Maximum Likelihood for Variance Estimation in High-Dimensional Linear Models

Lee H. Dicker
Rutgers University

Murat A. Erdogdu
Stanford University

Abstract

We study maximum likelihood estimators (MLEs) for the residual variance, the signal-to-noise ratio, and other variance parameters in high-dimensional linear models. These parameters are essential in many statistical applications involving regression diagnostics, inference, tuning parameter selection for high-dimensional regression, and other applications, including genetics. The estimators that we study are not new, and have been widely used for variance component estimation in linear random-effects models. However, our analysis is new and it implies that the MLEs, which were devised for *random-effects* models, may also perform very well in high-dimensional linear models with *fixed-effects*, which are more commonly studied in some areas of high-dimensional statistics. The MLEs are shown to be consistent and asymptotically normal in fixed-effects models with random design, in asymptotic settings where the number of predictors (p) is proportional to the number of observations (n). Moreover, the estimators' asymptotic variance can be given explicitly in terms moments of the Marčenko-Pastur distribution. A variety of analytical and empirical results show that the MLEs outperform other, previously proposed estimators for variance parameters in high-dimensional linear models with fixed-effects. More broadly, the results in this paper illustrate a strategy for drawing connections between fixed- and random-effects models in high dimensions, which may be useful in other applications.

1 INTRODUCTION

Variance parameters — including the residual variance, the proportion of explained variation, and, under some interpretations (e.g. Dicker, 2014; Hartley & Rao, 1967; Janson et al., 2015), the signal-to-noise ratio — are important parameters in many statistical models. While they are often not the primary focus of a given data analysis or prediction routine, variance parameters are fundamental for a variety of tasks, including:

(i) Regression diagnostics, e.g. risk estimation (Bayati et al., 2013; Mallows, 1973).
(ii) Inference, e.g. hypothesis testing and constructing confidence intervals (Fan et al., 2012; Javanmard & Montanari, 2014; Reid et al., 2013).
(iii) Optimally tuning other methods, e.g. tuning parameter selection for high-dimensional regression (Sun & Zhang, 2012).
(iv) Other applications, e.g. genetics (De los Campos et al., 2015; Listgarten et al., 2012; Loh et al., 2015; Yang et al., 2010).

This paper focuses on variance estimation in high-dimensional linear models. We study maximum likelihood-based estimators for the residual variance and other related variance parameters. The results in this paper show that — under appropriate conditions — the maximum likelihood estimators (MLEs) outperform several other previously proposed variance estimators for high-dimensional linear models, in theory and in numerical examples, with both real and simulated data.

The MLEs studied here are not new. Indeed, they are widely used for variance components estimation in linear *random-effects* models (e.g. Searle et al., 1992). However, the analysis in this paper is primarily concerned with the performance of the MLEs in *fixed-effects* models, which are more commonly studied in some areas of modern high-dimensional statistics (e.g. Bühlmann & Van de Geer, 2011). Thus, from one perspective, this paper may be viewed as an article on model misspecification: We show that a classical es-

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. Copyright 2016 by the authors.

timator for random-effects models may also be used effectively in fixed-effects models. More broadly, the results in this paper illustrate a strategy for drawing connections between fixed- and random-effects models in high dimensions, which may be useful in other applications; for instance, a similar strategy has been employed in Dicker's (2016) analysis of ridge regression.

1.1 Related Work

There is an abundant literature on variance components estimation for *random-effects* models, going back to at least the 1940s (Crump, 1946; Satterthwaite, 1946). More recently, a substantial literature has developed on the use of random-effects models in high throughput genetics (e.g. De los Campos et al., 2015; Listgarten et al., 2012; Loh et al., 2015; Yang et al., 2010). Maximum likelihood estimators for variance parameters, like those studied in this paper, are widely used in this research area. A variety of questions about the use of MLEs for variance parameters in modern genetics remain unanswered; including fundamental questions about whether fixed- or random-effects models are more appropriate (De los Campos et al., 2015). Our work here focuses primarily on fixed-effects models, and thus differs from much of this recent work on genetics. However, Theorem 2 (see, in particular, the discussion in Section 4.2) and the data analysis in Section 6 below suggest that fixed-effects analyses could possibly offer improved power over random-effects approaches to variance estimation, in some settings.

Switching focus to high-dimensional linear models with *fixed-effects*, previous work on variance estimation has typically been built on one of two sets of assumptions: Either (i) the underlying regression parameter [$\boldsymbol{\beta}$ in the model (1) below] is highly structured (e.g. sparse) or (ii) the design matrix $X = (x_{ij})$ is highly structured [e.g. x_{ij} are iid $N(0, 1)$]. Notable work on variance estimation under the "structured $\boldsymbol{\beta}$" assumption includes (Chatterjee & Jafarov, 2015; Fan et al., 2012; Sun & Zhang, 2012). Bayati et al. (2013), Dicker (2014), and Janson et al. (2015) have studied variance estimation in the "structured X" setting. This paper also focuses on the structured X setting. As is characteristic of the structured X setting, many of the results in this paper require strong assumptions on X; however, none of our results require any sparsity assumptions on $\boldsymbol{\beta}$.

Figure 1 illustrates some potential advantages of the structured X approach. In particular, it shows that structured $\boldsymbol{\beta}$ methods for estimating the residual variance σ_0^2 can be biased, if $\boldsymbol{\beta}$ is not extremely sparse, while the structured X methods are effective regardless of sparsity. Figure 1 was generated from datasets where the design matrix X was highly structured, i.e. $x_{ij} \sim N(0, 1)$ were all iid (a detailed description of the numerical simulations used to generate Figure 1 may be found in the Supplementary Material). In general, structured X methods work best when the entries of X are uncorrelated. Indeed, by generating X with highly correlated columns and taking $\boldsymbol{\beta}$ to be very sparse, it is easy to generate plots similar to those in Figure 1, where the structured X methods are badly biased and the structured $\boldsymbol{\beta}$ methods perform well. On the other hand, if the design matrix X is correlated, it may be possible to improve the performance of structured X methods by "decorrelating" X (i.e. right-multiplying X by a suitable positive definite matrix); this is not pursued in detail here, but may be an interesting area for future research.

The method labeled "MLE" in Figure 1 is the main focus of this paper. The other methods depicted in Figure 1 that rely on structured X are "MM," "EigenPrism," and "AMP." MM is a method-of-moments estimator for σ_0^2 proposed in (Dicker, 2014); EigenPrism was proposed in (Janson et al., 2015) and is the solution to a convex optimization problem; AMP is a method for estimating σ_0^2 based on approximate message passing algorithms and was proposed in (Bayati et al., 2013). In Figure 1, each of the structured X methods is evidently unbiased. Additionally, it is clear that the variability of MLE is uniformly smaller than that of MM and EigenPrism (detailed results are reported in Table S1 of the Supplementary Material). The AMP method is unique among the four structured X methods, because it is the only one that can adapt to sparsity in $\boldsymbol{\beta}$; consequently, while the variability of MLE is smaller than that of AMP when $\boldsymbol{\beta}$ is not sparse (e.g. when sparsity is 40%), the AMP estimator has smaller variance when $\boldsymbol{\beta}$ is sparse. This is certainly an attractive property of the AMP estimator. However, little is known about the asymptotic distribution of the AMP estimator. By contrast, one important feature of the MLE is that it is asymptotically normal in high dimensions and its asymptotic variance is given by the simple formula in Theorem 2. This is a useful property for performing inference on σ_0^2 and related parameters, and for better understanding the asymptotic performance of the MLE.

1.2 Overview of the Paper

Section 2 covers preliminaries. We introduce the statistical model and define a bivariate MLE for the residual variance and the signal-to-noise ratio — this is the main estimator of interest. We also give some additional background on the "structured X" model that is studied here.

In Section 3, we present a coupling argument, which

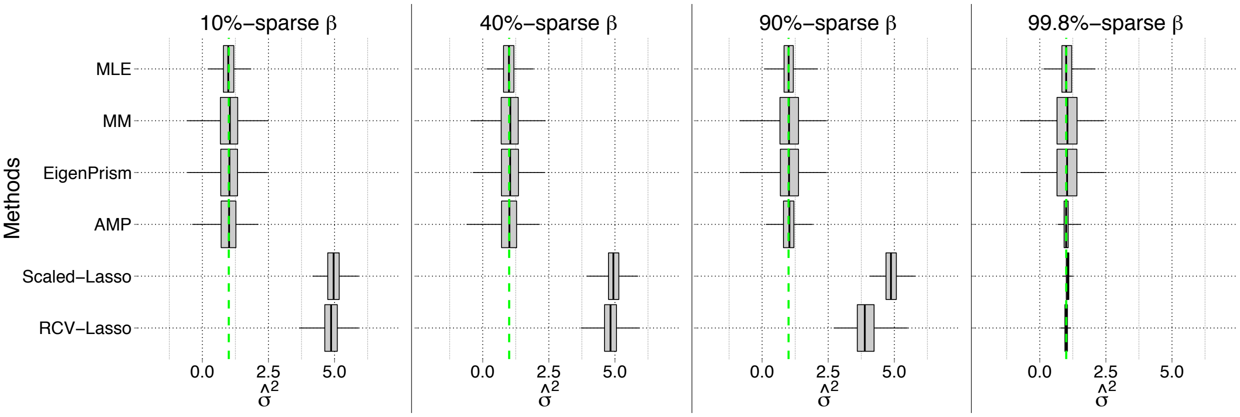

Figure 1: Estimates of the residual variance σ_0^2 from 500 independent datasets. Structured X simulations: $x_{ij} \sim N(0,1)$ are iid. Parameter values: $\sigma_0^2 = 1$, $\eta_0^2 = 4$, $n = 500$, $p = 1000$. "α%-sparse" indicates that $(\alpha/100) \times p$ coordinates of $\boldsymbol{\beta} \in \mathbb{R}^p$ are set equal to 0. Structured X estimators: MLE, MM (Dicker, 2014), EigenPrism (Janson et al., 2015), AMP (Bayati et al., 2013). Structured $\boldsymbol{\beta}$ estimators: Scaled-Lasso (Sun & Zhang, 2012), RCV-Lasso (Fan et al., 2012). Detailed description of settings and results may be found in the Supplementary Material.

is the key to relating the fixed-effects model of interest and the random-effects model that motivates the MLEs studied here.

Basic asymptotic properties of the MLEs are described in Section 4. Consistency and asymptotic normality results (given in Section 4.1) rely heavily on the coupling argument from Section 3: Theorems 1–2 follow by coupling the fixed-effects model with random predictors (the "structured X" model) to a random-effects model and then appealing to existing results for random-effects models and quadratic forms.

In Section 5, we use results from random matrix theory to derive explicit formulas for the asymptotic variance of the MLEs, which are determined by moments of the Marčenko-Pastur distribution. Using these formulas, we analytically compare the asymptotic variance of the MLEs to that of other estimators.

The results of a real data analysis involving gene expression data and single nucelotide polymorphism (SNP) data are described in Section 6. The results illustrate some potential benefits of our results in an important practical example.

A concluding discusion may be found in Section 7.

2 PRELIMINARIES

Consider a linear regression model, where outcomes $\mathbf{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n$ are related to an $n \times p$ matrix of predictors $X = (x_{ij})_{1 \le i \le n; 1 \le j \le p}$ via the equation

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}; \qquad (1)$$

$\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top \in \mathbb{R}^n$ is a random error vector, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top \in \mathbb{R}^p$ is an unknown p-dimensional parameter. The observed data consists of $(\mathbf{y}, X)$.

In this paper, we make the following distributional assumptions:

$$\epsilon_i \sim N(0, \sigma_0^2), \; x_{ij} \sim N(0,1), \qquad (2)$$

$1 \le i \le n$, $1 \le j \le p$ are all independent. The parameter $\sigma_0^2 > 0$ is the residual variance. Estimating σ_0^2 when p and n are large is one of the main objectives of the paper.[1]

The distributional assumptions (2) are strong. However, others following the "structured X" approach to variance estimation have made the same assumptions (e.g. Bayati et al., 2013; Dicker, 2014; Janson et al., 2015). There is also a substantial literature on estimating $\boldsymbol{\beta}$ in high-dimensions under similar distributional assumptions; see, for instance, research on approximate message passing (AMP) algorithms (Donoho et al., 2009; Rangan, 2011; Vila & Schniter, 2013). In the AMP literature, there has been some success at relaxing the Gaussian assumption (2) (e.g. Korada & Montanari, 2011; Rangan et al., 2014). Relaxing (2) for structured X approaches to variance estimation is an important topic for future research. (See also the discussion in Section 7.)

[1] If $p \le n$ and X has full rank, then the residual sum of squares $\hat{\sigma}^2_{\text{OLS}} = (n-p)^{-1}\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_{\text{OLS}}\|^2$ may be used to estimate σ_0^2, where $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (X^\top X)^{-1}X^\top \mathbf{y}$. Thus, estimating σ_0^2 when $p > n$ is more interesting and may be considered the main focus of this work (however, the results in Section 5.2 imply that the MLE $\hat{\sigma}^2$ outperforms $\hat{\sigma}^2_{\text{OLS}}$ even when $p < n$).

To estimate the residual variance, we introduce two auxiliary parameters that are closely related to σ_0^2: The signal strength, $\tau_0^2 = \|\boldsymbol{\beta}\|^2$, and the signal-to-noise ratio, $\eta_0^2 = \tau_0^2/\sigma_0^2$ (here and throughout the paper, $\|\cdot\|$ denotes the ℓ^2-norm). Now define the bivariate parameter $\boldsymbol{\theta}_0 = (\sigma_0^2, \eta_0^2)$. Consider the estimator

$$\hat{\boldsymbol{\theta}} = \underset{\sigma^2,\eta^2\geq 0}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta} = (\sigma^2, \eta^2)$ and

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2n}\log\det\left(\frac{\eta^2}{p}XX^\top + I\right) - \frac{1}{2\sigma^2 n}\mathbf{y}^\top\left(\frac{\eta^2}{p}XX^\top + I\right)^{-1}\mathbf{y}. \tag{4}$$

The estimator $\hat{\boldsymbol{\theta}}$ is the main focus of this paper.[2] It is the MLE for $\boldsymbol{\theta}_0$ under a linear random-effects model, where the regression parameter $\boldsymbol{\beta} \sim N\{0, (\tau_0^2/p)I\}$ is Gaussian and independent of $\boldsymbol{\epsilon}$ and X; indeed, under this random-effects model, the distribution of $\mathbf{y}|X \sim N\{0, (\tau_0^2/p)XX^\top + \sigma_0^2 I\}$ is Gaussian.

3 A COUPLING ARGUMENT

Theoretical properties of $\hat{\boldsymbol{\theta}}$ have already been studied extensively in the literature, *within the context of random-effects models* (e.g. Dicker & Erdogdu, 2016; Hartley & Rao, 1967; Jiang, 1996; Searle et al., 1992). In this paper we are primarily concerned with the performance of $\hat{\boldsymbol{\theta}}$ in the fixed-effects model (1)–(2), where $\boldsymbol{\beta}$ is non-random. The main objective of this section is to explain why one might expect $\hat{\boldsymbol{\theta}}$ to perform well in the fixed-effects model. To do this, we show that the model (1)–(2) can be coupled to a random-effects model. The key points of the coupling argument are that the distribution of the design matrix X and the estimator $\hat{\boldsymbol{\theta}}$ are both invariant under orthogonal transformations, i.e.

$$X \overset{\mathscr{D}}{=} XU^\top, \tag{5}$$

$$\hat{\boldsymbol{\theta}}(\mathbf{y}, X) = \hat{\boldsymbol{\theta}}(\mathbf{y}, XU^\top) \tag{6}$$

for any $p \times p$ orthogonal matrix U ("$\overset{\mathscr{D}}{=}$" denotes equality in distribution). The argument is given in more detail below.

Let $\mathcal{U}$ be an independent uniformly distributed $p \times p$ orthogonal matrix, define $\tilde{\boldsymbol{\beta}} = \mathcal{U}\boldsymbol{\beta}$, and let

$$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}. \tag{7}$$

Then $(\tilde{\mathbf{y}}, X)$ may be viewed as data drawn from a random-effects linear model, with the regression parameter $\tilde{\boldsymbol{\beta}} \sim \text{uniform}\{S^{p-1}(\tau_0^2)\}$ uniformly distribution on the sphere in $\mathbb{R}^p$ of radius τ_0, $S^{p-1}(\tau_0^2) = \{\mathbf{u} \in \mathbb{R}^p;\ \|\mathbf{u}\|^2 = \tau_0^2\}$. Next define the MLE based on (7),

$$\tilde{\boldsymbol{\theta}} = \underset{\sigma^2,\eta^2\geq 0}{\operatorname{argmax}}\, \tilde{\ell}(\boldsymbol{\theta}),$$

where $\tilde{\ell}(\boldsymbol{\theta})$ is defined just as in (4), except that $\mathbf{y}$ is replaced by $\tilde{\mathbf{y}}$. Then the distribution of $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y}, X)$ is the exact same as the distribution of $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\tilde{\mathbf{y}}, X)$; more precisely, for any Borel set $B \subseteq \mathbb{R}^2$, we have

$$\begin{aligned}\mathbb{P}\{\hat{\boldsymbol{\theta}}(\mathbf{y}, X) \in B\} &= \mathbb{P}\{\hat{\boldsymbol{\theta}}(X\mathcal{U}^\top\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, X) \in B\} \\ &= \mathbb{P}\{\hat{\boldsymbol{\theta}}(X\mathcal{U}^\top\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, X\mathcal{U}^\top) \in B\} \\ &= \mathbb{P}\{\hat{\boldsymbol{\theta}}(X\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, X) \in B\} \\ &= \mathbb{P}\{\tilde{\boldsymbol{\theta}}(\tilde{\mathbf{y}}, X) \in B\},\end{aligned}$$

where the second equality holds because of (6) and the third inequality follows from (5). Thus, the distributional properties of $\hat{\boldsymbol{\theta}}$ — the estimator based on the original data from the fixed-effects model — are the exact same as the distributional properties of $\tilde{\boldsymbol{\theta}}$ — the estimator based on the random-effects data, i.e.

$$\hat{\boldsymbol{\theta}} \overset{\mathscr{D}}{=} \tilde{\boldsymbol{\theta}}. \tag{8}$$

The distributional identity (8) links the fixed-effects model (1)–(2) with the random-effects model (7). However, one remaining issue is that the random-effects vector $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^\top \in \mathbb{R}^p$ has correlated components. While most of the existing work on random-effects models applies to models with independent random-effects, Dicker & Erdogdu (2016) recently derived concentration bounds for variance components estimators in models with correlated random-effects, which can be applied in the present setting. Dicker & Erdogdu's bounds rely on the existence of a tight independent coupling; this means finding a random vector $\boldsymbol{\beta}^\star$ with independent components, such that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|$ is small with high probability. Such a coupling is described in the following paragraph.

Let $\mathbf{z} \sim N(0, I)$ be an independent p-dimensional Gaussian random vector. Without loss of generality, we assume that

$$\tilde{\boldsymbol{\beta}} = \frac{\tau_0\mathbf{z}}{\|\mathbf{z}\|} \sim \text{uniform}\{S^{p-1}(\tau_0^2)\}.$$

Now define $\boldsymbol{\beta}^\star = (\tau_0/p^{1/2})\mathbf{z} \sim N\{0, (\tau_0^2/p)I\}$. Then $\boldsymbol{\beta}^\star$ has independent components and, when p is large, $\|\mathbf{z}\| \approx p^{1/2}$ and $\tilde{\boldsymbol{\beta}} \approx \boldsymbol{\beta}^\star$. This is the required coupling. More precisely, applying Chernoff's bound to the chi-squared random-variable $\|\mathbf{z}\|^2$ implies that if $0 \leq r <$

[2] If $\ell(\boldsymbol{\theta})$ has multiple maximizers, then use any predetermined rule to select $\hat{\boldsymbol{\theta}}$. Maximizing $\ell(\boldsymbol{\theta})$ is a nonconvex problem; however, it has been extremely well-studied and is straightforward, e.g. (Demidenko, 2013).

τ_0^2, then

$$\begin{aligned}\mathbb{P}\left\{\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\| > r\right\} &= \mathbb{P}\left\{\tau_0 \left|\frac{\|\mathbf{z}\|}{\sqrt{p}} - 1\right| > r\right\} \\ &= \mathbb{P}\left\{\|\mathbf{z}\|^2 > p\left(1 + \frac{r}{\tau_0}\right)^2\right\} \\ &\quad + \mathbb{P}\left\{\|\mathbf{z}\|^2 < p\left(1 - \frac{r}{\tau_0}\right)^2\right\} \\ &\leq 2\exp\left(-\frac{pr^2}{2\tau_0^2}\right), \qquad (9)\end{aligned}$$

for $0 \leq r < \tau_0$.

4 ASYMPTOTIC PROPERTIES

4.1 Consistency and Asymptotic Normality

In this section, we derive consistency and asymptotic normality results for $\hat{\boldsymbol{\theta}}$, under the assumption that $p/n \to \rho \in (0, \infty) \setminus \{1\}$.[3] Consistency, i.e. convergence in probability, is a direct consequence of (9) and Proposition 2 from (Dicker & Erdogdu, 2016).

Theorem 1. *Assume that* (1)–(2) *holds. Let $\mathcal{K} \subseteq (0, \infty)$ be a compact set and suppose that $\sigma_0^2, \eta_0^2 \in \mathcal{K}$. Additionally suppose that $\rho \in (0, \infty) \setminus 1$. Then $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ in probability, as $p/n \to \rho$.*

It takes a little bit more work to show that $\hat{\boldsymbol{\theta}}$ is asymptotically normal, but the approach is similar to the standard argument for asymptotic normality of maximum likelihood and M-estimators (e.g. Chapter 5 of Van der Vaart, 2000). The key fact is that $\hat{\boldsymbol{\theta}}$ solves the score equation $S(\hat{\boldsymbol{\theta}}) = 0$, where

$$\begin{aligned}S(\boldsymbol{\theta}) &= \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \left(\frac{\partial \ell}{\partial \sigma^2}(\boldsymbol{\theta}),\ \frac{\partial \ell}{\partial \eta^2}(\boldsymbol{\theta})\right)^\top, \\ \frac{\partial \ell}{\partial \sigma^2}(\boldsymbol{\theta}) &= \frac{1}{2\sigma^4 n}\mathbf{y}^\top\left(\frac{\eta^2}{p}XX^\top + I\right)^{-1}\mathbf{y} - \frac{1}{2\sigma^2}, \\ \frac{\partial \ell}{\partial \eta^2}(\boldsymbol{\theta}) &= \frac{1}{2\sigma^2 n}\mathbf{y}^\top\left(\frac{1}{p}XX^\top\right)\left(\frac{\eta^2}{p}XX^\top + I\right)^{-2}\mathbf{y} \\ &\quad - \frac{1}{2n}\mathrm{tr}\left\{\left(\frac{1}{p}XX^\top\right)\left(\frac{\eta^2}{p}XX^\top + I\right)^{-1}\right\}.\end{aligned}$$

Taylor expanding $S(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$, we obtain

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx -n^{1/2}\left\{\frac{\partial S}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right\}^{-1} S(\boldsymbol{\theta}_0). \qquad (10)$$

The right-hand side of (10) (in particular, $S(\boldsymbol{\theta}_0)$) involves a quadratic form in $\mathbf{y}$. If the data were drawn from a random-effects model with independent random-effects, then standard theory (for random-effects models or quadratic forms) could be used to argue that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is approximately normal. However, since we are interested in the fixed-effects model (1)–(2) there is an additional step: We must go through a variant of the coupling argument from Section 3 again.

Define $\tilde{S}(\boldsymbol{\theta})$ to be the same as $S(\boldsymbol{\theta})$, except replace $\mathbf{y}$ with $\tilde{\mathbf{y}}$. Then, using (8),

$$\begin{aligned}n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &\overset{\mathscr{D}}{=} n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\approx -n^{1/2}\left\{\frac{\partial \tilde{S}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right\}^{-1}\tilde{S}(\boldsymbol{\theta}_0). \qquad (11)\end{aligned}$$

The entries of $\tilde{S}(\boldsymbol{\theta})$ are quadratic forms in

$$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon} = X\frac{\tau_0 \mathbf{z}}{\|\mathbf{z}\|} + \boldsymbol{\epsilon}. \qquad (12)$$

Thus, the approximation (11) allows us to shift focus to the random-effects model (7) and quadratic forms in $\tilde{\mathbf{y}}$. This is an important step, because of the relative simplicity of random-effects models. However, the coordinates of $\tilde{\boldsymbol{\beta}}$ are correlated, which is an additional obstacle that must be addressed. To overcome this issue, we use the fact that $\tilde{\mathbf{y}}$ can be represented in terms of $\mathbf{z}$ and $\boldsymbol{\epsilon}$, as in (12), and apply the delta method to $\tilde{S}(\boldsymbol{\theta}_0)$ (that is, we use another Taylor expansion; cf. Ch. 3 of Van der Vaart, 2000). Following this strategy, $\tilde{S}(\boldsymbol{\theta}_0)$ can be approximated by quadratic forms in the random vectors $\mathbf{z}$ and $\boldsymbol{\epsilon}$, which have independent components. Asymptotic normality for $\tilde{S}(\boldsymbol{\theta}_0)$ then follows from existing normal approximation results for quadratic forms in independent random variables.

Theorem 2. *Assume that the conditions of Theorem 1 hold. Additionally, define*

$$\mathcal{I}_N(\boldsymbol{\theta}_0) = \begin{bmatrix} \iota_2(\boldsymbol{\theta}_0) & \iota_3(\boldsymbol{\theta}_0) \\ \iota_3(\boldsymbol{\theta}_0) & \iota_4(\boldsymbol{\theta}_0) \end{bmatrix},$$

where

$$\iota_k(\boldsymbol{\theta}_0) = \frac{1}{2n\sigma_0^{2(4-k)}}\mathrm{tr}\left\{\left(\frac{1}{p}XX^\top\right)^{k-2}\left(\frac{\eta_0^2}{p}XX^\top + I\right)^{2-k}\right\},$$

for $k = 2, 3, 4$, and

$$\mathcal{J}(\boldsymbol{\theta}_0)^{-1} = \mathcal{I}_N(\boldsymbol{\theta}_0)^{-1} - \frac{2\eta_0^4}{(p/n)}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \qquad (13)$$

Then $n^{1/2}\mathcal{J}(\boldsymbol{\theta}_0)^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N(0, I)$ in distribution, as $p/n \to \rho$.

[3] If $\rho = 1$, then $\inf \lambda_{\min}(p^{-1}XX^\top) = 0$, where $\lambda_{\min}(p^{-1}XX^\top)$ is the smallest nonzero eigenvalue of $p^{-1}XX^\top$. This causes some technical difficulties, which arise frequently in random matrix theory, but can often be overcome with some additional work, e.g. (Bai et al., 2003).

As described above, the proof of Theorem 2 relies on three steps: (i) The distributional identity and Taylor approximation (11), (ii) applying the delta method to handle the correlated coordinates of $\tilde{\boldsymbol{\beta}}$, and (iii) normal approximation results for quadratic forms in $\mathbf{z}$ and $\boldsymbol{\epsilon}$. These steps can be rigorously justified using Theorem 2 and Proposition 2 from (Dicker & Erdogdu, 2016).

4.2 Random-Effects Model

The matrix $\mathcal{I}_N(\boldsymbol{\theta}_0)$ defined in Theorem 2 is the Fisher information matrix for $\boldsymbol{\theta}_0$ under the Gaussian random-effects linear model, where $\boldsymbol{\beta} \sim N\{0, (\tau_0^2/p)I\}$. Well-known results on random-effects models (Jiang, 1996) and standard likelihood theory (e.g. Ch. 6 of Lehmann & Casella, 1998) imply that under this random-effects model, $\hat{\boldsymbol{\theta}}$ is asymptotically efficient with asymptotic variance $\mathcal{I}_N(\boldsymbol{\theta}_0)^{-1}$. Thus, Theorem 2 implies that $\hat{\eta}^2$ has smaller asymptotic variance under the fixed-effects model than under the random-effects model (on the other hand, $\hat{\sigma}^2$ has the same asymptotic variance in both models). As a consequence, confidence intervals for variance parameters related to η_0^2 tend to be smaller under the fixed-effects model than under the random-effects models, and corresponding tests may have more power in the fixed-effects model. A practical illustration of this observation may be found in the data analysis in Section 6.

5 MARČENKO-PASTUR APPROXIMATIONS

5.1 Formulas for the Asymptotic Variance of $\hat{\boldsymbol{\theta}}$

Let $\mathbb{F} = \mathbb{F}_{n,p}$ denote the empirical distribution of the eigenvalues of $p^{-1}XX^\top$. Marčenko & Pastur (1967) showed that under (2), if $p/n \to \rho \in (0, \infty)$, then $\mathbb{F}_{n,p} \to F_\rho$, where F_ρ is the Marčenko-Pastur distribution with density

$$f_\rho(s) = \max\{1-\rho, 0\}\delta_0(s) + \frac{\rho}{2\pi}\frac{\sqrt{(b-s)(s-a)}}{s},$$

for $a \le s \le b$ and $a = (1-\rho^{-1/2})^2$, $b = (1+\rho^{-1/2})^2$. In Theorem 2, observe that the entries of $\mathcal{I}_N(\boldsymbol{\theta}_0)$ can be reexpressed as

$$\iota_k(\boldsymbol{\theta}_0) = \frac{1}{2\sigma_0^{2(4-k)}}\int\left(\frac{s}{\eta_0^2 s+1}\right)^{k-2} d\mathbb{F}(s).$$

Thus, it is evident that if $\rho \in (0, \infty)$, then $\lim_{p/n\to\rho}\iota_k(\boldsymbol{\theta}_0) = \iota_k^\rho(\boldsymbol{\theta}_0)$, where

$$\iota_k^\rho(\boldsymbol{\theta}_0) = \frac{1}{2\sigma_0^{2(4-k)}}\int_a^b\left(\frac{s}{\eta_0^2 s+1}\right)^{k-2} f_\rho(s)\ ds. \quad (14)$$

The quantities $\iota_k^\rho(\boldsymbol{\theta}_0)$ — and, consequently, the asymptotic variance of $\hat{\boldsymbol{\theta}}$, given in Theorem 2 — can be computed explicitly using properties of the Stieltjes transform of the Marčenko-Pastur distribution. The Stieltjes transform of the Marčenko-Pastur distribution is defined by

$$\begin{aligned} m_\rho(z) &= \int_a^b \frac{1}{s+z} f_\rho(s)\ ds \quad (15)\\ &= \frac{1}{2z}\left[1-\rho z-\rho+\{(1-\rho z-\rho)^2+4\rho z\}^{1/2}\right], \end{aligned}$$

where the second equality holds for $z > 0$ (see, e.g., Ch. 3 of Bai & Silverstein, 2010). Comparing (14)–(15) and differentiating $m_\rho(z)$ as necessary implies that

$$\iota_2^\rho(\boldsymbol{\theta}_0) = \frac{1}{2\sigma_0^4}, \quad (16)$$

$$\iota_3^\rho(\boldsymbol{\theta}_0) = \frac{1}{2\sigma_0^2}\left\{\frac{1}{\eta_0^2} - \frac{1}{\eta_0^4}m_\rho\left(\frac{1}{\eta_0^2}\right)\right\}, \quad (17)$$

$$\iota_4^\rho(\boldsymbol{\theta}_0) = \frac{1}{2}\left\{\frac{1}{\eta_0^4} - \frac{2}{\eta_0^6}m_\rho\left(\frac{1}{\eta_0^2}\right) - \frac{1}{\eta_0^8}m_\rho'\left(\frac{1}{\eta_0^2}\right)\right\}. \quad (18)$$

To compute the asymptotic variance of $\hat{\boldsymbol{\theta}}$ in terms of $m_\rho(z)$, we use (13) and the expressions for $\iota_k^\rho(\boldsymbol{\theta}_0)$ given above. In particular, let $\psi_{k+l}^\rho(\boldsymbol{\theta}_0)$ denote the kl-element of $\lim_{p/n\to\rho}\mathcal{J}(\boldsymbol{\theta}_0)^{-1}$. It follows from basic matrix algebra that

$$\psi_2^\rho(\boldsymbol{\theta}_0) = 2\sigma_0^4\left[1 - \frac{\{m(\eta_0^{-2}) - \eta_0^2\}^2}{m'(\eta_0^{-2}) + m(\eta_0^{-2})^2}\right], \quad (19)$$

$$\psi_3^\rho(\boldsymbol{\theta}_0) = -\frac{2\sigma_0^2\eta_0^4\{m(\eta_0^{-2}) - \eta_0^2\}}{m'(\eta_0^{-2}) + m(\eta_0^{-2})^2}, \quad (20)$$

$$\psi_4^\rho(\boldsymbol{\theta}_0) = -\frac{2\eta_0^8}{m'(\eta_0^{-2}) + m(\eta_0^{-2})^2} - \frac{2\eta_0^4}{\rho}. \quad (21)$$

From (15) and (19)–(21), it is apparent that the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is an easily computed algebraic function in σ_0^2, η_0^2, and ρ.

5.2 Comparison to Previously Proposed Estimators

Define the proportion of explained variation[4] $r_0^2 = \tau_0^2/(\tau_0^2+\sigma_0^2) = \eta_0^2/(\eta_0^2+1)$ and the estimator $\hat{r}^2 = \hat{\eta}^2/(\hat{\eta}^2+1)$. Consistency and asymptotic normality for $\hat{r}^2$ follows easily from Theorems 1–2 and the delta method. In this section, we compare the asymptotic variance of $\hat{\sigma}^2$ and $\hat{r}^2$ to that of other previously proposed estimators for σ_0^2 and r_0^2.

[4] The proportion of explained variation is an important parameter for regression diagnostics. It is also important in genetics; see Section 6

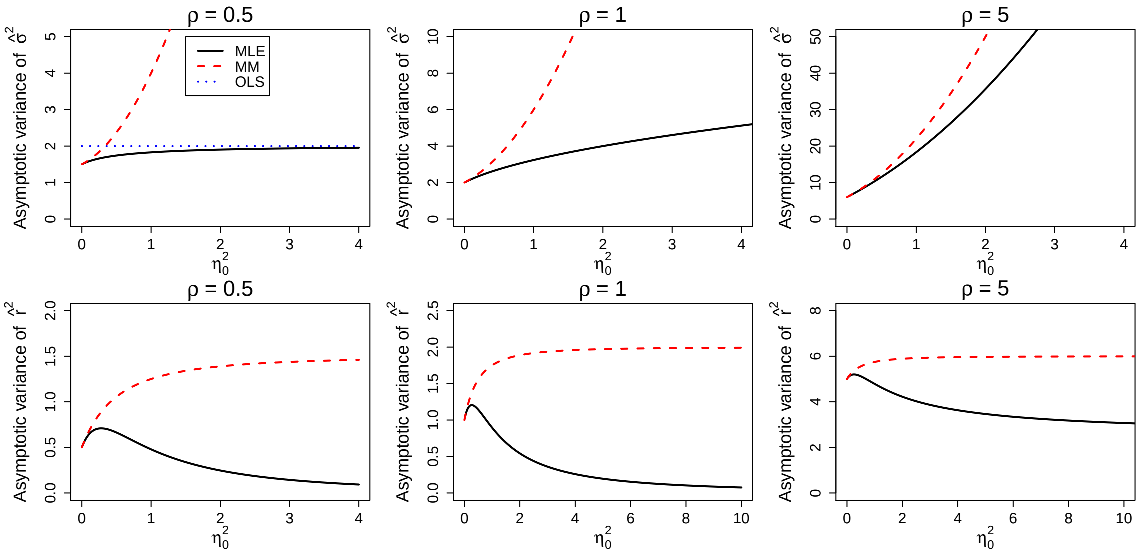

Figure 2: Plots of the asymptotic variance of various estimators for σ_0^2 and r_0^2, as functions of η_0^2 (the signal-to-noise ratio) for $\rho = 0.5, 1, 5$. Top row: Asymptotic variances of estimators for σ^2 (v_{MLE}, v_{MM}, and v_{OLS}). Bottom row: Asymptotic variances of estimators for r^2 (w_{MLE} and w_{MM}). Theorem 2 does not apply to the case where $\rho = 1$; the asymptotic variances for the MLEs in this case are conjectured to be $v_{\mathrm{MLE}}(\eta_0^2, 1)$ and $w_{\mathrm{MLE}}(\eta_0^2, 1)$. The values v_{MM}, v_{OLS}, and w_{MM} are given explicitly in the Supplementary Material.

First define the standardized asymptotic variance of $\hat{\sigma}^2$,

$$\begin{aligned} v_{\mathrm{MLE}}(\eta_0^2, \rho) &= \frac{n}{2\sigma_0^4} \lim_{p/n \to \rho} \mathrm{Var}(\hat{\sigma}^2) = \frac{1}{2\sigma_0^4} \psi_2^{\rho}(\boldsymbol{\theta}_0) \\ &= 1 - \frac{\{m(\eta_0^{-2}) - \eta_0^2\}^2}{m'(\eta_0^{-2}) + m(\eta_0^{-2})^2}. \end{aligned}$$

The asymptotic variance of $\hat{r}^2$ can be derived using (21) and the delta method; assuming $p/n \to \rho \in (0, \infty) \setminus \{1\}$, it is given by

$$\begin{aligned} &w_{\mathrm{MLE}}(\eta_0^2, \rho) \\ &\quad = \frac{n}{2} \lim_{p/n \to \rho} \mathrm{Var}(\hat{r}^2) = \frac{1}{2} \left(\frac{1}{\eta_0^2 + 1} \right)^4 \psi_4^{\rho}(\boldsymbol{\theta}_0) \\ &\quad = - \left(\frac{\eta_0^2}{\eta_0^2 + 1} \right)^4 \left\{ \frac{1}{m'(\eta_0^{-2}) + m(\eta_0^{-2})^2} + \frac{1}{\rho \eta_0^2} \right\}. \end{aligned}$$

The asymptotic variances v_{MLE} and w_{MLE} are plotted in Figure 2, along with the asymptotic variances of method-of-moments (MM) estimators for σ_0^2 and r_0^2 (Dicker, 2014), and, in the top-left plot, the asymptotic variance of $\hat{\sigma}_{\mathrm{OLS}}^2 = (n-p)^{-1} \|\mathbf{y} - X\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\|^2$, which was defined in Footnote 1. Explicit formulas for the asymptotic variances of the MM and OLS estimators are given in the Supplementary Material (these formulas have been derived elsewhere previously). We note that the OLS estimator appears only in the top-left plot of Figure 2 because $\hat{\sigma}_{\mathrm{OLS}}^2$ is undefined when $p > n$ (i.e. $\rho > 1$) and has infinite variance when $p = n$; furthermore, even for $p < n$, it is unclear how a corresponding estimate of r_0^2 should be defined. From Figure 2, it is clear that the MLEs have smaller asymptotic variance (i.e. they are more efficient) than the MLE and OLS estimators.

In other work on variance estimation within the "structured X" paradigm, Janson et al. (2015) proposed the EigenPrism method for estimating σ_0^2 and η_0^2. While (Janson et al., 2015) do describe methods for constructing confidence intervals for σ_0^2, they do not give an explicit formula for the asymptotic variance of EigenPrism; instead, they gives bounds on the asymptotic variance in terms of the solution to a convex optimization problem. Given this discrepancy, we omit the EigenPrism estimator from Figure 2. However, the performance of EigenPrism relative to the MLE and the MM estimator for σ_0^2 is depicted in Figure 1 (see also Table S1); in these settings, the MLE appears to have significantly smaller variance than EigenPrism.

6 DATA ANALYSIS: ESTIMATING HERITABILITY

To illustrate how the results from this paper could be used in a practical application and to further investigate the MLE's performance vis-à-vis other previously proposed methods for variance estimation, we conducted a real data analysis involving publicly available gene expression data and single nucelotide polymorphism (SNP) data.

6.1 Background

In genetics, heritability measures the fraction of variability in an observed trait (the phenotype) that can be explained by genetic variability (the genotype). In the context of linear models like (1)–(2), it is natural to identify heritability with the parameter $r_0^2 = \eta_0^2/(\eta_0^2+1)$ (De los Campos et al., 2015; Yang et al., 2010). More specifically, if y_i is the i-th subject's phenotype and $x_{ij} \in \{0,1,2\}$ is a measure of their genotype at location j (i.e. the minor allele count at a given location in the i-th subject's DNA), then r^2 measures the fraction of variability in the phenotype y_i that is explained by the genotype $(x_{ij})_{j=1,\ldots p}$. Heritability is often studied in relation to easily observable phenotypes, such as human height (Yang et al., 2010). However, other important work has focused on more basic phenotypes, such as gene expression levels, which may be measured by mRNA abundance (Stranger et al., 2007, 2012).

6.2 Analysis and Results

This data analysis was based on publicly available SNP data from the International HapMap Project and gene expression data collected by Stranger and coauthors[5] on $n = 80$ individuals from the Han Chinese HapMap population. For each of 100 different genes, we estimated the heritability r_0^2 of the gene's expression level using genotype data from nearby SNPs, and then computed corresponding confidence intervals (CIs). In other words, we estimated r_0^2 separately for 100 different genes, based on a different collection of SNPs (ranging from $p = 17$ to $p = 190$) for each gene, and then computed CIs for each r_0^2. The 100 genes were selected from genes that (Stranger et al., 2007) had previously identified as having significant association between gene expression levels and nearby SNPs in the Japanese HapMap population. More detailed preprocessing steps are described in the Supplementary Material.[6] The main objective of this data analysis is to compare the lengths of the various CIs for r_0^2. Given that the genes under consideration were already identified as significant in another study, we take shorter CIs to be an indicator of more efficient inference on r_0^2.

For each gene, we calculated two estimates for r_0^2: (i) The MLE $\hat{r}^2$ and (ii) the method-of-moments estimate $\hat{r}^2_{\mathrm{MM}}$ proposed in (Dicker, 2014). After computing these estimates, we then constructed three Wald-type 95% CIs for the r_0^2 of each gene using: (i) $\hat{r}^2$ and the asymptotic standard error suggested by Theorem 2 (referred to as MLE-FE); (ii) $\hat{r}^2$ and the asymptotic standard errors corresponding to a Gaussian random-effects model, where $\boldsymbol{\beta} \sim N\{0, (\tau_0^2/p)I\}$ (MLE-RE); and (iii) $\hat{r}^2_{\mathrm{MM}}$ and the asymptotic standard error given by Proposition 2 of (Dicker, 2014) (MM). Summary statistics from our analysis are reported in Table 1. The results in Table 1 indicate that the MLE-FE intervals are typically the shortest, which potentially points towards the improved power of this approach.

Table 1: Mean and empirical SD (in parentheses) of 95% CI length and corresponding estimates $\hat{r}^2$, computed over 100 genes of interest.

	MLE-FE	MLE-RE	MM
CI length	0.43 (0.13)	0.47 (0.14)	0.45 (0.26)
$\hat{r}^2$	0.21 (0.21)	0.21 (0.21)	0.22 (0.23)

7 DISCUSSION

The MLEs studied in this paper outperform a variety of other previously proposed methods for variance parameter estimation in high-dimensional linear models. Additionally, our methods highlight connections between fixed- and random-effects models in high dimensions, which may form the basis for further research in this area. Investigating the extent to which the distributional assumptions (2) may be relaxed will be important for determining the breadth of applicability of the results in this paper, and that of other "structured X" methods for high-dimensional data analysis. In this paper, the key implication of assumption (2) is the invariance property (5). Two possible directions relaxing (5) are: (i) Replacing the identity (5) with an approximate identity (as could be satisfied by certain classes of sub-Gaussian random vectors $(x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$) and (ii) assuming that (5) holds only for U belonging to certain subsets G of $p \times p$ orthogonal matrices, such as $G = \{p \times p \text{ permutation matrices}\}$ (this relaxation seems relevant for iid, but not necessarily Gaussian, x_{ij}).

Acknowledgements

L.H. Dicker's work on this project was supported by NSF Grant DMS-1208785 and NSF CAREER Award DMS-1454817.

[5] Accessed at http://hapmap.ncbi.nlm.nih.gov/ and http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264/

[6] We emphasize that the entries of X are discrete and thus violate the normality assumptions required for most of the theoretical results in this paper, even after preprocessing. However, additional experiments with simulated data (not reported here) suggest that many of these results remain valid for iid discrete x_{ij}. Pursuing theoretical support for these observations is of great interest.

References

Bai, Z. and Silverstein, J. W. *Spectral Analysis of Large Dimensional Random Matrices.* Springer, 2010.

Bai, Z. D., Miao, B., and Yao, J. F. Convergence rates of spectral distributions of large sample covariance matrices. *SIAM J. Matrix Anal. A.*, 25:105–127, 2003.

Bayati, M., Erdogdu, M. A., and Montanari, A. Estimating the LASSO risk and noise level. *NIPS 2013*, 2013.

Bühlmann, P. and Van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011.

Chatterjee, S. and Jafarov, J. Prediction error of cross-validated Lasso. *arXiv preprint arXiv:1502.06291*, 2015.

Crump, S. L. The estimation of variance components in analysis of variance. *Biometrics Bull.*, 2:7–11, 1946.

De los Campos, G., Sorensen, D., and Gianola, D. Genomic heritability: What is it? *PLoS Genet.*, 11:e1005048, 2015.

Demidenko, E. *Mixed Models: Theory and Applications with R.* Wiley, 2013.

Dicker, L. H. Variance estimation in high-dimensional linear models. *Biometrika*, 101:269–284, 2014.

Dicker, L. H. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

Dicker, L. H. and Erdogdu, M. A. Flexible results for quadratic forms with applications to variance components estimation. *Ann. Stat.*, 2016. To appear.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *P. Natl. Acad. Sci. USA*, 106:18914–18919, 2009.

Fan, J., Guo, S., and Hao, N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. Roy. Stat. Soc. B*, 74:37–65, 2012.

Hartley, H. O. and Rao, J. N. K. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54:93–108, 1967.

Janson, L., Barber, R. F., and Candès, E. EigenPrism: Inference for high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1505.02097*, 2015.

Javanmard, A. and Montanari, A. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014.

Jiang, J. REML estimation: Asymptotic behavior and related topics. *Ann. Stat.*, 24:255–286, 1996.

Korada, S. B. and Montanari, A. Applications of the Lindeberg principle in communications and statistical learning. *IEEE T. Inform. Theory*, 57:2440–2450, 2011.

Lehmann, E. L. and Casella, G. *Theory of Point Estimation.* Springer, 2nd edition, 1998.

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., and Heckerman, D. Improved linear mixed models for genome-wide association studies. *Nat. Methods*, 9:525–526, 2012.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., and Price, A. L. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, 47:285–290, 2015.

Mallows, C. L. Some comments on C_p. *Technometrics*, 15: 661–675, 1973.

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Math. U.S.S.R. Sb.*, 1:457–483, 1967.

Rangan, S. Generalized approximate message passing for estimation with random linear mixing. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 2168–2172. IEEE, 2011.

Rangan, S., Schniter, P., and Fletcher, A. On the convergence of approximate message passing with arbitrary matrices. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 236–240. IEEE, 2014.

Reid, S., Tibshirani, R., and Friedman, J. A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274*, 2013.

Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics Bull.*, 2: 110–114, 1946.

Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components.* Wiley, 1992.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A. S., Bird, C. P., Beazley, C., Inglem, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. Population genomics of human gene expression. *Nat. Genet.*, 39:1217–1224, 2007.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., Price, A. L., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P., and Dermitzakis, E. T. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.*, 8:e1002639, 2012.

Sun, T. and Zhang, C.-H. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.

Van der Vaart, A. W. *Asymptotic Statistics.* Cambridge University Press, 2000.

Vila, J. P. and Schniter, P. Expectation-maximization Gaussian-mixture approximate message passing. *IEEE T. Signal Proces.*, 61:4658–4672, 2013.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42:565–569, 2010.