambda = 0$. The time complexity of the entire procedure is therefore $\mathcal{O}(p^3 + np^2)$. The standard algorithm becomes very expensive for $p > 100$ and can be intractable for $p > 2000$ features.

5.9 Accelerating transform using properties of linear algebra

Theorem 1. *Let C be a $p \times p$ symmetric positive semi-definite matrix, with $k \leq p$ positive eigen values. Let the eigen decoposition of C be $U\,diag(s)U^T$ where $s_i = 0$ for $i > k$. Let the eigen-vectors be partitioned as $U = [U_1\, U_2]$, and the eigen-values be partitioned as $s = [s_1\, s_2]$ to separate the nonzero and zero components. Let raising a matrix to an exponent α correspond to raising the eigen-values of the matrix to α. Then*

$$A(C + \tau I_p)^\alpha B \;=\; (AU_1)\,diag(s_1^\alpha)U_1^T B + \tau^\alpha \left(AB - (AU_1)U_1^T B\right) \tag{28}$$

where U_1 and s_1 indicate the first k eigen-vectors and values, respectively, and $A_{a\times p}$ and $B_{p\times b}$ are arbitrary matrices, τ is a positive scalar, and I_p is the p-dimensional identity matrix.

Proof. The derivation follows from standard properties of linear algebra, the fact that the matrix of eigen-vectors, U, is orthonormal, and Lemma 1.1 below.

$$\begin{aligned}
A(C + \tau I_p)^\alpha B &= A\left(U\mathrm{diag}(s)U^T + \tau I_p\right)^\alpha B && (29)\\
&= A\left(U\left(\mathrm{diag}(s) + \tau I_p\right)U^T\right)^\alpha B && (30)\\
&= AU\mathrm{diag}(s+\tau)^\alpha U^T B. && (31)
\end{aligned}$$

While this formula removes U from within the exponentiated term, it still includes $p - k$ eigen-values that are zero along with their corresponding eigen-vectors. Partitioning the non-zero and zero terms of s and U and into separate components produces an expression using only non-zero terms.

$$\begin{aligned}
&= A[U_1\, U_2]\begin{bmatrix} s_1 I_k + \tau & 0 \\ 0 & \tau I_{p-k}\end{bmatrix}^\alpha [U_1\, U_2]^T B && (32)\\
&= AU_1\mathrm{diag}(s_1+\tau)^\alpha U_1^T B + AU_2(\tau^\alpha I_{p-k})U_2^T B && (33)\\
&= AU_1\mathrm{diag}(s_1+\tau)^\alpha U_1^T B + \tau^\alpha AU_2U_2^T B && (34)\\
&= AU_1\mathrm{diag}(s_1+\tau)^\alpha U_1^T B + \tau^\alpha A(I_p - U_1U_1^T)B && (35)\\
&= AU_1\mathrm{diag}(s_1+\tau)^\alpha U_1^T B + \tau^\alpha(AB - (AU_1)U_1^T B) && (36)
\end{aligned}$$

□

Corollary 1.1. *This allows efficient evaluation of $A(C+\tau I_p)^\alpha B$ for many values of τ by precomputing s_1, U_1, AU_1 and $U_1^T B$.*

Lemma 1.1. *Since U is orthonormal,*

$$\begin{aligned} I_p &= UU^T && (37)\\ &= [U_1\, U_2][U_1\, U_2]^T && (38)\\ &= U_1U_1^T + U_2U_2^T && (39) \end{aligned}$$

Rearranging, it follows that $U_2U_2^T = I_p - U_1U_1^T$

5.10 Application to imputing z-statistics

Consider a matrix $X_{n\times p}$ of p genetic variants from n individuals, and a vector of continuous trait values y. Standard analysis in genome-wide association studies performs a linear regression to test association of each genetic variant with the trait. Ignoring covariates here for simplicity, this estimates a regression coefficient for the jth variant as $\hat{\beta}_j = (x_j^T x_j)^{-1} x_j^T y$, and produces a corresponding z-statistic, z_j. Within a local genomic region, genetic variants can be highly correlated due to linkage disequilibrium (LD). Therefore, even if the kth genetic variant is not directly observed, its z-statistic can be imputed from the set of observed z-statistics and the correlation structure between variants in the region. Following Pasaniuc, et al. [16], let $\Sigma_{p\times p}$ be the correlation matrix between all pairs of genetic variants, and partition the genetic variants into two sets with j storing indices with observed z-statistics and k storing indices without z-statistics. We would like to impute the values of z_k given observed values of z_j and Σ. If we model the z-statistics as a multivariate Gaussian

$$z \sim \mathcal{N}(0, \Sigma), \tag{40}$$

then the conditional distribution of z_k given z_j is

$$z_k | z_j \sim \mathcal{N}\left(\Sigma_{k,j}\Sigma_{j,j}^{-1} z_j,\ \Sigma_{j,j} - \Sigma_{k,j}\Sigma_{j,j}^{-1}\Sigma_{k,j}^T\right). \tag{41}$$

The imputed z-statistics is then

$$\hat{z}_k | z_j = \Sigma_{k,j}\Sigma_{j,j}^{-1} z_j. \tag{42}$$

This also works for logistic regression in the case / control studies.

So far, we have assumed that the correlation matrix Σ is known. But in practice, this 'LD matrix' is estimated from a 'reference panel' of a modest number of individuals that are not included in the genome-wide association study. Therefore, inverting $\Sigma_{j,j}$ requires some regularization as p becomes large in comparison to n.

In the context of the current work, Equation (42) can be in interpreted as an out-of-sample whitening transformation of the observed z-statistics, z_j, and the cross-correlation matrix, $\Sigma_{k,j}$. Letting $\hat{\Sigma}_{\lambda,j,j}$ be the regularized estimate of the correlation matrix, then the imputed z-statistic is

$$\left[\Sigma_{k,j}\hat{\Sigma}_{\lambda,j,j}^{-1/2}\right]\left[\hat{\Sigma}_{\lambda,j,j}^{-1/2}z_j\right]. \tag{43}$$

Applying this transformation by producing $\hat{\Sigma}_{\lambda,j,j}$ directly is cubic time in p_j, the number of genetic variants in set j. But applying the high dimensional whitening transformation of Equation (19) reduces this time substantially to $\mathcal{O}\left(n^2 p_j\right)$ when $n < p_j$. Imputation then depends on the observed z-statistics the cross-correlation matrix, and the SVD of X.

6 Supplementary Figures

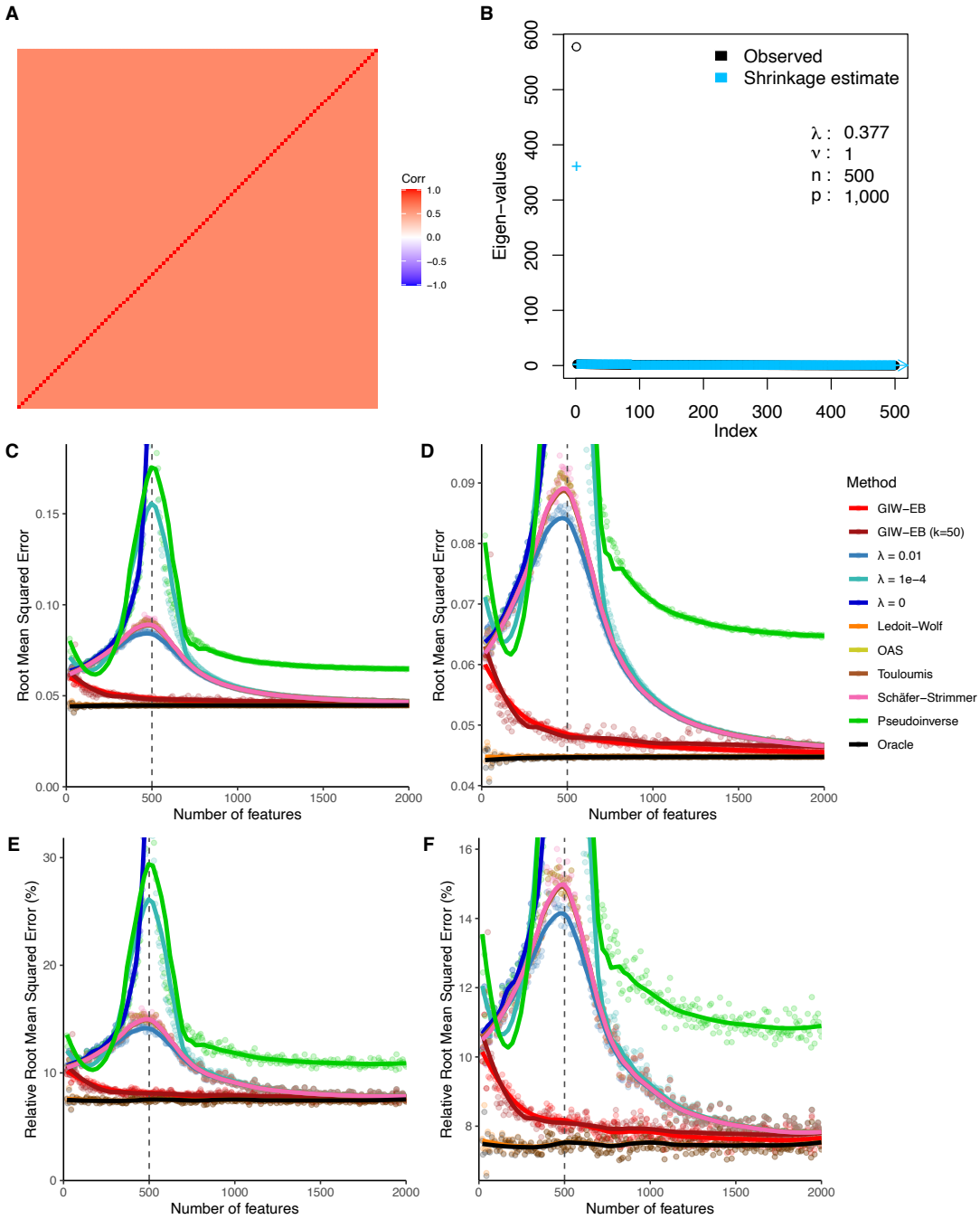

Figure S1: Out-of-sample performance of whitening transformation. Scenario 1: constant correlation. Details are given in **Figure 3**. **A**) Constant correlation matrix with $\rho = 0.6$ used in this simulation. **B**) Observed and shrunk eigen-values from simulation of $p = 1000$ features. **C,D**) rMSE of the whitening transformation for an increasing number of features with a zoom-in on the y-axis. **E,F**) rMSE scaled by the rMSE of the untransformed data with a zoom-in on the y-axis. Points indicate observed values and lines show loess smooth.

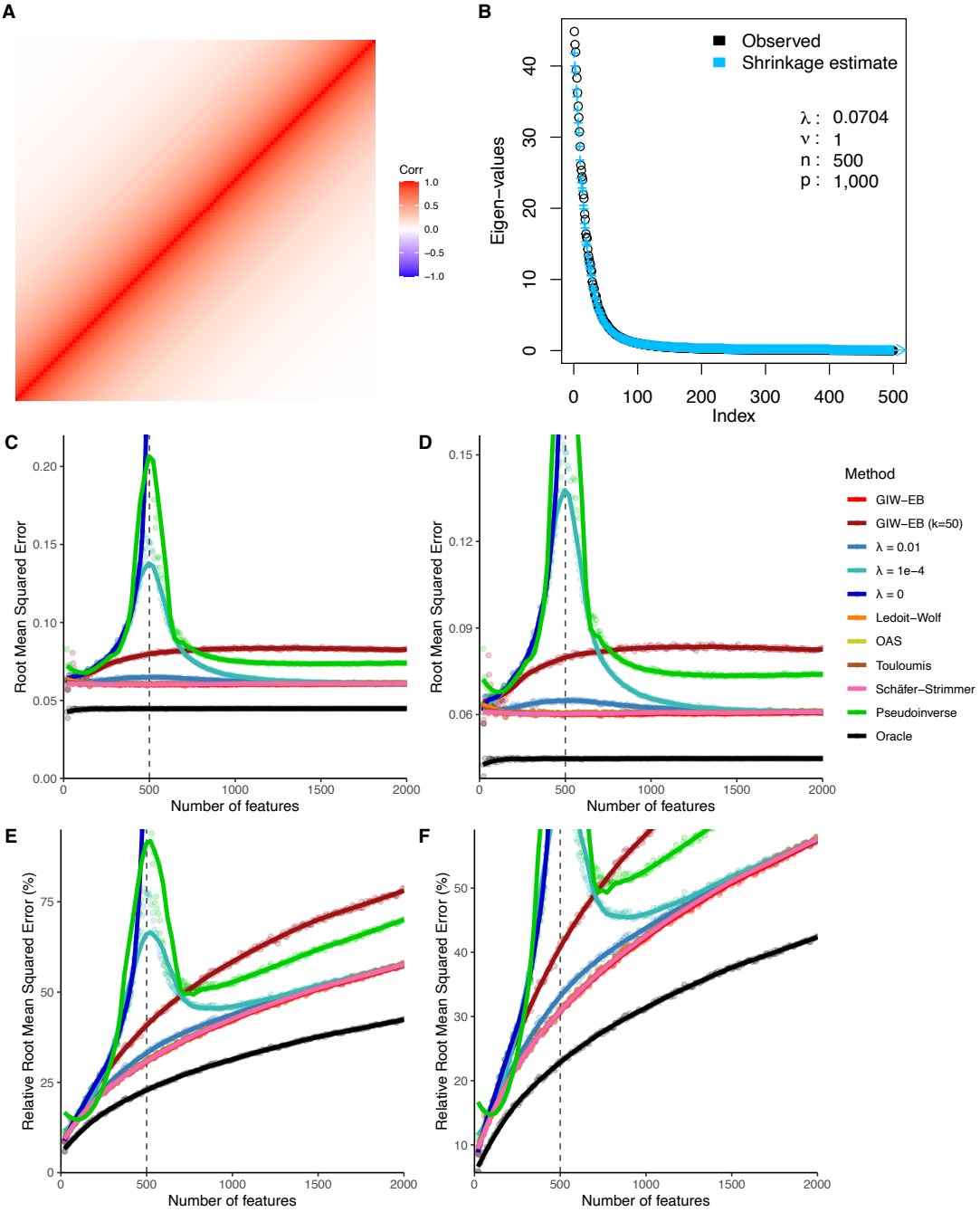

Figure S 2: Out-of-sample performance of whitening transformation. Scenario 2: auto-correlation. Details are given in **Figure 3**. **A**) Auto-correlation matrix with $\rho = 0.95$ used in this simulation. **B**) Observed and shrunk eigen-values from simulation of $p = 1000$ features. **C,D**) rMSE of the whitening transformation for an increasing number of features with a zoom-in on the y-axis. **E,F**) rMSE scaled by the rMSE of the untransformed data with a zoom-in on the y-axis. Points indicate observed values and lines show loess smooth.

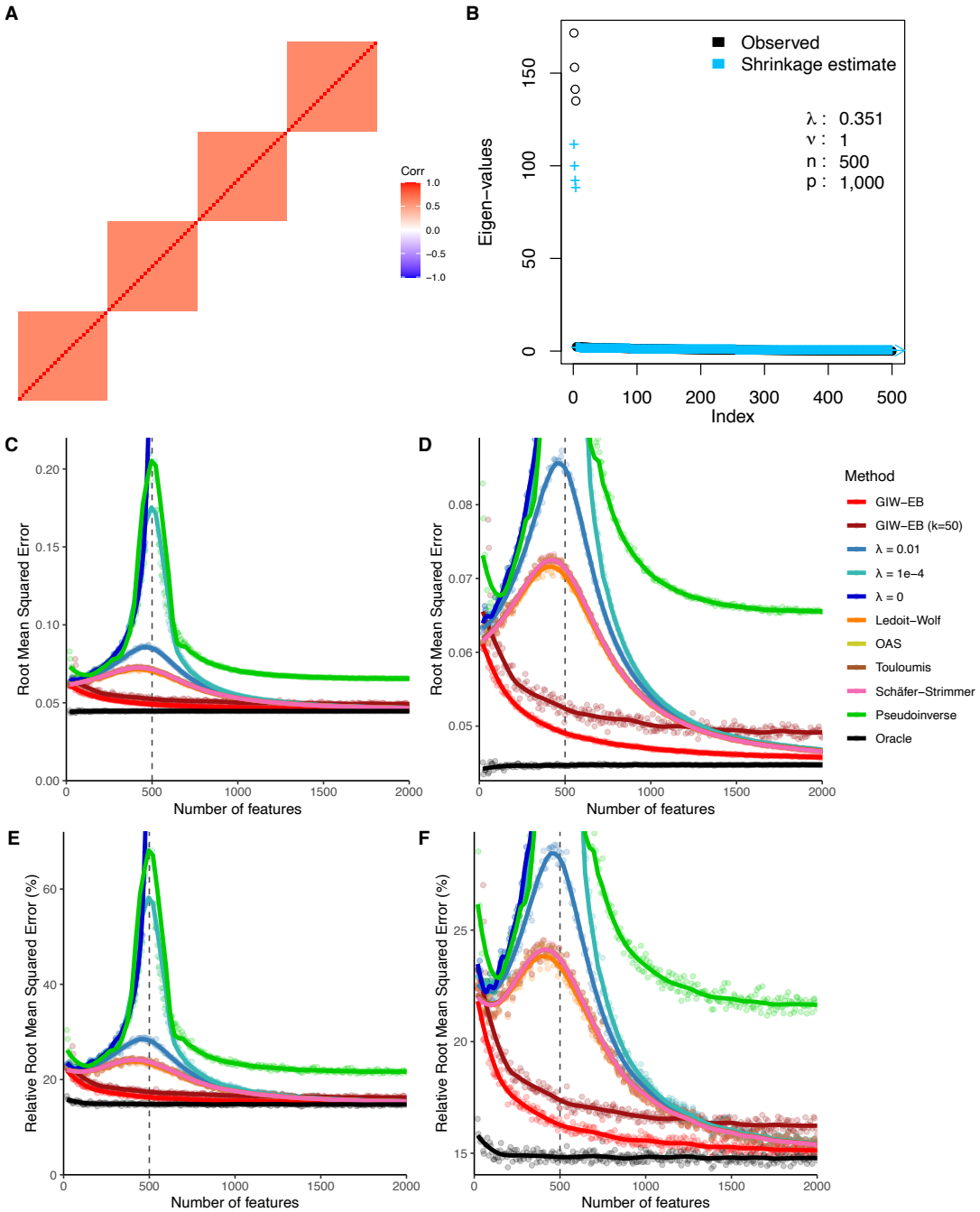

Figure S3: Out-of-sample performance of whitening transformation. Scenario 3: block constant correlation. Details are given in **Figure 3**. **A**) Block constant correlation matrix with $\rho = 0.6$ used in this simulation. **B**) Observed and shrunk eigen-values from simulation of $p = 1000$ features. **C,D**) rMSE of the whitening transformation for an increasing number of features with a zoom-in on the y-axis. **E,F**) rMSE scaled by the rMSE of the untransformed data with a zoom-in on the y-axis. Points indicate observed values and lines show loess smooth.

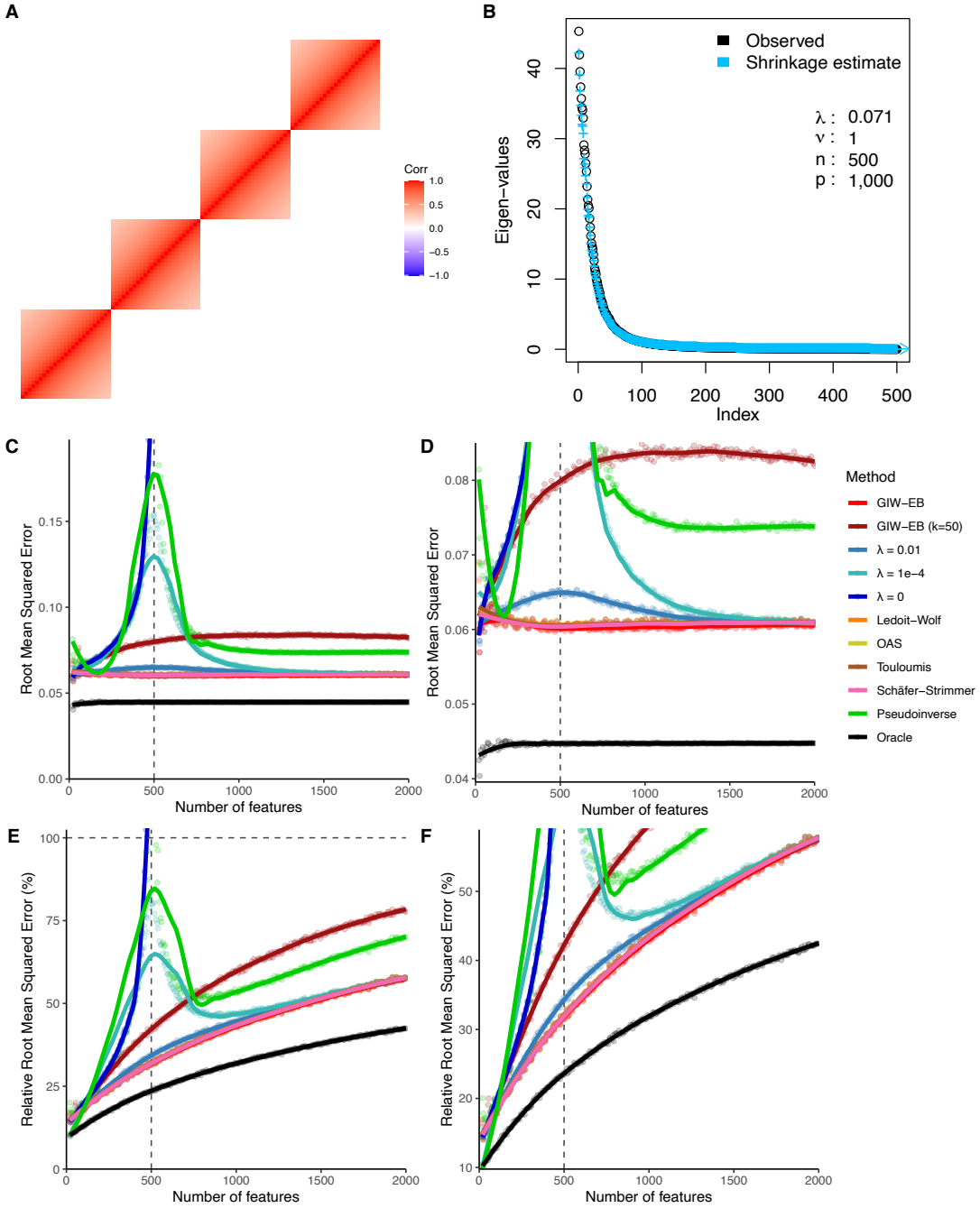

Figure S4: Out-of-sample performance of whitening transformation. Scenario 4: block auto-correlation. Details are given in **Figure 3**. **A**) Block auto-correlation matrix with $\rho = 0.95$ used in this simulation. **B**) Observed and shrunk eigen-values from simulation of $p = 1000$ features. **C,D**) rMSE of the whitening transformation for an increasing number of features with a zoom-in on the y-axis. **E,F**) rMSE scaled by the rMSE of the untransformed data with a zoom-in on the y-axis. Points indicate observed values and lines show loess smooth.

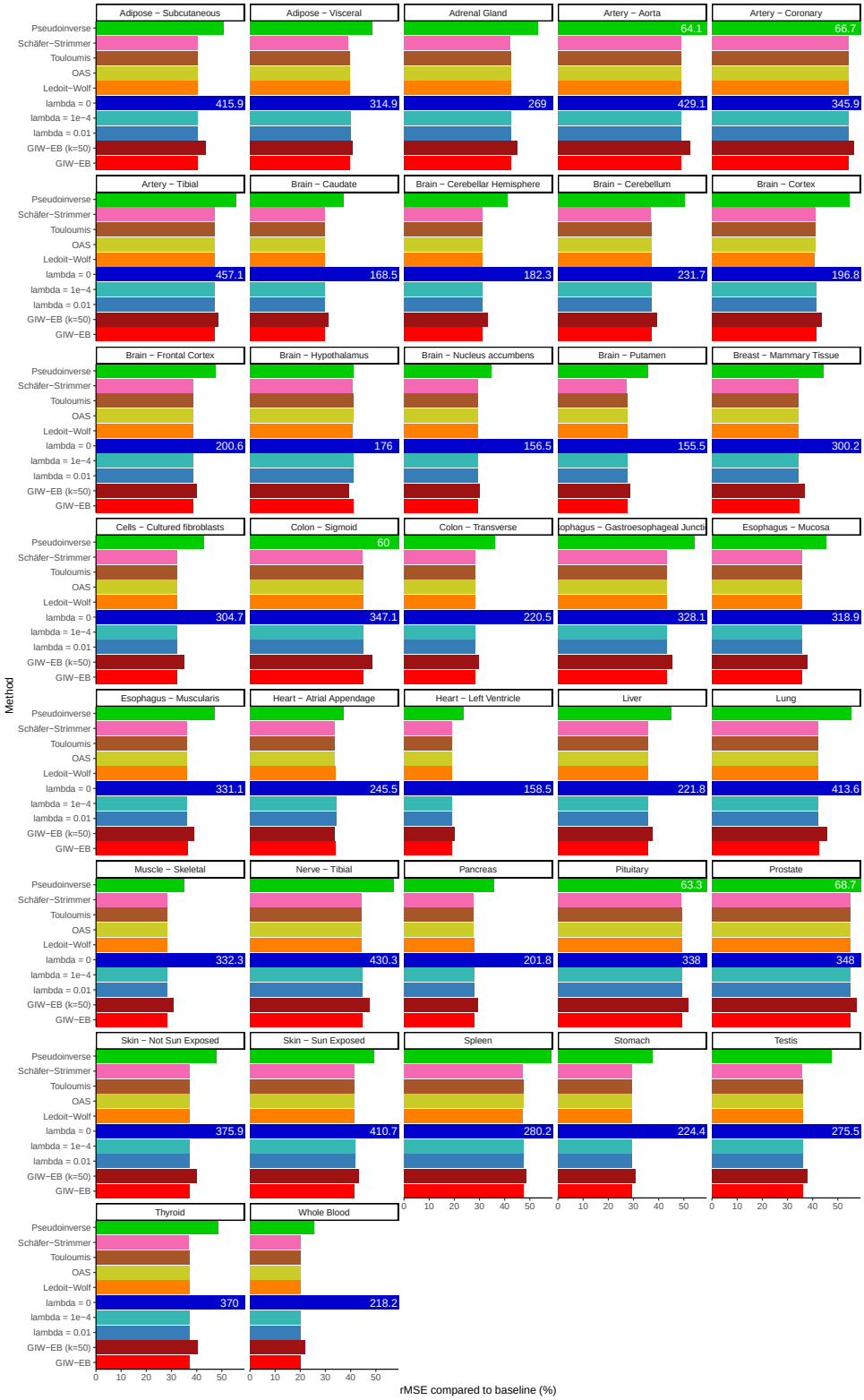

Figure S5: Application of whitening transformations to gene expression data from GTEx. The root mean square error (rMSE) of the empirical covariance after the whitening transformation is normalized by the rMSE of the covariance matrix without applying the transformation. Therefore, a value of 40% indicates that the whitening transformation reduced the rMSE (compared to the identity matrix) of the covariance to 40% its original value. Results are shown for 37 tissues. When normalized rMSE values exceed the axes, the value is shown in white text.

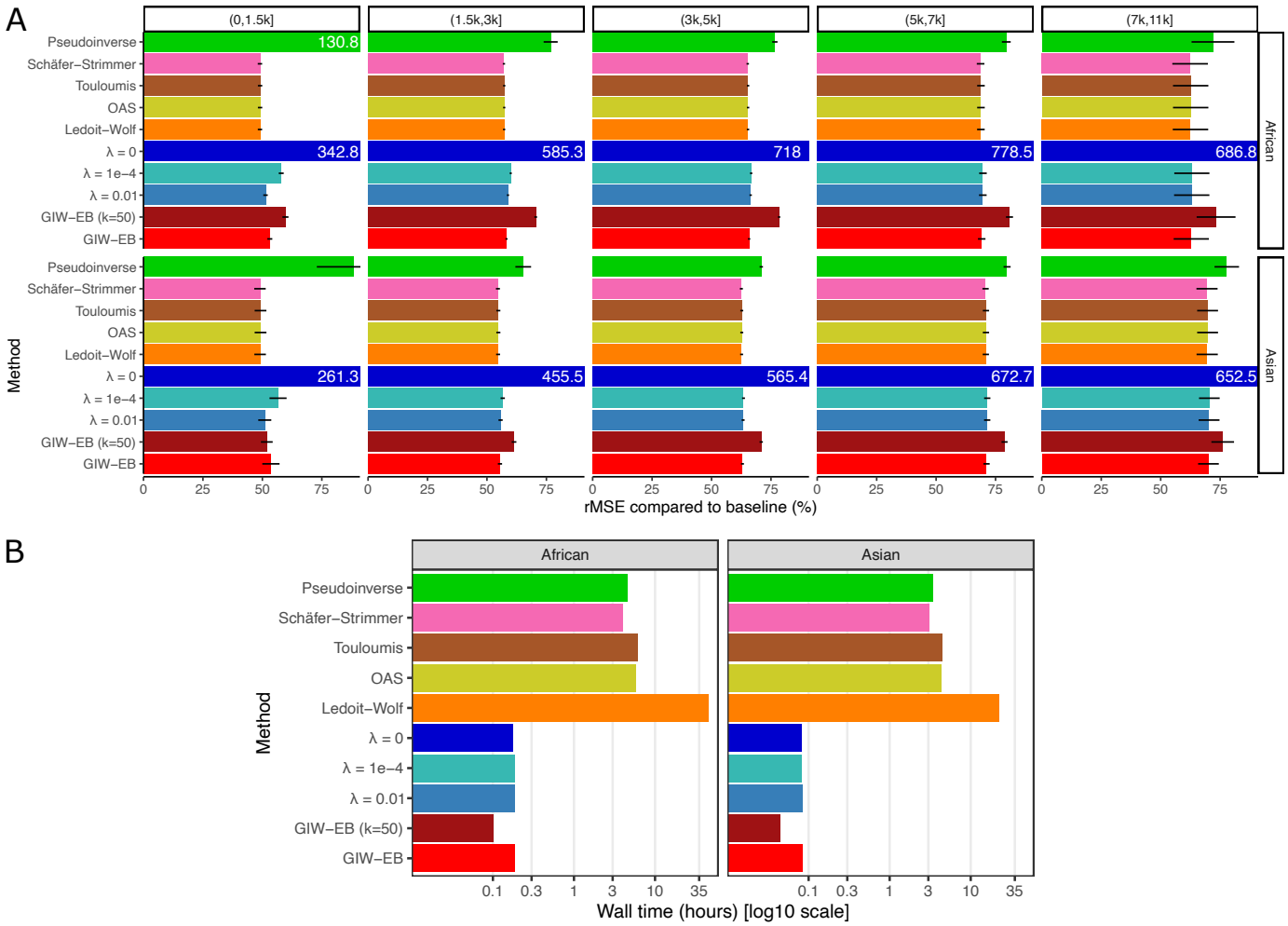

Figure S6: Out-of-sample performance of whitening transformation on genotype data from the 1000 Genomes Project. A) Root mean squared error (rMSE) compared to baseline for African and Asian individuals stratified by the number of genetic variants in the LD window. Values indicate the values averaged across windows, and bars indicate the 95% confidence interval. When normalized rMSE values exceed the axes, the value is shown in white text. **B**) Run time of each method shown on a $\log_{10}$ scale.