# Variance estimation in high-dimensional linear models

**Lee H. Dicker***

*Department of Statistics and Biostatistics*
*Rutgers University*
*501 Hill Center, 110 Frelinghuysen Road*
*Piscataway, NJ 08854*
*e-mail:* ldicker@stat.rutgers.edu

**Abstract:** The residual variance and the proportion of explained variation are important quantities in many statistical models and model fitting procedures. They play an important role in regression diagnostics, model selection procedures, and in determining the performance limits in many problems. In this paper, we propose new method-of-moments based estimators for the residual variance, the proportion of explained variation and other related quantities, such as the $\ell^2$-signal strength. The proposed estimators are consistent and asymptotically normal in high-dimensional linear models with Gaussian predictors and errors, where the number of predictors $d$ is proportional to the number of observations $n$; in fact, consistency holds even in settings where $d/n \to \infty$. Existing results on residual variance estimation in high-dimensional linear models depend on sparsity in the underlying signal. Our results require no sparsity assumptions and imply that the residual variance and the proportion of explained variation may be consistently estimated even when $d > n$ and the underlying signal itself is non-estimable. Basic numerical work suggests that some of our distributional assumptions may be relaxed. A real data analysis involving gene expression data and single nucleotide polymorphism data further illustrates the performance of the proposed methods.

**AMS 2000 subject classifications:** Primary 62J05; secondary 62F12, 15B52.
**Keywords and phrases:** Asymptotic normality, Proportion of explained variation, Random matrix theory, Residual variance, Signal-to-noise ratio.

## 1. Introduction

Consider the linear model

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where $y_1, \ldots, y_n \in \mathbb{R}$ and $x_1 = (x_{11}, \ldots, x_{1d})^{\mathrm{T}}, \ldots, x_n = (x_{n1}, \ldots, x_{nd})^{\mathrm{T}} \in \mathbb{R}^d$ are observed outcomes and $d$-dimensional predictors, respectively, $\epsilon_1, \ldots, \epsilon_n \in \mathbb{R}$ are unobserved independent and identically distributed errors with $E(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2 > 0$, and

*Supported by NSF Grant DMS-1208785

$\beta = (\beta_1, \ldots, \beta_d)^{\mathrm{T}} \in \mathbb{R}^d$ is an unknown $d$-dimensional parameter. Let $y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ denote the $n$-dimensional vector of outcomes and $X = (x_1, \ldots, x_n)^{\mathrm{T}}$ denote the $n \times d$ matrix of predictors. Also let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}} \in \mathbb{R}^n$. Then (1) may be re-expressed as $y = X\beta + \epsilon$. In this paper, we focus on the case where the predictors $x_i$ are random. More specifically, we assume that $x_1, \ldots, x_n$ are independent and identically distributed random vectors with mean $E(x_i) = 0$ and $d \times d$ positive definite covariance matrix $\mathrm{cov}(x_i) = \Sigma$. The $x_i$ are additionally assumed to be independent of $\epsilon$. In practice, the assumption $E(x_i) = 0$ is often untenable and it may be appropriate to add an intercept term to the linear model (1). All of our theoretical results in this paper remain valid in cases where $E(x_i) \neq 0$ and an intercept term is included in the model, upon centering the data and replacing $n$ with $n - 1$.

Let $\tau^2 = \beta^{\mathrm{T}}\Sigma\beta = ||\Sigma^{1/2}\beta||^2$, where $||\cdot||$ denotes the $\ell^2$-norm. Then $\tau^2$ is a measure of the overall $\ell^2$-signal strength and $\sigma^2 = \mathrm{var}(\epsilon_i) = \mathrm{var}\{E(y_i \mid x_i)\}$ is the residual variance. This paper is concerned with identifying effective estimators for $\sigma^2$, $\tau^2$, and related quantities in high-dimensional linear models, where $d$ and $n$ are large; we are particularly interested in settings where $d > n$. The parameters $\sigma^2$ and $\tau^2$ are important in many problems in statistics. In estimation and prediction problems, $\sigma^2$ frequently determines the scale of an estimator's risk under quadratic loss. Reliable estimates of $\sigma^2$ may be required to compute popular model selection statistics, such as AIC, BIC, or RIC (Akaike, 1974; Foster and George, 1994; Schwarz, 1978; Zou et al., 2007). Good estimates of $\sigma^2$ and $\tau^2$ may be used to derive plug-in estimates of other quantities, such as the proportion of explained variation $r^2 = \tau^2/(\sigma^2+\tau^2)$ and the signal-to-noise ratio $\tau^2/\sigma^2$. The proportion of explained variation is important for various regression diagnostics, including goodness-of-fit testing, and is related to important practical concepts, such as heritability in genetics; the signal-to-noise ratio is important for regularization parameter selection and determines the performance limits in certain high-dimensional regression problems, e.g. Dicker (2013) and a 2013 technical report available from the author.

If the predictors $x_i$ are nondegenerate and $n - d$ is large, then estimating $\sigma^2$ and $\tau^2$ is straightforward. Indeed, if $d \leq n$ and $X$ has full rank, let $\hat{\beta}_{\mathrm{ols}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ be the ordinary least squares estimator for $\beta$; then

$$\hat{\sigma}_0^2 = \frac{1}{n-d}||y - X\hat{\beta}_{\mathrm{ols}}||^2 = \frac{1}{n-d}||y||^2 - \frac{1}{n-d}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \tag{2}$$
$$\hat{\tau}_0^2 = \frac{1}{n}||y||^2 - \hat{\sigma}_0^2 = \frac{1}{n-d}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y - \frac{d}{n(n-d)}||y||^2$$

are unbiased estimators for $\sigma^2$ and $\tau^2$, respectively. Furthermore, if $n - d \to \infty$, then $\hat{\sigma}_0^2$ and $\hat{\tau}_0^2$ are consistent and asymptotically normal, under fairly mild conditions.

When $d > n$, it is more challenging to construct reliable estimators for $\sigma^2$ and $\tau^2$; indeed, if $d > n$, then $X^{\mathrm{T}}X$ is not invertible and the estimator $\hat{\sigma}_0^2$ breaks down. Fan et al. (2012)

and Sun and Zhang (2012) have proposed methods for estimating $\sigma^2$ that are effective when $d \geq n$ and $\beta$ is sparse, e.g., the $\ell^0$- or $\ell^1$-norm of $\beta$ is small. Fan et al.'s (2012) and Sun and Zhang's (2012) results apply in settings where $d/n \to \infty$. However, their underlying sparsity assumptions may be untenable in certain instances and this can dramatically affect the performance of the proposed estimators, as demonstrated by our numerical simulations in Section 5.2.

In this paper, we propose new method of moments-based estimators for $\sigma^2$ and $\tau^2$ that are consistent when $d/n^2 \to 0$. Some of our results require $\sigma^2$ and $\tau^2$ to be bounded, but we make no additional sparsity assumptions on $\beta$. When $d/n \to \rho \in [0, \infty)$, the proposed estimators are shown to be asymptotically normal with rate $n^{-1/2}$. We also derive consistent and asymptotically normal estimators for $r^2 = \tau^2/(\sigma^2 + \tau^2)$; moreover, the same techniques may be used to derive reliable estimators for other functions of $\sigma^2$ and $\tau^2$, like the signal-to-noise ratio. One consequence of our results is that $\sigma^2$ and $\tau^2$ may be consistently estimated when $d > n$, even if $\beta$ itself is non-estimable. In addition to providing theoretical results, we illustrate the performance of the proposed estimators through simulation studies and a real-data analysis involving gene expression data and single nucleotide polymorphism data.

## 2. Assumptions

### *2.1. Distributional assumptions*

Though sparsity is not required in this paper, we do make strong distributional assumptions about the data. Henceforth, we assume that

$$\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2), \quad x_1, \ldots, x_n \sim N(0, \Sigma) \tag{3}$$

are all independent. Normality is used repeatedly throughout our analysis. However, we expect that key aspects of many of the results in this paper remain valid under weaker distributional assumptions. This is explored via simulation in Section 5.1.

While it is of interest to relax (3), this assumption facilitates the use of a collection of tools for random matrices developed by Chatterjee (2009), which provides the means for a highly flexible theoretical analysis, especially regarding asymptotic normality. To give a sense of the relevant arguments, let $g(\sigma^2, \tau^2)$ be some function of $\sigma^2, \tau^2$. We use Chatterjee's results to bound the total variation distance between estimators for $g(\sigma^2, \tau^2)$ and a standard normal random variable, and then show that this distance vanishes asymptotically for certain $g(\sigma^2, \tau^2)$ of interest.

Alternative approaches to studying asymptotic normality in random matrix theory do not require an underlying normality assumption, e.g., Bai et al. (2007) and Pan and Zhou (2008). These could potentially be applied in the problems considered here, but this may be significantly more complex, especially when considering general estimands of the form $g(\sigma^2, \tau^2)$.

### 2.2. Correlation among predictors

The covariance matrix $\mathrm{cov}(x_i) = \Sigma$ plays an important role in our analysis. The initially proposed estimators for $\sigma^2$ and $\tau^2$ are derived under the assumption that $\Sigma$ is known, which is equivalent to assuming that $\Sigma = I$; see Section 3.1. These estimators are unbiased, consistent, and asymptotically normal. We subsequently propose modified estimators for $\sigma^2$ and $\tau^2$ in cases where $\Sigma$ is unknown, but a norm-consistent estimator for $\Sigma$ is available, or $\Sigma$ and $\beta$ satisfy conditions described in Section 4.2, which are closely related to other conditions appearing in the random matrix theory literature (Bai et al., 2007; Pan and Zhou, 2008). It remains of interest to find estimators for $\sigma^2$ and $\tau^2$ that are consistent for broader classes of $\Sigma$.

## 3. Independent predictors: $\Sigma = I$

### 3.1. The basic estimators

We assume $\Sigma = I$ throughout the discussion in Section 3. However, conditions on $\Sigma$ will be stated explicitly in all formal results; in particular, Theorem 1 in Section 3.3, on asymptotic normality, holds for arbitrary positive definite $\Sigma$. If $\Sigma \neq I$, but $\Sigma$ is known, then one easily reduces to the case where $\Sigma = I$ by replacing $(X, \beta)$ with $(X\Sigma^{-1/2}, \Sigma^{1/2}\beta)$.

The estimators for $\sigma^2$ and $\tau^2$ proposed in this paper are based on the method of moments. Using basic facts about moments of the normal and Wishart distributions, which may be found in Section S3 of the Supplementary Material, one finds that

$$E\left(\frac{1}{n}||y||^2\right) = \tau^2 + \sigma^2, \quad E\left(\frac{1}{n^2}||X^{\mathrm{T}}y||^2\right) = \frac{d+n+1}{n}\tau^2 + \frac{d}{n}\sigma^2. \tag{4}$$

The key observation is that these expressions are non-degenerate linear combinations of $\tau^2$ and $\sigma^2$. It follows that unbiased estimators of $\sigma^2$ and $\tau^2$ may be found by taking linear combinations of $n^{-1}||y||^2$ and $n^{-2}||X^{\mathrm{T}}y||^2$. Define

$$\hat{\sigma}^2 = \frac{d+n+1}{n(n+1)}||y||^2 - \frac{1}{n(n+1)}||X^{\mathrm{T}}y||^2, \quad \hat{\tau}^2 = -\frac{d}{n(n+1)}||y||^2 + \frac{1}{n(n+1)}||X^{\mathrm{T}}y||^2.$$

for all positive integers $d, n$. Our first result follows directly from (4).

**Lemma 1.** *If $\Sigma = I$, then $E(\hat{\sigma}^2) = \sigma^2$ and $E(\hat{\tau}^2) = \tau^2$.*

### 3.2. Consistency

Let $\hat{\theta} = (\hat{\sigma}^2, \hat{\tau}^2)$ and let $S = (n^{-1}||y||^2, n^{-2}||X^{\mathrm{T}}y||^2)$. The covariance matrix of $\hat{\theta}$ is important for understanding the asymptotic properties of $\hat{\sigma}^2$ and $\hat{\tau}^2$. Since $\hat{\theta} = AS$, where

$$A = \begin{pmatrix} \frac{d+n+1}{n+1} & -\frac{n}{n+1} \\ -\frac{d}{n+1} & \frac{n}{n+1} \end{pmatrix}, \tag{5}$$

it follows that $\mathrm{cov}(\hat{\theta}) = A\mathrm{cov}(S)A^{\mathrm{T}}$. The covariance matrices for $\hat{\theta}$ and $S$ are computed explicitly in Lemma S1 and Corollary S1 of the Supplementary Material. Asymptotic approximations for the entries of $\mathrm{cov}(\hat{\theta})$ are given in the next result, which also gives basic consistency properties for $\hat{\sigma}^2, \hat{\tau}^2$. The result follows directly from Corollary S1 in the Supplementary Material.

**Lemma 2.** *If $\Sigma = I$, then*

$$\begin{aligned}
\mathrm{var}(\hat{\sigma}^2) &= \frac{2}{n}\left\{\frac{d}{n}(\sigma^2+\tau^2)^2 + \sigma^4 + \tau^4\right\}\left\{1 + O\left(\frac{1}{n}\right)\right\}, \\
\mathrm{var}(\hat{\tau}^2) &= \frac{2}{n}\left\{\left(1+\frac{d}{n}\right)(\sigma^2+\tau^2)^2 - \sigma^4 + 3\tau^4\right\}\left\{1 + O\left(\frac{1}{n}\right)\right\}, \\
\mathrm{cov}(\hat{\sigma}^2, \hat{\tau}^2) &= -\frac{2}{n}\left\{\frac{d}{n}(\sigma^2+\tau^2)^2 + 2\tau^4\right\}\left\{1 + O\left(\frac{1}{n}\right)\right\}.
\end{aligned}$$

*In particular,* $|\hat{\sigma}^2 - \sigma^2|,\ |\hat{\tau}^2 - \tau^2| = O_P[\{(d+n)/n^2\}^{1/2}(\sigma^2+\tau^2)]$.

*Remark* 1. If $\sigma^2, \tau^2$ are bounded, then Lemma 2 implies that $\hat{\sigma}^2$ and $\hat{\tau}^2$ converge to $\sigma^2$ and $\tau^2$, respectively, at rate $(d+n)^{1/2}/n$; in particular, $\hat{\sigma}^2$ and $\hat{\tau}^2$ are consistent whenever $d/n^2 \to 0$. Define the plug-in estimator for $r^2$, $\hat{r}^2 = \hat{\tau}^2/(\hat{\sigma}^2 + \hat{\tau}^2)$. If $\sigma^2, \tau^2$ are contained in some compact subset of $(0, \infty)$, then Lemma 2 and Slutsky's theorem imply that $\hat{r}^2$ is consistent for $r^2$, whenever $d/n^2 \to 0$. The analogous plug-in estimator for the signal-to-noise ratio $\tau^2/\sigma^2$ has similar properties.

*Remark* 2. It is instructive to compare the asymptotic variance of $\hat{\sigma}^2$ to that of $\hat{\sigma}_0^2$, defined in (2). If $n \to \infty$ and $d/n \to \rho \in [0, 1)$, then $\mathrm{var}(\hat{\sigma}_0^2) \sim 2\sigma^4/\{n(1-\rho)\}$ and $\mathrm{var}(\hat{\sigma}^2) \sim (2/n)\{\rho(\sigma^2+\tau^2)^2 + \sigma^4 + \tau^4\}$. Evidently, $\mathrm{var}(\hat{\sigma}^2)$ increases with the signal strength $\tau^2$, while $\mathrm{var}(\hat{\sigma}_0^2)$ does not depend on $\tau^2$; this may be an undesirable feature of $\hat{\sigma}^2$. On the other hand, $\mathrm{var}(\hat{\sigma}^2) < \mathrm{var}(\hat{\sigma}_0^2)$ when $\tau^2$ is small or $\rho$ is close to 1.

*Remark* 3. Suppose $c_1, c_2 > 0$ are fixed real numbers. Lemma 2 implies that if $d/n \to \rho \in [0, \infty)$, then $\hat{\sigma}^2$, $\hat{\tau}^2$ are consistent in the sense that

$$\lim_{d/n\to\rho} \sup_{\substack{0\le\sigma^2\le c_1 \\ 0\le\tau^2\le c_2}} E\{(\hat{\sigma}^2 - \sigma^2)^2\} = \lim_{d/n\to\rho} \sup_{\substack{0\le\sigma^2\le c_1 \\ 0\le\tau^2\le c_2}} E\{(\hat{\tau}^2 - \tau^2)^2\} = 0.$$

In a 2013 technical report available from the author, it is shown that if $\rho > 0$, then it is impossible to estimate $\beta$ in this setting. In particular, if $\rho > 0$, then

$$\liminf_{d/n \to \rho} \inf_{\hat{\beta}} \sup_{\substack{0 \le \sigma^2 \le c_1 \\ 0 \le \tau^2 \le c_2}} E(||\hat{\beta} - \beta||^2) > 0,$$

where the infimum is over all measurable estimators for $\beta$. Thus, Lemma 2 describes methods for consistently estimating $\sigma^2$ and $\tau^2$ in high-dimensional linear models, where $d > n$ and it is impossible to estimate $\beta$.

### *3.3. Asymptotic normality*

Define the total variation distance between random variables $u$ and $v$, $d_{TV}(u, v) = \sup_{B \in \mathcal{B}(\mathbb{R})} |\text{pr}(u \in B) - \text{pr}(v \in B)|$, where $\mathcal{B}(\mathbb{R})$ denotes the collection of Borel sets in $\mathbb{R}$. The next theorem is this paper's main result on asymptotic normality. It is a direct application of results due to Chatterjee (2009) and it is proved in the Supplementary Material.

**Theorem 1.** *Let $\lambda_1 = ||n^{-1}X^{\mathrm{T}}X||$ be the operator norm of $n^{-1}X^{\mathrm{T}}X$, i.e., $\lambda_1$ is the largest eigenvalue of $n^{-1}X^{\mathrm{T}}X$. Let $h : \mathbb{R}^2 \to \mathbb{R}$ be a function with continuous second order partial derivatives, let $\nabla h$ denote the gradient of $h$ and let $\nabla^2 h$ denote the Hessian of $h$. Suppose $\psi^2 = \text{var}\{h(S)\} < \infty$, where $S = (n^{-1}||y||^2, n^{-2}||X^{\mathrm{T}}y||^2)$, and let $w$ be a normal random variable with the same mean and variance as $h(S)$. Then*

$$d_{TV}\{h(S), w\} = O\left(\frac{||\Sigma||^{3/2}\xi\nu}{n^{3/2}\psi^2}\right), \tag{6}$$

*where*

$$\xi = \xi(\sigma^2, \tau^2, \Sigma, d, n) = \gamma_4^{1/4} + \gamma_2^{1/4} + \gamma_0^{1/4}\tau(\tau + 1),$$
$$\nu = \nu(\sigma^2, \tau^2, \Sigma, d, n) = \eta_8^{1/4} + \eta_4^{1/4} + \eta_0^{1/4}\tau^2(\tau^2 + 1) + \gamma_4^{1/4} + \gamma_0^{1/4}(\tau^2 + 1)$$

*and, for non-negative integers $k$,*

$$\gamma_k = \gamma_k(\sigma^2, \tau^2, \Sigma, d, n) = E\left\{||\nabla h(S)||^4 (\lambda_1 + 1)^6 \left(\frac{1}{n}||\epsilon||^2\right)^k\right\},$$
$$\eta_k = \eta_k(\sigma^2, \tau^2, \Sigma, d, n) = E\left\{\left|\left|\nabla^2 h(S)\right|\right|^4 (\lambda_1 + 1)^{12} \left(\frac{1}{n}||\epsilon||^2\right)^k\right\}.$$

*Remark* 1. If $||\Sigma||$ is bounded, then the asymptotic behavior of the upper bound (6) is determined by that of $\xi$, $\nu$, and $\psi^2$, which, in turn, is determined by the function $h$. For the functions $h$ considered in this paper, if $d/n \to \rho \in [0, \infty)$, then $\xi$, $\nu$, and $n\psi^2$ are bounded by rational functions in $\sigma^2$ and $\tau^2$. Thus, if $||\Sigma||$ is bounded, $d/n \to \rho \in [0, \infty)$ and $\sigma^2, \tau^2$ lie in some compact set, then we typically have $d_{TV}\{h(S), w\} = O(n^{-1/2})$. In other words, $h(S)$ converges to a normal random variable at rate $n^{-1/2}$. Under these conditions, if $\psi^2 = \text{var}\{h(S)\}$ is known or estimable, as it is for the $h$ studied here, then asymptotically valid confidence intervals for $E\{h(S)\}$ may be constructed using Theorem 1.

Now let $A$ be the matrix (5) and let $a_1^{\mathrm{T}}$, $a_2^{\mathrm{T}}$ denote the first and second rows of $A$, respectively. Then $\hat{\sigma}^2 = a_1^{\mathrm{T}} S$, $\hat{\tau}^2 = a_2^{\mathrm{T}} S$, and $\hat{r}^2 = \hat{\tau}^2/(\hat{\sigma}^2 + \hat{\tau}^2) = a_2^{\mathrm{T}} S/(a_1^{\mathrm{T}} S + a_2^{\mathrm{T}} S)$. Straightforward applications of Theorem 1 with $\Sigma = I$ and $h(S) = a_1^{\mathrm{T}} S = \hat{\sigma}^2$, $h(S) = a_2^{\mathrm{T}} S = \hat{\tau}^2$, and $h(S) = a_2^{\mathrm{T}} S/(a_1^{\mathrm{T}} S + a_2^{\mathrm{T}} S) = \hat{r}^2$ give bounds on the total variation distance between $\hat{\sigma}^2$, $\hat{\tau}^2$, and $\hat{r}^2$ and corresponding normal random variables; some additional care must be taken for $\hat{r}^2 = h(S) = a_2^{\mathrm{T}} S/(a_2^{\mathrm{T}} S + a_1^{\mathrm{T}} S)$ because it is undefined when $a_2^{\mathrm{T}} S + a_1^{\mathrm{T}} S = 0$. The asymptotic variance of the estimators follows from Lemma 2 and, in the case of $\hat{r}^2 = h(S) = a_2^{\mathrm{T}} S/(a_2^{\mathrm{T}} S + a_1^{\mathrm{T}} S)$, a basic Taylor expansion, i.e., the delta method. This is summarized in the following corollary to Theorem 1.

**Corollary 1.** *Suppose $\Sigma = I$ and $\sigma^2, \tau^2 \in D$ for some compact set $D \subseteq (0, \infty)$. Define*

$$\psi_1^2 = 2\left\{\frac{d}{n}(\sigma^2 + \tau^2)^2 + \sigma^4 + \tau^4\right\}, \tag{7}$$

$$\psi_2^2 = 2\left\{\left(1 + \frac{d}{n}\right)(\sigma^2 + \tau^2)^2 - \sigma^4 + 3\tau^4\right\}, \tag{8}$$

$$\psi_0^2 = \frac{2}{(\sigma^2 + \tau^2)^2}\left\{\left(1 + \frac{d}{n}\right)(\sigma^2 + \tau^2)^2 - \sigma^4\right\}. \tag{9}$$

*If $d/n \to \rho \in [0, \infty)$, then*

$$n^{1/2}\left(\frac{\hat{\sigma}^2 - \sigma^2}{\psi_1}\right), \quad n^{1/2}\left(\frac{\hat{\tau}^2 - \tau^2}{\psi_2}\right), \quad n^{1/2}\left(\frac{\hat{r}^2 - r^2}{\psi_0}\right) \to N(0, 1)$$

*in distribution.*

*Remark* 1. In Corollary 1, we require that $d/n \to \rho \in [0, \infty)$. By contrast, Lemma 2 only requires $d/n^2 \to 0$ in order to ensure consistency. It may be of interest to investigate how much the conditions on $d, n$ in Corollary 1 can be relaxed, while still ensuring asymptotic normality.

## 4. Unknown $\Sigma$

### 4.1. Estimable $\Sigma$

In this subsection, we consider the case where $\Sigma$ is unknown, but a norm consistent estimator for $\Sigma$ is available. An estimator $\hat{\Sigma}$ for $\Sigma$ is norm consistent if $||\hat{\Sigma}-\Sigma|| \to 0$, where $||\hat{\Sigma}-\Sigma||$ is the operator norm of $\hat{\Sigma}-\Sigma$ and the convergence holds in some appropriate sense, e.g. convergence in probability or squared-mean. If $d/n \to \rho > 0$, then the sample covariance matrix $n^{-1}X^{\mathrm{T}}X$ is not norm-consistent for $\Sigma$; furthermore, in the absence of additional information about $\Sigma$, it is generally not possible to find a norm-consistent estimator for $\Sigma$. However, Bickel and Levina (2008), El Karoui (2008), Cai et al. (2010), and others have shown that for wide classes of matrices $\Sigma$, norm-consistent estimators are available when $d/n \to \rho > 0$. Accurate estimation of $\Sigma$ may also be possible in semi-supervised learning situations (Lafferty and Wasserman, 2008), where additional $x_i$'s ($i = n+1, \ldots, N$) are available, but the corresponding $y_i$'s are unobserved.

Suppose $\hat{\Sigma}$ is a positive definite estimator for $\Sigma$ and define the estimators

$$\hat{\sigma}^2(\hat{\Sigma}) = \frac{d+n+1}{n(n+1)}||y||^2 - \frac{1}{n(n+1)}||\hat{\Sigma}^{-1/2}X^{\mathrm{T}}y||^2,$$
$$\hat{\tau}^2(\hat{\Sigma}) = -\frac{d}{n(n+1)}||y||^2 + \frac{1}{n(n+1)}||\hat{\Sigma}^{-1/2}X^{\mathrm{T}}y||^2.$$

Then $\hat{\sigma}^2(\Sigma)$ and $\hat{\tau}^2(\Sigma)$ are the known-$\Sigma$ analogues of $\hat{\sigma}^2$ and $\hat{\tau}^2$, respectively. In particular, all of the results from Section 3 remain valid with $\hat{\sigma}^2(\Sigma)$ and $\hat{\tau}^2(\Sigma)$ in place of $\hat{\sigma}^2$ and $\hat{\tau}^2$. Since

$$\hat{\sigma}^2(\hat{\Sigma}) = \hat{\sigma}^2(\Sigma) + O\left\{\frac{1}{n^2}||\Sigma^{-1/2}X^{\mathrm{T}}y||^2||\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I||\right\}, \tag{10}$$
$$\hat{\tau}^2(\hat{\Sigma}) = \hat{\tau}^2(\Sigma) + O\left\{\frac{1}{n^2}||\Sigma^{-1/2}X^{\mathrm{T}}y||^2||\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I||\right\}, \tag{11}$$

we conclude that if $||\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I||$ is small, then asymptotic properties of $\hat{\sigma}^2(\hat{\Sigma})$ and $\hat{\tau}^2(\hat{\Sigma})$ are determined by those of $\hat{\sigma}^2(\Sigma)$ and $\hat{\tau}^2(\Sigma)$. The following proposition is a direct consequence of (10)–(11) and the results of Section 3.

**Proposition 1.** *Let $\hat{\Sigma}$ be a positive definite estimator for $\Sigma$ and suppose $||\Sigma||$, $||\Sigma^{-1}||$, $||\hat{\Sigma}||$, $||\hat{\Sigma}^{-1}|| = O_P(1)$.*

*(i) The plug-in estimators $\hat{\sigma}^2(\hat{\Sigma})$ and $\hat{\tau}^2(\hat{\Sigma})$ satisfy*

$$|\hat{\sigma}^2(\hat{\Sigma}) - \sigma^2|,\ |\hat{\tau}^2(\hat{\Sigma}) - \tau^2| = O_P\left[\left\{\left(\frac{d+n}{n^2}\right)^{1/2} + ||\hat{\Sigma}-\Sigma||\right\}(\sigma^2+\tau^2)\right].$$

*(ii) Let $\psi_1^2$, $\psi_2^2$, and $\psi_0^2$ be as defined in (7)-(9) and let $\hat{r}^2(\hat{\Sigma}) = \hat{\tau}^2(\hat{\Sigma})/\{\hat{\sigma}^2(\hat{\Sigma}) + \hat{\tau}^2(\hat{\Sigma})\}$. Suppose $\sigma^2, \tau^2 \in D$ for some compact set $D \subseteq (0, \infty)$. If $d/n \to \rho \in [0, \infty)$ and $||\hat{\Sigma} - \Sigma|| = o_P(n^{-1/2})$, then*

$$n^{1/2}\left\{\frac{\hat{\sigma}^2(\hat{\Sigma}) - \sigma^2}{\psi_1}\right\}, \quad n^{1/2}\left\{\frac{\hat{\tau}^2(\hat{\Sigma}) - \tau^2}{\psi_2}\right\}, \quad n^{1/2}\left\{\frac{\hat{r}^2(\hat{\Sigma}) - r^2}{\psi_0}\right\} \to N(0, 1)$$

*in distribution.*

*Remark* 1. Part (i) of Proposition 1 implies if $\sigma^2, \tau^2$ are bounded, $d = o(n^2)$ and $||\hat{\Sigma} - \Sigma|| = o_P(1)$, then $\hat{\sigma}^2(\hat{\Sigma})$ and $\hat{\tau}^2(\hat{\Sigma})$ are consistent for $\sigma^2$ and $\tau^2$, respectively.

*Remark* 2. If $||\hat{\Sigma} - \Sigma|| = o_P(n^{-1/2})$ and the other conditions of Proposition 1 are met, then $\hat{\sigma}^2(\hat{\Sigma})$, $\hat{\tau}^2(\hat{\Sigma})$, and $\hat{r}^2(\hat{\Sigma})$ are asymptotically normal with the same asymptotic variance as $\hat{\sigma}^2(\Sigma)$, $\hat{\tau}^2(\Sigma)$, and $\hat{r}^2(\Sigma)$, respectively. The condition $||\hat{\Sigma} - \Sigma|| = o_P(n^{-1/2})$ is quite strong. However, important classes of high-dimensional covariance matrices can be estimated at this rate, e.g., certain classes of Toeplitz matrices (Cai et al., 2013).

### *4.2. Non-estimable $\Sigma$*

If $\mathrm{cov}(x_i) = \Sigma$ is unknown and a norm-consistent estimator is unavailable, then it is more challenging to find good estimators for $\sigma^2$, $\tau^2$ when $d > n$. To indicate where the estimators $\hat{\sigma}^2$, $\hat{\tau}^2$ from Section 3 break down when $\Sigma$ is unknown, define $\tau_k^2 = \beta^{\mathrm{T}}\Sigma^k\beta$ and $m_k = d^{-1}\mathrm{tr}(\Sigma^k)$, $k = 0, 1, 2, \ldots$. Then $\tau^2 = \tau_1^2$. If $\Sigma = I$, then $\tau^2 = \tau_k^2$ and $m_k = 1$ for all $k = 0, 1, 2, \ldots$. On the other hand, if $\Sigma \neq I$, then typically $\tau_k \neq \tau_{k'}$ and $m_k \neq m_{k'}$ for $k \neq k'$. One easily checks that

$$\frac{1}{n}E(||y||^2) = \sigma^2 + \tau_1^2, \tag{12}$$

$$\frac{1}{n^2}E(||X^{\mathrm{T}}y||^2) = \frac{d}{n}m_1\sigma^2 + \frac{d}{n}m_1\tau_1^2 + \left(1 + \frac{1}{n}\right)\tau_2^2, \tag{13}$$

$$E(\hat{\sigma}^2) = \frac{d(1 - m_1) + n + 1}{n + 1}\sigma^2 + \frac{d(1 - m_1) + n + 1}{n + 1}\tau_1^2 - \tau_2^2,$$

$$E(\hat{\tau}^2) = \frac{d(m_1 - 1)}{n + 1}\sigma^2 + \frac{d(m_1 - 1)}{n + 1}\tau_1^2 + \tau_2^2.$$

Thus, if $\Sigma \neq I$, then $\hat{\sigma}^2$, $\hat{\tau}^2$ are no longer unbiased.

In this subsection we propose alternative estimators for $\sigma^2, \tau^2$ that are consistent and asymptotically normal, provided $\beta, \Sigma$ satisfy conditions that simplify the relationship between

the various $\tau_k$ and $m_k$. Indeed, consider the approximation

$$\tau_k^2 = \beta^{\mathrm{T}}\Sigma^k\beta \approx ||\beta||^2\frac{1}{d}\mathrm{tr}(\Sigma^k) = \tau_0^2 m_k \tag{14}$$

for positive integers $k$. While (14) does not hold in general, related conditions have been proposed elsewhere in the random matrix theory literature (Bai et al., 2007; Pan and Zhou, 2008). If $\Sigma = \nu^2 I$, where $\nu^2 > 0$ is a positive real number, then (14) is an equality. More broadly, if $d$ is large and $\Sigma$ is an independent orthogonally invariant random matrix with well-behaved eigenvalues, then (14) may be a reasonable approximation for any $\beta$ (Bai et al., 2007). Furthermore, Bai et al. (2007) point out that for any given $\Sigma$ there must exist some $\beta$ such that (14) is an equality; for instance, take $\beta = \bar{u}$, where $\bar{u} = d^{-1/2}(u_1 + \cdots + u_d)$ and $u_1, \ldots, u_d$ are orthonormal eigenvectors of $\Sigma$.

Now assume that (14) holds for $k = 1, 2$. The method of moments suggests that

$$\hat{m}_1 = \frac{1}{d}\mathrm{tr}\left(\frac{1}{n}X^{\mathrm{T}}X\right), \quad \hat{m}_2 = \frac{1}{d}\mathrm{tr}\left\{\left(\frac{1}{n}X^{\mathrm{T}}X\right)^2\right\} - \frac{1}{dn}\left\{\mathrm{tr}\left(\frac{1}{n}X^{\mathrm{T}}X\right)\right\}^2 \tag{15}$$

are reasonable estimators for $m_1$ and $m_2$. Combining this with (14) yields $\tau_k^2 \approx \tau^2\hat{m}_k/\hat{m}_1$ $(k = 1, 2)$ and, by (12)–(13),

$$\frac{1}{n}E\left(||y||^2\right) = \sigma^2 + \tau^2, \quad \frac{1}{n^2}E\left(||X^{\mathrm{T}}y||^2\right) \approx \frac{d}{n}\hat{m}_1\sigma^2 + \left\{\frac{d}{n}\hat{m}_1 + \left(1 + \frac{1}{n}\right)\frac{\hat{m}_2}{\hat{m}_1}\right\}\tau^2.$$

In particular, up to the accuracy of approximation, $n^{-1}E(||y||^2)$ and $n^{-2}E(||X^{\mathrm{T}}y||^2)$ are linear combinations of $\sigma^2$ and $\tau^2$, with coefficients determined by the known quantities $d$, $n$, $\hat{m}_1$, and $\hat{m}_2$. Thus, we may obtain approximately unbiased estimators of $\sigma^2$ and $\tau^2$ by taking linear combinations of $n^{-1}||y||^2$ and $n^{-2}||X^{\mathrm{T}}y||^2$, with coefficients determined by $d$, $n$, $\hat{m}_1$, and $\hat{m}_2$. Indeed, we define the estimators

$$\tilde{\sigma}^2 = \left\{1 + \frac{d\hat{m}_1^2}{(n+1)\hat{m}_2}\right\}\frac{1}{n}||y||^2 - \frac{\hat{m}_1}{n(n+1)\hat{m}_2}||X^{\mathrm{T}}y||^2,$$
$$\tilde{\tau}^2 = -\frac{d\hat{m}_1^2}{n(n+1)\hat{m}_2}||y||^2 + \frac{\hat{m}_1}{n(n+1)\hat{m}_2}||X^{\mathrm{T}}y||^2$$

where $\hat{m}_1$, $\hat{m}_2$ are given in (15).

Proposition 2 summarizes some asymptotic properties of $\tilde{\sigma}^2$ and $\tilde{\tau}^2$, which depend on the accuracy of the approximation (14). An outline of the proof may be found in the Supplementary Material.

**Proposition 2.** *Suppose $D \subseteq (0,\infty)$ is a compact set and $\sigma^2, \tau^2 \in D$. Suppose further that there exist constants $m_-, m_+ \in \mathbb{R}$ such that $0 < m_- < m_1, m_4 < m_+$ for all $d$. Let $\Delta_k = |\tau_1^2 - \tau_0^2 m_1| + \cdots + |\tau_k^2 - \tau_0^2 m_k|$ $(k = 1, 2, \ldots)$.*

*(i) If $0 < d/n < c$ for some constant $c \in \mathbb{R}$, then $|\tilde{\sigma}^2 - \sigma^2|,\ |\tilde{\tau}^2 - \tau^2| = O_P(n^{-1/2} + \Delta_2)$.*

*(ii) Let $\tilde{r}^2 = \tilde{\tau}^2/(\tilde{\sigma}^2 + \tilde{\tau}^2)$ and let*

$$\tilde{\psi}_1^2 = 2\left\{\left(\frac{dm_1^2}{nm_2} + \frac{m_1 m_3}{m_2^2} - 1\right)(\sigma^2+\tau^2)^2 + \left(2 - \frac{m_1 m_3}{m_2^2}\right)\sigma^4 + \frac{m_1 m_3}{m_2^2}\tau^4\right\},$$
$$\tilde{\psi}_2^2 = 2\left\{\left(\frac{dm_1^2}{nm_2} + \frac{m_1 m_3}{m_2^2}\right)(\sigma^2+\tau^2)^2 - \frac{m_1 m_3}{m_2^2}\sigma^4 + \left(2 + \frac{m_1 m_3}{m_2^2}\right)\tau^4\right\},$$
$$\tilde{\psi}_0^2 = \frac{2}{(\sigma^2+\tau^2)^2}\left\{\frac{dm_1^2}{nm_2}(\sigma^2+\tau^2)^2 + 2\frac{m_1 m_3}{m_2^2}(\sigma^2\tau^2 + \tau^4) - \tau^4\right\}.$$

*If $d/n \to \rho \in [0,\infty)$, $d \to \infty$, and $\Delta_3 = o(n^{-1/2})$, then*

$$n^{1/2}\left(\frac{\tilde{\sigma}^2 - \sigma^2}{\tilde{\psi}_1}\right),\ n^{1/2}\left(\frac{\tilde{\tau}^2 - \tau^2}{\tilde{\psi}_2}\right),\ n^{1/2}\left(\frac{\tilde{r}^2 - r^2}{\tilde{\psi}_0}\right) \to N(0,1)$$

*in distribution.*

*Remark* 1. The condition that $0 < m_- < m_1, m_4 < m_+ < \infty$ for all $d$ ensures that the first four moments of the empirical distribution of the eigenvalues of $\Sigma = \Sigma_d$ are well-behaved.

*Remark* 2. Proposition 2 (i) requires that $\Delta_2$ is small in order to ensure consistency; Proposition 2 (ii) requires that $\Delta_3$ is small. If $\Delta_2$, $\Delta_3$ are small, then (14) is a reasonable approximation for $k = 2, 3$.

*Remark* 3. The condition $\Delta_3 = o(n^{-1/2})$ in Proposition 2 (ii) is quite strong. For instance, if $\Sigma$ is a sample covariance matrix formed from independent $N(0, \sigma_0^2)$ data with a constant aspect ratio, then $\Delta_2 = o(1)$, but $\Delta_3 \neq o(n^{-1/2})$. On the other hand, if $\Sigma$ is a constant multiple of the identity matrix, then $\Delta_3 = o(n^{-1/2})$.

*Remark* 4. If $\Sigma = I$, then $m_1 = m_2 = m_3 = 1$ and $\tilde{\psi}_j^2 = \psi_j^2$, $j = 0, 1, 2$, where $\psi_j^2$ are given in (7)–(9). In other words, if $\Sigma = I$, then the asymptotic variance of $\tilde{\sigma}^2$, $\tilde{\tau}^2$, and $\tilde{r}^2$ is the same as that of $\hat{\sigma}^2$, $\hat{\tau}^2$, and $\hat{r}^2$, respectively. This is driven by the fact that if $d/n \to \rho \in [0,\infty)$ and $d \to \infty$, then $|\hat{m}_k - m_k|$ converges at a rate faster than $n^{-1/2}$. This also highlights the necessity of the condition $d \to \infty$ in Proposition 2 (ii): if $d$ is bounded, then $|\hat{m}_k - m_k| = O(n^{-1/2})$ and this may affect the asymptotic distribution of the estimators.

*Remark* 5. In order to construct confidence intervals or conduct Wald tests based on Proposition 2 (ii), an estimate of $m_3$ is required. The method of moments and Proposition S1 in

the Supplementary Material suggest the estimator

$$\hat{m}_3 = \frac{1}{d}\text{tr}\left\{\left(\frac{1}{n}X^{\text{T}}X\right)^3\right\} + \frac{2}{dn^2}\left\{\text{tr}\left(\frac{1}{n}X^{\text{T}}X\right)\right\}^3 - \frac{3}{dn}\text{tr}\left(\frac{1}{n}X^{\text{T}}X\right)\text{tr}\left\{\left(\frac{1}{n}X^{\text{T}}X\right)^2\right\}. \quad (16)$$

## 5. Simulation studies

### 5.1. Example 1

In this example, we investigated basic properties of some of the estimators proposed above. We fixed $d = 500$ and considered settings with $n = 250, 500$. The predictors $x_1, \ldots, x_n \in \mathbb{R}^d$ were generated according to one of three distributions. In the first setting, $x_1, \ldots, x_n \sim N(0, I)$. In the second setting, we generated a $(2d) \times d$ random matrix $Z$ with independent $N(0,1)$ entries and took $\Sigma_w = (2d)^{-1}Z^{\text{T}}Z$; the predictors $x_1, \ldots, x_n$ were then generated according to a $N(0, \Sigma_w)$ distribution. In this predictor setting, a single matrix $\Sigma_w$ was used to generate all of the simulated datasets, i.e., $\Sigma_w$ was generated only once. In the third predictor setting, the individual $x_{ij}$ $(i = 1, \ldots, n;\ j = 1, \ldots, d)$ were independent random variables taking values in $\{\pm 1\}$ with $\text{pr}(x_{ij} = 1) = \text{pr}(x_{ij} = -1) = 0.5$.

To generate the parameter $\beta \in \mathbb{R}^d$, we created a $d$-dimensional vector with the first $d/2$ coordinates independent uniform$(0,1)$ and the remaining $d/2$ coordinates independent $N(0,1)$; $\beta$ was obtained by standardizing this vector so that $\beta^{\text{T}}\Sigma\beta = \tau^2 = 1$. Thus, $\beta$ corresponding to the settings where $\text{cov}(x_i) = I$ is scaled slightly differently from $\beta$ corresponding to $\text{cov}(x_i) = \Sigma \neq I$. The residual variance was fixed at $\sigma^2 = 1$.

For each setting in this example, we generated 1000 independent datasets and computed the estimators $\hat{\sigma}^2 = \hat{\sigma}^2(I)$, $\hat{\tau}^2 = \hat{\tau}^2(I)$, $\hat{r}^2 = \hat{r}^2(I)$ and $\tilde{\sigma}^2$, $\tilde{\tau}^2$, $\tilde{r}^2$ for each dataset. Recall that the estimators $\hat{\sigma}^2$, $\hat{\tau}^2$, $\hat{r}^2$, from Section 3.1, were derived under the assumption that $x_i \sim N(0, I)$; the estimators $\tilde{\sigma}^2$, $\tilde{\tau}^2$, $\tilde{r}^2$, from Section 4.2, were derived under the assumption that (14) is a reasonable approximation. Summary statistics for the various estimators are reported in Table 1.

Table 1 indicates that for $x_i \sim N(0, I)$, all of the methods perform as expected, given the theoretical results developed above: the estimators are essentially unbiased, and the standard errors match those predicted in Corollary 1 and Proposition 2. For $x_i \sim N(0, \Sigma_w)$, Table 1 indicates that $\tilde{\sigma}^2$, $\tilde{\tau}^2$, and $\tilde{r}^2$ are approximately unbiased and their standard errors are comparable to the setting where $x_i \sim N(0, I)$. On the other hand, $\hat{\sigma}^2(I)$, $\hat{\tau}^2(I)$, and $\hat{r}^2(I)$ are badly biased when $x_i \sim N(0, \Sigma_w)$; this is not unexpected, given the discussion in Section 4.2. While this paper contains no theoretical results describing the behavior of our estimators for non-normal data, the numerical results in this example suggest that the methods proposed here might be successfully applied in broader circumstances. Indeed, the results in Table 1

TABLE 1

*Summary statistics for Example 1; $d = 500$. Means and standard errors of various estimators, computed over 1000 independent datasets for each configuration. In each setting, $\sigma^2 = \tau^2 = 1$ and $r^2 = \tau^2/(\sigma^2 + \tau^2) = 0.5$. In the standard error column corresponding to $x_i \sim N(0, I)$, numbers in parentheses are theoretically predicted standard errors, which are denoted $\psi_1$, $\psi_2$, and $\psi_0$ in the text; see Corollary 1 and Proposition 2. Theoretically predicted standard errors for $x_i \sim N(0, \Sigma_w)$ and $x_i \in \{\pm 1\}$ binary are not known.*

| | | $x_i \sim N(0, I)$ | | $x_i \sim N(0, \Sigma_w)$ | | $x_i \in \{\pm 1\}$ binary | |
|---|---|---|---|---|---|---|---|
| Estimator | $n$ | Mean | Std. Error | Mean | Std. Error | Mean | Std. Error |
| $\hat{\sigma}^2(I)$ | 250 | 0.99 | 0.29 (0.28) | 0.50 | 0.41 | 1.00 | 0.29 |
| | 500 | 1.00 | 0.16 (0.15) | 0.51 | 0.24 | 1.00 | 0.15 |
| $\tilde{\sigma}^2$ | 250 | 1.00 | 0.28 (0.28) | 1.01 | 0.26 | 1.00 | 0.29 |
| | 500 | 1.00 | 0.16 (0.15) | 1.01 | 0.15 | 1.00 | 0.15 |
| $\hat{\tau}^2(I)$ | 250 | 1.00 | 0.33 (0.33) | 1.50 | 0.48 | 0.99 | 0.34 |
| | 500 | 1.00 | 0.20 (0.20) | 1.48 | 0.29 | 0.99 | 0.19 |
| $\tilde{\tau}^2$ | 250 | 1.00 | 0.33 (0.33) | 1.00 | 0.32 | 1.00 | 0.34 |
| | 500 | 1.00 | 0.20 (0.20) | 0.98 | 0.19 | 1.00 | 0.19 |
| $\hat{r}^2(I)$ | 250 | 0.50 | 0.15 (0.15) | 0.75 | 0.21 | 0.49 | 0.15 |
| | 500 | 0.50 | 0.08 (0.08) | 0.74 | 0.12 | 0.50 | 0.08 |
| $\tilde{r}^2$ | 250 | 0.50 | 0.15 (0.15) | 0.49 | 0.14 | 0.49 | 0.15 |
| | 500 | 0.50 | 0.08 (0.08) | 0.49 | 0.08 | 0.50 | 0.08 |

for $x_i \in \{\pm 1\}$ binary shows that all of the estimators are nearly unbiased and have standard errors that are similar to the corresponding standard errors in the case where $x_i \sim N(0, I)$.

One of the more striking aspects of Table 1 is the accuracy and robustness of the estimators $\tilde{\sigma}^2$, $\tilde{\tau}^2$, and $\tilde{r}^2$. Proposition 2 suggests that these estimators might be expected to perform well when $x_i \sim N(0, I)$ and $x_i \sim N(0, \Sigma_w)$; however, none of our theoretical results apply to the case where $x_i \in \{\pm 1\}$ is binary.

The quantities $\sigma^2, \tau^2 \geq 0$ are all non-negative. However, there is no guarantee that the estimators proposed in this paper are non-negative. Two factors contributing to negative estimates of $\sigma^2$ are $\tau^2$ are (i) random fluctuations in the data and (ii) significant violations of assumptions about $\mathrm{cov}(x_i) = \Sigma$. For most of the estimators and settings considered in this example, the percentage of datasets with negative estimates of $\tau^2$ or $\sigma^2$ was less than 1%; in these instances, it appears that negative estimates were largely explained by random fluctuations in the data and that one could still reasonably appeal to asymptotic results in order to construct confidence intervals for $\sigma^2$ and $\tau^2$, for instance. The major exception involved the estimate $\hat{\sigma}^2(I)$ in cases where $x_i \sim N(0, \Sigma_w)$. The estimators $\hat{\sigma}^2(I)$, $\hat{\tau}^2(I)$, and $\hat{r}^2(I)$ were derived under the assumption that $\mathrm{cov}(x_i) = I$ and, as noted above, are significantly biased when $x_i \sim N(0, \Sigma_w)$; in particular, $\hat{\sigma}^2(I)$ is biased towards 0, which led to a substantially higher fraction of estimates $\hat{\sigma}^2(I) < 0$. Indeed, $\hat{\sigma}^2(I) < 0$ in 11% of the datasets with $x_i \sim N(0, \Sigma_w)$ and $n = 250$. An even more extreme example is considered in Example

2 below, where the mean value of an estimator for $\sigma^2$ is negative. In these settings it seems challenging to make valid inferences about $\sigma^2$ and $\tau^2$. These observations further highlight the importance of understanding and validating the assumptions underlying the proposed methods.

### 5.2. Example 2

Sun and Zhang (2012) proposed scaled lasso and scaled minimax concave penalty methods for estimating $\sigma^2$ in high-dimensional linear models. These methods, which simultaneously estimate $\sigma^2$ and $\beta$, are very effective when $\beta$ is sparse. Let $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$ denote the scaled lasso and scaled minimax concave penalty estimators for $\sigma^2$, respectively. In this example, we compare the performance of $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$ with some of the estimators for $\sigma^2$ proposed in this paper, in settings where $\beta$ is both sparse and non-sparse, i.e., dense.

With $d = 3000$, the predictors in this example were generated according to $x_i \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})$ and $\sigma_{ij} = 0.5^{|i-j|}$. We fixed $\sigma^2 = 1$. Sparse and dense parameters $\beta \in \mathbb{R}^d$ were generated as follows. First, to generate the sparse $\beta$, five random multiples of 25 between 25 and $d - 25 = 2975$ were selected. That is, we selected $k_1, \ldots, k_5$ from $\{25, 50, 75 \ldots, 2975\}$ uniformly at random. Next, we took $\beta_0 \in \mathbb{R}^d$ to be the vector with the 7-dimensional subvector $(1, 2, 3, 4, 3, 2, 1)$ centered at the coordinates corresponding to $k_1, \ldots, k_5$, so that the $k_j$-th entry of $\beta_0$ was 4, the $(k_j \pm 1)$-th was 3, etc.; the remaining entries in $\beta_0$ were set equal to 0. We then set $\beta = \{3/(\beta_0^{\mathrm{T}}\Sigma\beta_0)\}^{1/2}\beta_0$, so that $\tau^2 = \beta^{\mathrm{T}}\Sigma\beta = 3$. This sparse $\beta$ was generated only once; in other words, the same sparse $\beta$ was use throughout the simulations in this example. To generate the dense $\beta$ used in this example, we followed the same procedure as for the sparse $\beta$, except that in $\beta_0$, the 7-dimensional subvector $(1, 2, 3, 4, 3, 2, 1)$ was centered at coordinates corresponding to each multiple of 25 between 25 and 2975. Thus, for the sparse $\beta$, we had $||\beta||_0 = 7 \times 5 = 35$, where $||\beta||_0$ denotes the number of non-zero coordinates in $\beta$, and for the dense $\beta$ we had $||\beta||_0 = 7 \times (d/25 - 1) = 833$; however, $\tau^2 = \beta^{\mathrm{T}}\Sigma\beta = 3$ was the same for both the sparse and dense $\beta$. In this simulation study, we considered datasets with $n = 600$ and $n = 2400$ observations. With sparse $\beta$ and $n = 600$, the simulation settings in this example are very similar to those in Example 2 from Section 4.1 of Sun and Zhang (2012).

Under each of the settings described above, we generated 100 independent datasets and, for each simulated dataset, we computed $\hat{\sigma}^2_{\text{lasso}}$, $\hat{\sigma}^2_{\text{MCP}}$, $\hat{\sigma}^2(\hat{\Sigma})$, $\hat{\sigma}^2(\Sigma)$, and $\tilde{\sigma}^2$. For the $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$, we used the shrinkage parameter $\lambda_0 = \{\log(d)/n\}^{1/2}$; this value of $\lambda_0$ yielded the best performance in the numerical examples described by Sun and Zhang (2012). The estimator $\hat{\sigma}_{\text{MCP}}$ requires specification of an additional parameter $\gamma$; following Sun and Zhang (2012), we took $\gamma = 2/[1 - \max_{i,j}\{X_i^{\mathrm{T}}X_j/(||X_i||||X_j||)\}]$, where $X_j$ denotes the $j$-th column of $X$. The estimator $\hat{\sigma}^2(\hat{\Sigma})$ was introduced in Section 4.1 of this paper. Here we take advantage of the AR(1)

TABLE 2

*Example 2; $d = 3000$, $\sigma^2 = 1$. Means and standard errors of estimators for $\sigma^2$, based on 100 independent datasets. Left, sparse $\beta$; right, dense $\beta$*

| Sparse $\beta$ | | Mean | Std. Err. |
|---|---|---|---|
| $n = 600$ | $\hat{\sigma}^2_{\text{lasso}}$ | 1.11 | 0.07 |
| | $\hat{\sigma}^2_{\text{MCP}}$ | 1.05 | 0.06 |
| | $\hat{\sigma}^2(\hat{\Sigma})$ | 0.97 | 0.50 |
| | $\hat{\sigma}^2(\Sigma)$ | 0.97 | 0.50 |
| | $\tilde{\sigma}^2$ | −0.60 | 0.52 |
| $n = 2400$ | $\hat{\sigma}^2_{\text{lasso}}$ | 1.03 | 0.03 |
| | $\hat{\sigma}^2_{\text{MCP}}$ | 1.01 | 0.03 |
| | $\hat{\sigma}^2(\hat{\Sigma})$ | 0.98 | 0.16 |
| | $\hat{\sigma}^2(\Sigma)$ | 0.98 | 0.16 |
| | $\tilde{\sigma}^2$ | −0.59 | 0.21 |

| Dense $\beta$ | | Mean | Std. Err. |
|---|---|---|---|
| $n = 600$ | $\hat{\sigma}^2_{\text{lasso}}$ | 3.26 | 0.21 |
| | $\hat{\sigma}^2_{\text{MCP}}$ | 3.10 | 0.21 |
| | $\hat{\sigma}^2(\hat{\Sigma})$ | 0.98 | 0.56 |
| | $\hat{\sigma}^2(\Sigma)$ | 0.98 | 0.56 |
| | $\tilde{\sigma}^2$ | −0.57 | 0.59 |
| $n = 2400$ | $\hat{\sigma}^2_{\text{lasso}}$ | 2.32 | 0.07 |
| | $\hat{\sigma}^2_{\text{MCP}}$ | 2.00 | 0.08 |
| | $\hat{\sigma}^2(\hat{\Sigma})$ | 1.01 | 0.15 |
| | $\hat{\sigma}^2(\Sigma)$ | 1.01 | 0.15 |
| | $\tilde{\sigma}^2$ | −0.57 | 0.22 |

structure of $\Sigma$ and set $\hat{\Sigma} = (\hat{\sigma}_{ij})$, where $\hat{\sigma}_{i,j} = \hat{\alpha}^{|i-j|}$ and $\hat{\alpha} = \{n(d-1)\}^{-1}\sum_{i=1}^{n}\sum_{j=2}^{d} x_{ij}x_{i(j-1)}$. We view the estimator $\hat{\sigma}^2(\Sigma)$ as an oracle estimator, which utilizes full knowledge of actual covariance matrix $\Sigma$; following the discussion in Section 3.1, this estimator should perform similarly to the estimator $\hat{\sigma}^2(I)$ in settings where $\text{cov}(x_i) = I$ and $\tau^2 = 3$. Finally, the estimator $\tilde{\sigma}^2$ is the unknown covariance estimator from Section 4.2. Recall that the theoretical performance guarantees for $\tilde{\sigma}^2$ given in Proposition 2 require $|\beta^{\text{T}}\Sigma^k\beta - ||\beta||^2\text{tr}(\Sigma^k)/d| \approx 0$, for $k = 1, 2$. In this example, for the sparse $\beta$ we had

$$||\beta||^2\frac{1}{d}\text{tr}(\Sigma) - \beta^{\text{T}}\Sigma\beta = -1.76, \quad ||\beta||^2\frac{1}{d}\text{tr}(\Sigma^2) - \beta^{\text{T}}\Sigma^2\beta = -5.54; \tag{17}$$

the corresponding quantities are essentially the same for the dense $\beta$. Summary statistics for the various estimators computed in this numerical study are reported in Table 2.

For sparse $\beta$, the results in Table 2 indicate that $\hat{\sigma}^2_{\text{lasso}}$, $\hat{\sigma}^2_{\text{MCP}}$, $\hat{\sigma}^2(\hat{\Sigma})$ and $\hat{\sigma}^2(\Sigma)$ are all nearly unbiased. However, the empirical standard errors for the scaled lasso and minimax concave penalty estimators are considerably smaller than the standard errors for $\hat{\sigma}^2(\hat{\Sigma})$ and $\hat{\sigma}^2(\Sigma)$. In this example, the performance of $\hat{\sigma}^2(\hat{\Sigma})$ is very similar to that of the oracle estimator $\hat{\sigma}^2(\Sigma)$.

The estimator $\tilde{\sigma}^2$ is significantly biased in this example. Indeed, the mean value of $\tilde{\sigma}^2$ is negative, while $\sigma^2 > 0$. The poor performance of $\tilde{\sigma}^2$ in this example is not completely unexpected, given that $|\beta^{\text{T}}\Sigma^k\beta - ||\beta||^2\text{tr}(\Sigma^k)/d|$ is substantially larger than 0 for $k = 1, 2$; see (17). In fact, more can be said. Using the approximation $\hat{m}_k \approx m_k = \text{tr}(\Sigma^k)/d$, one can check that $E(\tilde{\sigma}^2) \approx \sigma^2 + \tau_1^2 - (m_1/m_2)\tau_2^2$. In this example, $\tau_1^2 - (m_1/m_2)\tau_2^2 = -1.57$ and, by the previous approximation, $E(\tilde{\sigma}^2) \approx -0.57$; this calculation is for the sparse $\beta$, but the result is almost exactly the same for the dense $\beta$. Note the similarity between this approximation and the empirical means of $\tilde{\sigma}^2$ in Table 2.

TABLE 3

*Example 2; $d = 3000$, $\sigma^2 = 1$, $||\beta||^2 = 1.24$. Empirical mean squared error $||\hat{\beta} - \beta||^2$ of the scaled lasso and minimax concave penalty estimators for $\beta$, based on 100 independent datasets.*

| Sparse $\beta$ | | |
|---|---|---|
| $n$ | lasso | MCP |
| 600 | 0.19 | 0.37 |
| 2400 | 0.05 | 0.09 |

| Dense $\beta$ | | |
|---|---|---|
| $n$ | lasso | MCP |
| 600 | 1.22 | 1.25 |
| 2400 | 0.90 | 0.93 |

For dense $\beta$, the performance of $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$ breaks down, while the performance of $\hat{\sigma}^2(\hat{\Sigma})$, $\hat{\sigma}^2(\Sigma)$ and $\tilde{\sigma}^2$ remains virtually unchanged, as compared to the sparse $\beta$ case. When $n = 600$, the empirical means of $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$ are both greater than 3; when $n = 2400$, the empirical means of $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$ are both nearly greater than 2. Both $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$ depend on associated lasso and minimax concave penalty estimators for $\beta$. The performance breakdown of $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$ when $\beta$ is dense is likely related to the fact that the corresponding estimators for $\beta$ perform poorly when $\beta$ is dense and $d/n$ is large. Indeed, in Table 3 we report the empirical mean squared error for the lasso and minimax concave penalty estimators for $\beta$ that are associated with $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{MCP}}$; mean squared error is substantially higher for estimating dense $\beta$.

The results of this simulation study suggest that estimators proposed in this paper may be useful for estimating $\sigma^2$ in settings where $d/n$ is large and little is known about sparsity in $\beta$. However, we emphasize two important points: the predictor covariance must be handled appropriately, which, in this example, means utilizing the fact that $\Sigma$ has AR(1) structure; and the estimators for $\sigma^2$ proposed in this paper may have larger standard error than estimators derived from a reliable estimate of $\beta$.

## 6. Data analysis

A series of studies have investigated associations between array-based gene expression data and single nucleotide polymorphisms from individuals in the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/). One of the major hypotheses validated by Stranger et al. (2007a, 2012, 2007b) and others is that single nucleotide polymorphisms located near a given gene may be associated with gene expression levels. These findings may have significant implications for understanding disease susceptibility, treatment, and public health. In this analysis, we use methods proposed in this paper to estimate the proportion of variation in gene expression levels that is explained by single nucleotide polymorphisms; that is, we estimate $r^2$. In the present context, estimating $r^2$ is closely related to estimating heritability, an important concept in genetics (Yang et al., 2010).

In Stranger et al. (2007b), the authors identified 341 genes that were significantly associated

with single nucleotide polymorphisms from 45 individuals in the Japanese HapMap population. We randomly selected 100 of these genes and, using a more recent gene expression dataset made publicly available by Stranger et al. (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264/), we investigated associations between these 100 genes and nearby single nucleotide polymorphisms from $n = 80$ individuals in the Han Chinese HapMap population. Single nucleotide polymorphism data was obtained from the International HapMap Project, release 28. For each gene, we restricted our attention to single nucleotide polymorphisms located within 1 Mbp of the gene midpoint and with minor allele frequency $> 5\%$. Single nucleotide polymorphisms with missingness $> 10\%$ in the Han Chinese population were excluded from our analysis; for single nucleotide polymorphisms included in the analysis, missing values, i.e., minor allele counts, were imputed using the marginal mean from the Han Chinese population.

For each gene, we sought to estimate $r^2 = \tau^2/(\sigma^2 + \tau^2)$ for regressing the gene expression level on the minor allele counts of single nucleotide polymorphisms within 1 Mbp of the gene midpoint. Some preprocessing of the data was required in order to ensure appropriate application of the methods proposed in this paper. First, we centered the gene expression levels and single nucleotide polymorphism minor allele counts, so that each variable had mean 0. Next, we standardized the single nucleotide polymorphisms minor allele counts so that each allele count variable had standard deviation 1. As discussed at length above, correlation between the predictors is a key issue for the methods proposed in this paper. In this analysis, we de-correlated the predictors by simply removing highly correlated single nucleotide polymorphisms; we refer to this as thinning the predictors. To explain in more detail, Jiang (2004) proved that if $\mathrm{cov}(x_i) = I$, then the maximum absolute sample correlation between the predictors is approximately $2\{\log(n)/n\}^{1/2}$; thus, for each pair of single nucleotide polymorphism allele count variables with absolute sample correlation greater than $2\{\log(n)/n\}^{1/2}$, we removed one of them, chosen at random, from the dataset. After this thinning process, the maximum absolute correlation between single nucleotide polymorphisms was no more than that predicted for independent single nucleotide polymorphisms and, by this measure, the thinned single nucleotide polymorphisms appeared to be roughly independent. We subsequently applied the methods proposed in this paper to the standardized and thinned dataset. The thinning process described above is practical, but fairly crude. It may be desirable to seek other methods for handling correlation among the single nucleotide polymorphisms, e.g., estimating the correlation between single nucleotide polymorphisms.

Let $\mathrm{Gene}_{ik}$ denote the centered expression level for the $k$th gene ($k = 1, \ldots, 100$) in the $i$th individual ($i = 1, \ldots, n = 80$) and let $\mathrm{SNP}_{ijk}$ denote the standardized minor allele count for the $j$th single nucleotide polymorphism corresponding to the $k$th gene in individual $i$; so $j = 1, \ldots, d_k$, where $d_k$ is the number of thinned single nucleotide polymorphisms within 1 Mbp of gene $k$. For the genes included in this analysis, $d_k$ ranged from 17 to 190. The mean

TABLE 4

*Significant genes at false discovery rate 0.05 for testing $H_0 : r^2 = 0$. $d_k$ is the number of single nucleotide polymorphisms included in the analysis for the corresponding gene; p-values were computed using the Wald test suggested by Proposition 2 (ii); q-values refer to false discovery rate for the Benjamini-Hochberg step-up procedure (Benjamini and Hochberg, 1995; Storey, 2002). Recall that $n = 80$ individuals were included in this analysis.*

| Gene | $d_k$ | $\tilde{r}^2$ | $p$-value ($\times 10^2$) | $q$-value ($\times 10^2$) |
|---|---|---|---|---|
| HLA-DRB5 | 102 | 0.95 | 0.01 | 0.59 |
| MRPL43 | 56 | 0.79 | 0.01 | 0.59 |
| WBSCR27 | 43 | 0.71 | 0.02 | 0.71 |
| PEX6 | 112 | 0.78 | 0.06 | 1.35 |
| LOC197322 | 95 | 0.74 | 0.07 | 1.35 |
| FLJ10781 | 110 | 0.73 | 0.08 | 1.35 |
| TSGA10 | 38 | 0.59 | 0.10 | 1.42 |
| NQO2 | 127 | 0.73 | 0.12 | 1.49 |
| CHPT1 | 74 | 0.61 | 0.20 | 2.20 |
| CCNDBP1 | 24 | 0.45 | 0.25 | 2.53 |
| FN3KRP | 48 | 0.53 | 0.29 | 2.61 |
| SLC2A8 | 71 | 0.54 | 0.36 | 3.02 |
| IRF5 | 62 | 0.50 | 0.54 | 4.01 |
| UGT2B17 | 32 | 0.42 | 0.56 | 4.01 |
| UTS2 | 100 | 0.55 | 0.63 | 4.22 |

value of $d_k$ was 82.6; recall that $n = 80$, so that on average $d_k > n$. We used the estimator $\tilde{r}^2$ to estimate $r^2$ for the model

$$\text{Gene}_{ik} = \sum_{j=1}^{d_k} \text{SNP}_{ijk}\beta_{jk} + \epsilon_{ik}, \quad k = 1, \ldots, 100.$$

For each gene, we also computed $p$-values for the null hypothesis $H_0 : r^2 = 0$, using the Wald test suggested by Proposition 2 (ii). Genes significant at false discovery rate 0.05 using the Benjamini and Hochberg (1995) step-up procedure are reported in Table 4; more complete results may be found in Table S1 of the Supplementary Material.

Overall, 26 out of 100 genes were significant at false discovery rate 0.1 and 15 were significant at false discovery rate 0.05. Among genes significant at false discovery rate 0.05, estimates of $r^2$ range from 0.42 to 0.95, while $p$-values ranged from 0.0001 to 0.0056. Confidence intervals for $r^2$ corresponding to each of the reported estimates can be easily constructed using Propostion 2 (ii).

We note that 17 genes in this analysis had $p$-value smaller than 0.01, which far exceeds the 1 that would be expected if $r^2 = 0$ for each of the genes under consideration. Thus, since our initial list of 100 genes consisted of genes found to be significant in the Japanese HapMap population (Stranger et al., 2007b), this analysis partially validates Stranger et al.'s findings in a different HapMap population, Han Chinese. However, not all of Stranger et al.'s findings were validated in our analysis, i.e., not all of the genes under consideration were found to be significant. While there are many potential explanations for this discrepancy, e.g., random variation in the data or differences between individuals in the Han Chinese and Japanese

populations, it raises questions about the power of the proposed Wald tests and the efficiency of the estimator $\tilde{r}^2$ that are of interest for future research; questions about the thinning process for handling correlated predictors and its effect on the power of the proposed tests may also be of interest. On the other hand, we emphasize that our analysis goes beyond just identifying significant genes and provides estimates of the fraction of variability in gene expression levels that is explained by single nucleotide polymorphisms, even in settings where there are more single nucleotide polymorphisms than observations, i.e., $d_k > n$.

## Acknowledgements

The author is grateful to the Editor, the Associate Editor, and three reviewers for many insightful comments and suggestions that helped to substantially improve the paper. Thanks to Xihong Lin, Liming Liang, and Sihai Zhao for very helpful suggestions regarding the data analysis. This work was supported by the U.S. National Science Foundation.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE T. Automat. Contr.* **19** 716–723.

Bai, Z. D., Miao, B. Q. and Pan, G. M. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.* **35** 1532–1572.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57** 289–300.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Stat.* **36** 199–227.

Cai, T. T., Ren, Z. and Zhou, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Rel.* **156** 101–143.

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Stat.* **38** 2118–2144.

Chatterjee, S. (2009). Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Rel.* **143** 1–40.

Dicker, L. H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electron. J. Stat.* **7** 1806–1834.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Stat.* **36** 2717–2756.

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

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Stat.* 1947–1975.

Graczyk, P., Letac, G. and Massam, H. (2005). The hyperoctahedral group, symmetric group representations and the moments of the real Wishart distribution. *J. Theor. Prob.* **18** 1–42.

Jiang, T. (2004). The asymptotic distributions of the largest entries of sample correlation matrices. *Ann. Appl. Probab.* **14** 865–880.

Lafferty, J. and Wasserman, L. (2008). Statistical analysis of semi-supervised regression. *Adv. Neur. In.* **20** 801–808.

Letac, G. and Massam, H. (2004). All invariant moments of the Wishart distribution. *Scand. J. Stat.* **31** 295–318.

Pan, G. M. and Zhou, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.* **18** 1232–1270.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6** 461–464.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. B* **64** 479–498.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., De Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E. and Dermitzakis, E. T. (2007a). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315** 848–853.

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., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P. and Dermitzakis, E. T. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* **8** e1002639.

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

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

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. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.

Zou, H., Hastie, T. J. and Tibshirani, R. J. (2007). On the "degrees of freedom" of the lasso. *Ann. Stat.* **35** 2173–2192.

# Supplementary Material

## S1. Additional results for Section 6 of the main text (data analysis)

Results for all 100 genes included in the data analysis may be found in Table S1.

## S2. Results from the main text

### *Proof of Lemma 2*

Lemma 2 follows from a corollary of the next result.

**Lemma S1.** *Suppose that* $\Sigma = I$. *Then*

$$\text{var}\left(\frac{1}{n}||y||^2\right) = \frac{2}{n}(\sigma^2+\tau^2)^2 \tag{18}$$

$$\begin{aligned}\text{var}\left(\frac{1}{n^2}||X^{\mathrm{T}}y||^2\right) = \frac{2}{n}\Bigg[&\left\{\left(\frac{d}{n}\right)^2 + \frac{d}{n} + \frac{2d}{n^2}\right\}\sigma^4 \\ &+ \left\{2\left(\frac{d}{n}\right)^2 + \frac{6d}{n} + 2 + \frac{10d}{n^2} + \frac{10}{n} + \frac{12}{n^2}\right\}\sigma^2\tau^2 \\ &+ \left\{\left(\frac{d}{n}\right)^2 + \frac{5d}{n} + 4 + \frac{8d}{n^2} + \frac{15}{n} + \frac{15}{n^2}\right\}\tau^4\Bigg]\end{aligned} \tag{19}$$

$$\text{cov}\left(\frac{1}{n}||y||^2, \frac{1}{n^2}||X^{\mathrm{T}}y||^2\right) = \frac{2}{n}\left\{\frac{d}{n}\sigma^4 + \left(\frac{2d}{n} + 2 + \frac{3}{n}\right)\sigma^2\tau^2 + \left(\frac{d}{n} + 2 + \frac{3}{n}\right)\tau^4\right\}. \tag{20}$$

*Proof.* Equation (18) is obvious because $||y||^2 \sim (\sigma^2+\tau^2)\chi^2_n$. To prove (19), we condition on $X$ and use properties of expectations involving quadratic forms and normal random vectors to obtain

$$\begin{aligned}\text{var}(||X^{\mathrm{T}}y||^2) &= E\left\{\text{var}(||X^{\mathrm{T}}y||^2 \mid X)\right\} + \text{var}\left\{E(||X^{\mathrm{T}}y||^2 \mid X)\right\} \\ &= 2\sigma^4 E\left[\text{tr}\left\{(X^{\mathrm{T}}X)^2\right\}\right] + 4\sigma^2 E\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^3\beta\right\} \\ &\quad + \text{var}\left\{\sigma^2\text{tr}(X^{\mathrm{T}}X) + \beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\} \\ &= 2\sigma^4 E\left[\text{tr}\left\{(X^{\mathrm{T}}X)^2\right\}\right] + 4\sigma^2 E\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^3\beta\right\} + \sigma^4 E\left[\left\{\text{tr}(X^{\mathrm{T}}X)\right\}^2\right] \\ &\quad + 2\sigma^2 E\left\{\text{tr}(X^{\mathrm{T}}X)\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\} + E\left[\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\}^2\right]\end{aligned}$$

TABLE S1

*Data analysis. $d_k$ is the number of SNPs included in the analysis for the corresponding gene (recall that $n = 80$ for each gene); p-values (for testing $H_0 : r^2 = 0$) were computed using the Wald test suggested by Proposition 2 (ii) of the main text; q-values refer to false discover rate for the Benjamini-Hochberg step-up procedure (Benjamini and Hochberg, 1995; Storey, 2002).*

| Gene | Probe | $d_k$ | $\tilde{r}^2$ | $p$-value ($\times 10^2$) | $q$-value ($\times 10^2$) |
|---|---|---|---|---|---|
| HLA-DRB5 | ILMN_1697499 | 102 | 0.95 | 0.01 | 0.59 |
| MRPL43 | ILMN_1700477 | 56 | 0.79 | 0.01 | 0.59 |
| WBSCR27 | ILMN_1719170 | 43 | 0.71 | 0.02 | 0.71 |
| PEX6 | ILMN_1683279 | 112 | 0.78 | 0.06 | 1.35 |
| LOC197322 | ILMN_1716832 | 95 | 0.74 | 0.07 | 1.35 |
| FLJ10781 | ILMN_1794490 | 110 | 0.73 | 0.08 | 1.35 |
| TSGA10 | ILMN_1680430 | 38 | 0.59 | 0.10 | 1.42 |
| NQO2 | ILMN_1712918 | 127 | 0.73 | 0.12 | 1.49 |
| CHPT1 | ILMN_1729112 | 74 | 0.61 | 0.20 | 2.20 |
| CCNDBP1 | ILMN_1711459 | 24 | 0.45 | 0.25 | 2.53 |
| FN3KRP | ILMN_1652333 | 48 | 0.53 | 0.29 | 2.61 |
| SLC2A8 | ILMN_1724609 | 71 | 0.54 | 0.36 | 3.02 |
| IRF5 | ILMN_1670576 | 62 | 0.50 | 0.54 | 4.01 |
| UGT2B17 | ILMN_1752214 | 32 | 0.42 | 0.56 | 4.01 |
| UTS2 | ILMN_1748520 | 100 | 0.55 | 0.63 | 4.22 |
| SNX16 | ILMN_1787415 | 61 | 0.44 | 0.94 | 5.73 |
| F25965 | ILMN_1656886 | 106 | 0.51 | 0.97 | 5.73 |
| LZIC | ILMN_1661627 | 111 | 0.52 | 1.04 | 5.79 |
| UGT2B7 | ILMN_1679194 | 33 | 0.37 | 1.10 | 5.81 |
| NUDT2 | ILMN_1778347 | 48 | 0.38 | 1.57 | 7.39 |
| TFAM | ILMN_1715661 | 53 | 0.39 | 1.61 | 7.39 |
| LTBR | ILMN_1667476 | 115 | 0.47 | 1.66 | 7.39 |
| OAS1 | ILMN_1672606 | 67 | 0.40 | 1.70 | 7.39 |
| TIMM10 | ILMN_1765332 | 32 | 0.32 | 2.01 | 8.38 |
| NDUFS5 | ILMN_1776104 | 104 | 0.42 | 2.23 | 8.93 |
| SLC27A5 | ILMN_1725366 | 29 | 0.30 | 2.43 | 9.35 |
| BCAS1 | ILMN_1776647 | 151 | 0.42 | 3.84 | 13.84 |
| SURF1 | ILMN_1663407 | 120 | 0.38 | 3.87 | 13.84 |
| SFXN2 | ILMN_1795976 | 28 | 0.24 | 4.29 | 14.80 |
| KIAA1913 | ILMN_1778400 | 88 | 0.32 | 4.70 | 15.30 |
| CRYZ | ILMN_1672389 | 49 | 0.26 | 4.96 | 15.30 |
| DPYSL4 | ILMN_1792356 | 113 | 0.35 | 5.02 | 15.30 |
| RABGEF1 | ILMN_1760884 | 33 | 0.23 | 5.14 | 15.30 |
| PHACS | ILMN_1757950 | 122 | 0.35 | 5.20 | 15.30 |
| STAT6 | ILMN_1763198 | 46 | 0.25 | 5.44 | 15.55 |
| CD151 | ILMN_1661589 | 75 | 0.29 | 5.70 | 15.84 |
| PAQR6 | ILMN_1737631 | 55 | 0.25 | 6.45 | 17.43 |
| RAD51C | ILMN_1760635 | 45 | 0.22 | 7.25 | 19.07 |
| ACN9 | ILMN_1771348 | 69 | 0.24 | 8.07 | 20.62 |
| ABCD2 | ILMN_1652959 | 35 | 0.18 | 8.40 | 20.62 |
| DDX17 | ILMN_1724114 | 82 | 0.25 | 8.45 | 20.62 |
| ENTPD1 | ILMN_1773125 | 65 | 0.21 | 10.03 | 23.87 |
| PLTP | ILMN_1711748 | 93 | 0.23 | 10.57 | 24.57 |
| UGT2B10 | ILMN_1742444 | 31 | 0.15 | 10.91 | 24.79 |
| IPP | ILMN_1789106 | 28 | 0.15 | 11.29 | 25.08 |
| IVD | ILMN_1724207 | 60 | 0.18 | 11.76 | 25.56 |
| STIM2 | ILMN_1738449 | 98 | 0.22 | 12.35 | 26.18 |
| RAMP1 | ILMN_1764754 | 102 | 0.22 | 12.56 | 26.18 |
| CDK5RAP2 | ILMN_1725235 | 55 | 0.17 | 13.29 | 27.12 |
| RPS16 | ILMN_1651850 | 73 | 0.18 | 14.62 | 29.24 |

| Gene | Probe | $d_k$ | $\tilde{r}^2$ | $p$-value ($\times 10^2$) | $q$-value ($\times 10^2$) |
|---|---|---|---|---|---|
| C1QTNF3 | ILMN_1687260 | 103 | 0.19 | 15.13 | 29.66 |
| MPHOSPH1 | ILMN_1712452 | 79 | 0.15 | 18.01 | 34.64 |
| DNASE1L3 | ILMN_1762084 | 30 | 0.10 | 18.79 | 35.42 |
| SUPT3H | ILMN_1813277 | 95 | 0.15 | 19.13 | 35.42 |
| SNRPB2 | ILMN_1771620 | 140 | 0.17 | 19.63 | 35.69 |
| NIPSNAP3B | ILMN_1786308 | 73 | 0.13 | 20.99 | 37.48 |
| RPL37A | ILMN_1711222 | 115 | 0.14 | 23.21 | 40.72 |
| MRPL53 | ILMN_1813682 | 72 | 0.10 | 24.59 | 42.39 |
| TLR6 | ILMN_1749287 | 124 | 0.13 | 25.30 | 42.87 |
| SLC7A7 | ILMN_1810275 | 166 | 0.14 | 26.56 | 43.83 |
| RPS6KB2 | ILMN_1761175 | 34 | 0.07 | 26.74 | 43.83 |
| C10orf68 | ILMN_1662791 | 56 | 0.08 | 27.32 | 44.07 |
| PPA2 | ILMN_1755041 | 41 | 0.07 | 28.05 | 44.29 |
| GSTP1 | ILMN_1679809 | 38 | 0.06 | 28.34 | 44.29 |
| ZNF175 | ILMN_1675788 | 168 | 0.12 | 29.21 | 44.94 |
| UMPS | ILMN_1757437 | 80 | 0.07 | 32.79 | 49.68 |
| ATP6V1C1 | ILMN_1659801 | 74 | 0.06 | 34.65 | 51.71 |
| C10orf88 | ILMN_1775423 | 141 | 0.06 | 37.36 | 54.94 |
| RPH3AL | ILMN_1693717 | 108 | 0.04 | 40.57 | 58.80 |
| ACY1L2 | ILMN_1766000 | 77 | 0.01 | 47.97 | 67.67 |
| CIAO1 | ILMN_1792837 | 21 | 0.00 | 48.59 | 67.67 |
| MRPL21 | ILMN_1744835 | 82 | 0.00 | 48.72 | 67.67 |
| EIF2S1 | ILMN_1739821 | 46 | -0.02 | 57.03 | 77.29 |
| STOM | ILMN_1696419 | 71 | -0.02 | 57.19 | 77.29 |
| LOC339229 | ILMN_1805131 | 74 | -0.03 | 59.85 | 78.89 |
| NPEPL1 | ILMN_1724194 | 155 | -0.05 | 60.17 | 78.89 |
| C14orf130 | ILMN_1667839 | 117 | -0.04 | 60.75 | 78.89 |
| PRIC285 | ILMN_1787509 | 85 | -0.04 | 62.33 | 79.91 |
| NDUFV3 | ILMN_1765500 | 145 | -0.07 | 66.44 | 84.11 |
| MUM1 | ILMN_1764764 | 185 | -0.10 | 69.81 | 86.43 |
| TNFRSF18 | ILMN_1743100 | 56 | -0.06 | 71.05 | 86.43 |
| LOC144097 | ILMN_1795564 | 60 | -0.06 | 71.40 | 86.43 |
| B3GALT3 | ILMN_1732822 | 39 | -0.05 | 72.53 | 86.43 |
| G0S2 | ILMN_1691846 | 115 | -0.09 | 72.60 | 86.43 |
| ABCF2 | ILMN_1709222 | 144 | -0.13 | 78.30 | 92.11 |
| MTRR | ILMN_1702684 | 97 | -0.13 | 82.96 | 96.16 |
| SFRS10 | ILMN_1742798 | 108 | -0.14 | 83.66 | 96.16 |
| PLXDC2 | ILMN_1753312 | 85 | -0.12 | 86.03 | 97.76 |
| POLE4 | ILMN_1660063 | 92 | -0.13 | 87.67 | 98.20 |
| KIAA0265 | ILMN_1702635 | 79 | -0.13 | 88.38 | 98.20 |
| TCL1B | ILMN_1805857 | 190 | -0.26 | 93.67 | 99.61 |
| POLR3F | ILMN_1673966 | 109 | -0.19 | 94.27 | 99.61 |
| GPX4 | ILMN_1734353 | 181 | -0.27 | 94.51 | 99.61 |
| IL16 | ILMN_1813572 | 104 | -0.19 | 95.37 | 99.61 |
| RAD18 | ILMN_1707548 | 137 | -0.24 | 95.52 | 99.61 |
| DHX30 | ILMN_1813719 | 17 | -0.05 | 96.43 | 99.61 |
| SOS1 | ILMN_1767135 | 54 | -0.12 | 96.62 | 99.61 |
| LOC400696 | ILMN_1759989 | 81 | -0.23 | 99.87 | 100.00 |
| ALDH8A1 | ILMN_1699258 | 77 | -0.23 | 100.00 | 100.00 |
| STK25 | ILMN_1668090 | 74 | -0.25 | 100.00 | 100.00 |

$$- \sigma^4 \left[E\left\{\mathrm{tr}(X^{\mathrm{T}}X)\right\}\right]^2 - 2\sigma^2 E\left\{\mathrm{tr}(X^{\mathrm{T}}X)\right\} E\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\}$$
$$- \left[E\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\}\right]^2.$$

Equation (19) now follows from Proposition S1 below (found in Section S3). Equation (20) is proved similarly: we have

$$\begin{aligned}
\mathrm{cov}(||y||^2, ||X^{\mathrm{T}}y||^2) &= E\left\{\mathrm{cov}(||y||^2, ||X^{\mathrm{T}}y||^2 \mid X)\right\} + \mathrm{cov}\left\{E(||y||^2 \mid X), E(||X^{\mathrm{T}}y||^2 \mid X)\right\} \\
&= 2\sigma^4 E\left\{\mathrm{tr}(X^{\mathrm{T}}X)\right\} + 4\sigma^2 E\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\} \\
&\quad + \mathrm{cov}\left\{\beta^{\mathrm{T}}X^{\mathrm{T}}X\beta, \sigma^2\mathrm{tr}(X^{\mathrm{T}}X) + \beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\} \\
&= 2\sigma^4 E\left\{\mathrm{tr}(X^{\mathrm{T}}X)\right\} + 4\sigma^2 E\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\} \\
&\quad + \sigma^2 E\left\{\mathrm{tr}(X^{\mathrm{T}}X)\beta^{\mathrm{T}}X^{\mathrm{T}}X\beta\right\} + E\left\{\beta^{\mathrm{T}}X^{\mathrm{T}}X\beta\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\} \\
&\quad - \sigma^2 E\left(\beta^{\mathrm{T}}X^{\mathrm{T}}X\beta\right) E\left\{\mathrm{tr}(X^{\mathrm{T}}X)\right\} - E\left(\beta^{\mathrm{T}}X^{\mathrm{T}}X\beta\right) E\left\{\beta^{\mathrm{T}}(X^{\mathrm{T}}X)^2\beta\right\}
\end{aligned}$$

and (20) follows from Proposition S1.

□

**Corollary S1.** *Suppose that* $\Sigma = I$. *Then*

$$\begin{aligned}
\mathrm{var}(\hat{\sigma}^2) &= \frac{2n}{(n+1)^2}\left\{\left(\frac{d}{n} + 1 + \frac{2d}{n^2} + \frac{2}{n} + \frac{1}{n^2}\right)\sigma^4 + \left(\frac{2d}{n} + \frac{4d}{n^2} + \frac{4}{n} + \frac{8}{n^2}\right)\sigma^2\tau^2 \right. \\
&\qquad \left. + \left(\frac{d}{n} + 1 + \frac{2d}{n^2} + \frac{7}{n} + \frac{10}{n^2}\right)\tau^4\right\} \\
\mathrm{var}(\hat{\tau}^2) &= \frac{2n}{(n+1)^2}\left\{\left(\frac{d}{n} + \frac{2d}{n^2}\right)\sigma^4 + \left(\frac{2d}{n} + 2 + \frac{4d}{n^2} + \frac{10}{n} + \frac{12}{n^2}\right)\sigma^2\tau^2 \right. \\
&\qquad \left. + \left(\frac{d}{n} + 4 + \frac{2d}{n^2} + \frac{15}{n} + \frac{15}{n^2}\right)\tau^4\right\} \\
\mathrm{cov}(\hat{\sigma}^2, \hat{\tau}^2) &= -\frac{2n}{(n+1)^2}\left\{\left(\frac{d}{n} + \frac{2d}{n^2}\right)\sigma^4 + \left(\frac{2d}{n} + \frac{4d}{n^2} + \frac{5}{n} + \frac{9}{n^2}\right)\sigma^2\tau^2 \right. \\
&\qquad \left. + \left(\frac{d}{n} + 2 + \frac{2d}{n^2} + \frac{10}{n} + \frac{12}{n^2}\right)\tau^4\right\}.
\end{aligned}$$

Corollary S1 follows from Lemma S1 and the fact that

$$\begin{pmatrix} \hat{\sigma}^2 \\ \hat{\tau}^2 \end{pmatrix} = \begin{pmatrix} \frac{d+n+1}{n+1} & -\frac{n}{n+1} \\ -\frac{d}{n+1} & \frac{n}{n+1} \end{pmatrix} \begin{pmatrix} n^{-1}||y||^2 \\ n^{-2}||X^{\mathrm{T}}y||^2 \end{pmatrix}.$$

Lemma 2 follows immediately from Corollary S1.

***Proof of Theorem 1***

Theorem 1 is a direct application of Theorem 2.2 from (Chatterjee, 2009), which is stated here for ease of reference.

**Theorem S1.** [Theorem 2.2, (Chatterjee, 2009)] *Let $v = (v_1, \ldots, v_m) \sim N(0, \Psi)$. Suppose that $g \in C^2(\mathbb{R}^m)$ and let $\nabla g$ and $\nabla^2 g$ denote the gradient and the Hessian of $g$, respectively. Let*

$$\kappa_1 = \left[E\left\{||\nabla g(v)||^4\right\}\right]^{1/4},$$
$$\kappa_2 = \left[E\left\{||\nabla^2 g(v)||^4\right\}\right]^{1/4},$$

*where $||\nabla^2 g(v)||$ is the operator norm of $\nabla^2 g(v)$. Suppose that $E\{g(v)^4\} < \infty$ and let $\psi^2 = \mathrm{var}\{g(v)\}$. Let $w$ be a normal random variable having the same mean and variance as $g(v)$. Then*

$$d_{TV}\{g(v), w\} \leq \frac{2\sqrt{5}||\Psi||^{3/2}\kappa_1\kappa_2}{\psi^2}. \tag{21}$$

*Remark* 1. Chatterjee's Theorem 2.2 does not actually require Gaussian $v$. However, for non-Gaussian $v$, an additional term appears in the bound (21), which is not sufficiently small for our purposes. Furthermore, the class of distributions covered by the full version of Chatterjee's Theorem 2.2 is not all-encompassing: $v_i$ must be a $C^2$-function of a normal random variable.

To prove Theorem 1, we apply Theorem S1 with $v = (X, \epsilon) \in \mathbb{R}^{(d+1)n}$. Let $h \in C^2(\mathbb{R}^2)$ and let

$$g(X, \epsilon) = h(S),$$

where $S = S(X, \epsilon) = (n^{-1}||y||^2, n^{-2}||X^{\mathrm{T}}y||^2)$. First, we bound the quantities $\kappa_1$, $\kappa_2$ in Theorem S1. In order to bound $\kappa_1$, we compute the gradient of $g$. Let $h_1$, $h_2$ denote the partial derivatives of $h$ with respect to the first and second variables, respectively. Then

$$\frac{\partial g}{\partial x_{ij}}(X, \epsilon) = h_1(S)\frac{\partial}{\partial x_{ij}}\frac{1}{n}||y||^2 + h_2(S)\frac{\partial}{\partial x_{ij}}\frac{1}{n^2}||X^{\mathrm{T}}y||^2$$

for $i = 1, \ldots, n$, $j = 1, \ldots, d$. Let $\mathsf{E}_{ij}$ denote the $n \times d$ matrix with $i'j'$-entry $\delta_{ii'}\delta_{jj'}$ ($\delta_{ii'} = 1$ if $i = i'$ and 0 otherwise). Since

$$\frac{\partial}{\partial x_{ij}}||y||^2 = 2\beta^{\mathrm{T}}\mathsf{E}_{ij}^{\mathrm{T}}y$$

and

$$\frac{\partial}{\partial x_{ij}}||X^{\mathrm{T}}y||^2 = 2y^{\mathrm{T}}\mathsf{E}_{ij}X^{\mathrm{T}}y + 2\beta^{\mathrm{T}}\mathsf{E}_{ij}^{\mathrm{T}}XX^{\mathrm{T}}y,$$

it follows that

$$\frac{\partial g}{\partial x_{ij}}(X,\epsilon) = 2h_1(S)\left(\frac{1}{n}\beta^{\mathrm{T}}\mathsf{E}_{ij}^{\mathrm{T}}y\right) + 2h_2(S)\left(\frac{1}{n^2}y^{\mathrm{T}}\mathsf{E}_{ij}X^{\mathrm{T}}y + \frac{1}{n^2}\beta^{\mathrm{T}}\mathsf{E}_{ij}^{\mathrm{T}}XX^{\mathrm{T}}y\right). \tag{22}$$

For $1 \leq k \leq n$, the partial derivative of $g$ with respect to $\epsilon_k$ is

$$\begin{aligned}\frac{\partial g}{\partial \epsilon_k}(X,\epsilon) &= h_1(S)\frac{\partial}{\partial \epsilon_k}\frac{1}{n}||y||^2 + h_2(S)\frac{\partial}{\partial \epsilon_k}\frac{1}{n^2}||X^{\mathrm{T}}y||^2 \\ &= 2h_1(S)\left(\frac{1}{n}e_k^{\mathrm{T}}y\right) + 2h_2(S)\left(\frac{1}{n^2}e_k^{\mathrm{T}}XX^{\mathrm{T}}y\right),\end{aligned} \tag{23}$$

where $e_k = (0,...,0,1,0,...,0)^{\mathrm{T}} \in \mathbb{R}^n$ is the $k$-th standard basis vector in $\mathbb{R}^n$ and we have used the facts

$$\begin{aligned}\frac{\partial}{\partial \epsilon_k}||y||^2 &= 2e_k^{\mathrm{T}}y \\ \frac{\partial}{\partial \epsilon_k}||X^{\mathrm{T}}y||^2 &= 2e_kXX^{\mathrm{T}}y.\end{aligned}$$

Now recall that $\kappa_1 = [E\{||\nabla g(X,\epsilon)||^4\}]^{1/4}$. Equations (22)–(23) and the elementary inequality

$$(a+b)^2 \leq 2a^2 + 2b^2, \quad a,b \in \mathbb{R}, \tag{24}$$

imply that

$$\begin{aligned}||\nabla g(X,\epsilon)||^2 &= \sum_{i=1}^n\sum_{j=1}^d\left\{\frac{\partial}{\partial x_{ij}}g(X,\epsilon)\right\}^2 + \sum_{k=1}^n\left\{\frac{\partial}{\partial \epsilon_k}g(X,\epsilon)\right\}^2 \\ &\leq 8h_1(S)^2\sum_{i=1}^n\sum_{j=1}^d\left(\frac{1}{n}\beta^{\mathrm{T}}\mathsf{E}_{ij}^{\mathrm{T}}y\right)^2 \\ &\quad + 16h_2(S)^2\sum_{i=1}^n\sum_{j=1}^d\left\{\left(\frac{1}{n^2}y^{\mathrm{T}}\mathsf{E}_{ij}X^{\mathrm{T}}y\right)^2 + \left(\frac{1}{n^2}\beta^{\mathrm{T}}\mathsf{E}_{ij}^{\mathrm{T}}XX^{\mathrm{T}}y\right)^2\right\} \\ &\quad + 8h_1(S)^2\sum_{k=1}^n\left(\frac{1}{n}e_k^{\mathrm{T}}y\right)^2 + 8h_2(S)^2\sum_{k=1}^n\left(\frac{1}{n^2}e_k^{\mathrm{T}}XX^{\mathrm{T}}y\right)^2 \\ &= \frac{8}{n^2}(\tau^2+1)h_1(S)^2||y||^2 \\ &\quad + \frac{16}{n^4}h_2(S)^2\left\{||y||^2||X^{\mathrm{T}}y||^2 + \left(\tau^2+\frac{1}{2}\right)||XX^{\mathrm{T}}y||^2\right\}.\end{aligned}$$

Let $\lambda_1 = ||n^{-1}X^{\mathrm{T}}X||$ be the largest eigenvalue of $n^{-1}X^{\mathrm{T}}X$. Applying the triangle inequality and (24) yields

$$\begin{aligned}
||\nabla g(X,\epsilon)||^2 &\leq \frac{16}{n^2}(\tau^2+1)h_1(S)^2\left(||X^{\mathrm{T}}X||\tau^2+||\epsilon||^2\right) \\
&\quad + \frac{128}{n^4}h_2(S)^2||X^{\mathrm{T}}X||\left(||X^{\mathrm{T}}X||^2\tau^4+||\epsilon||^4\right) \\
&\quad + \frac{32}{n^4}h_2(S)^2||X^{\mathrm{T}}X||^2\left(\tau^2+\frac{1}{2}\right)\left(||X^{\mathrm{T}}X||\tau^2+||\epsilon||^2\right) \\
&\leq \frac{16}{n}||\nabla h(S)||^2\Bigg\{8\lambda_1\left(\frac{1}{n}||\epsilon||^2\right)^2+\left(2\lambda_1^2+1\right)\frac{1}{n}||\epsilon||^2\tau^2+\left(10\lambda_1^3+\lambda_1\right)\tau^4 \\
&\quad + (\lambda_1^2+1)\frac{1}{n}||\epsilon||^2+(\lambda_1^3+\lambda_1)\tau^2\Bigg\} \\
&\leq \frac{264}{n}||\nabla h(S)||^2(\lambda_1+1)^3\left\{\frac{1}{n}||\epsilon||^2\left(\frac{1}{n}||\epsilon||^2+1\right)+\tau^2(\tau^2+1)\right\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\kappa_1 &= \left[E\left\{||\nabla g(X,\epsilon)||^4\right\}\right]^{1/4} \\
&\leq \sqrt{\frac{264}{n}}\left(E\left[||\nabla h(S)||^4(\lambda_1+1)^6\left\{\frac{1}{n}||\epsilon||^2\left(\frac{1}{n}||\epsilon||^2+1\right)+\tau^2(\tau^2+1)\right\}^2\right]\right)^{1/4} \\
&= O\left[\frac{1}{\sqrt{n}}\left\{\gamma_4^{1/4}+\gamma_2^{1/4}+\gamma_0^{1/4}\tau(\tau+1)\right\}\right], \qquad (25)
\end{aligned}$$

where

$$\gamma_k = E\left[||\nabla h(S)||^4(\lambda_1+1)^6\left(\frac{1}{n}||\epsilon||^2\right)^k\right].$$

To bound $\kappa_2 = [E\{||\nabla^2 g(X,\epsilon)||^4\}]^{1/4}$, we bound the operator norm of the Hessian $||\nabla^2 g(X,\epsilon)||$. Let

$$\mathcal{U} = \left\{\tilde{U} = (u\ U);\ u = (u_1,\ldots,u_n)\in\mathbb{R}^n,\ U = (u_{ij})_{1\leq i\leq n,\ 1\leq j\leq d},\ \sum_{k=1}^n u_k^2+\sum_{i=1}^n\sum_{j=1}^d u_{ij}^2 = 1\right\}$$

be the collection of partitioned $n\times(d+1)$ matrices with Frobenius norm equal to one. For $\tilde{U} = (u\ U)\in\mathcal{U}$, define the differential operator

$$D_{\tilde{U}} = \sum_{i=1}^n\sum_{j=1}^d u_{ij}\frac{\partial}{\partial x_{ij}}+\sum_{k=1}^n u_k\frac{\partial}{\partial\epsilon_k}.$$

Then

$$\begin{aligned}||\nabla^2 g(X,\epsilon)|| &= \sup_{\tilde{U}\in\mathcal{U}} D^2_{\tilde{U}} g(X,\epsilon) \\ &= \sup_{\tilde{U}\in\mathcal{U}} \left\{\nabla h(S)^{\mathrm{T}} D^2_{\tilde{U}} S(X,\epsilon) + \left\{D_{\tilde{U}} S(X,\epsilon)\right\}^{\mathrm{T}} \nabla^2 h(S) D_{\tilde{U}} S(X,\epsilon)\right\} \\ &\leq \sup_{\tilde{U}\in\mathcal{U}} \left\{||\nabla h(S)|| \left|\left|D^2_{\tilde{U}} S(X,\epsilon)\right|\right| + \left|\left|\nabla^2 h(S)\right|\right| ||D_{\tilde{U}} S(X,\epsilon)||^2\right\}. \end{aligned} \tag{26}$$

From our previous calculations,

$$\begin{aligned} D_{\tilde{U}} S(X,\epsilon) &= \sum_{i=1}^n \sum_{j=1}^d u_{ij} \frac{\partial}{\partial x_{ij}} \begin{pmatrix} \frac{1}{n}||y||^2 \\ \frac{1}{n^2}||X^{\mathrm{T}} y||^2 \end{pmatrix} + \sum_{k=1}^n u_k \frac{\partial}{\partial \epsilon_k} \begin{pmatrix} \frac{1}{n}||y||^2 \\ \frac{1}{n^2}||X^{\mathrm{T}} y||^2 \end{pmatrix} \\ &= \sum_{i=1}^n \sum_{j=1}^d u_{ij} \begin{pmatrix} \frac{2}{n}\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} y \\ \frac{2}{n^2} y^{\mathrm{T}} \mathsf{E}_{ij} X^{\mathrm{T}} y + \frac{2}{n^2}\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} X X^{\mathrm{T}} y \end{pmatrix} + \sum_{k=1}^n u_k \begin{pmatrix} \frac{2}{n} e_k^{\mathrm{T}} y \\ \frac{2}{n^2} e_k^{\mathrm{T}} X X^{\mathrm{T}} y \end{pmatrix} \\ &= \begin{pmatrix} \frac{2}{n}\beta^{\mathrm{T}} U^{\mathrm{T}} y + \frac{2}{n} u^{\mathrm{T}} y \\ \frac{2}{n^2} y^{\mathrm{T}} U X^{\mathrm{T}} y + \frac{2}{n^2}\beta^{\mathrm{T}} U^{\mathrm{T}} X X^{\mathrm{T}} y + \frac{2}{n^2} u^{\mathrm{T}} X X^{\mathrm{T}} y \end{pmatrix}. \end{aligned}$$

To compute $D^2_{\tilde{U}} S(X,\epsilon)$, we need the second order partial derivatives of $||y||^2$ and $||X^{\mathrm{T}} y||^2$; these are

$$\begin{aligned} \frac{\partial^2}{\partial x_{i'j'} \partial x_{ij}} ||y||^2 &= 2\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} \mathsf{E}_{i'j'} \beta, \\ \frac{\partial^2}{\partial \epsilon_k \partial x_{ij}} ||y||^2 &= 2\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} e_k, \\ \frac{\partial^2}{\partial \epsilon_{k'} \partial \epsilon_k} ||y||^2 &= 2 e_k^{\mathrm{T}} e_{k'} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2}{\partial x_{i'j'} \partial x_{ij}} ||X^{\mathrm{T}} y||^2 &= 2\beta^{\mathrm{T}} \mathsf{E}_{i'j'}^{\mathrm{T}} \mathsf{E}_{ij} X^{\mathrm{T}} y + 2\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} \mathsf{E}_{i'j'} X^{\mathrm{T}} y + 2 y^{\mathrm{T}} \mathsf{E}_{ij} \mathsf{E}_{i'j'}^{\mathrm{T}} y \\ &\quad + 2 y^{\mathrm{T}} \mathsf{E}_{ij} X^{\mathrm{T}} \mathsf{E}_{i'j'} \beta + 2\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} X \mathsf{E}_{i'j'}^{\mathrm{T}} y + 2\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} X X^{\mathrm{T}} \mathsf{E}_{i'j'} \beta, \\ \frac{\partial^2}{\partial \epsilon_k \partial x_{ij}} ||X^{\mathrm{T}} y||^2 &= 2 e_k^{\mathrm{T}} \mathsf{E}_{ij} X^{\mathrm{T}} y + 2 y^{\mathrm{T}} \mathsf{E}_{ij} X^{\mathrm{T}} e_k + 2\beta^{\mathrm{T}} \mathsf{E}_{ij}^{\mathrm{T}} X X^{\mathrm{T}} e_k \\ \frac{\partial^2}{\partial \epsilon_{k'} \partial \epsilon_k} ||X^{\mathrm{T}} y||^2 &= 2 e_k^{\mathrm{T}} X X^{\mathrm{T}} e_{k'}, \end{aligned}$$

for $1 \leq i, k \leq d$ and $1 \leq j \leq d$. It follows that the entries of $D^2_{\tilde{U}} S(X, \epsilon)$ are

$$\frac{1}{n} D^2_{\tilde{U}} ||y||^2 = \frac{2}{n} \beta^{\mathrm{T}} U^{\mathrm{T}} U \beta + \frac{4}{n} \beta^{\mathrm{T}} U^{\mathrm{T}} u + \frac{2}{n} ||u||^2$$

and

$$\begin{aligned} \frac{1}{n^2} D^2_{\tilde{U}} ||X^{\mathrm{T}} y||^2 = &\frac{2}{n^2} y^{\mathrm{T}} U U^{\mathrm{T}} y + \frac{4}{n^2} \beta^{\mathrm{T}} U^{\mathrm{T}} U X^{\mathrm{T}} y + \frac{4}{n^2} \beta^{\mathrm{T}} U^{\mathrm{T}} X U^{\mathrm{T}} y + \frac{2}{n^2} \beta^{\mathrm{T}} U^{\mathrm{T}} X X^{\mathrm{T}} U \beta \\ &+ \frac{4}{n^2} u^{\mathrm{T}} U X^{\mathrm{T}} y + \frac{4}{n^2} y^{\mathrm{T}} U X^{\mathrm{T}} u + \frac{4}{n^2} \beta^{\mathrm{T}} U^{\mathrm{T}} X X^{\mathrm{T}} u + \frac{2}{n^2} u^{\mathrm{T}} X X^{\mathrm{T}} u. \end{aligned}$$

We conclude that

$$\begin{aligned} ||D_{\tilde{U}} S(X, \epsilon)||^2 &= \frac{4}{n^2} \left( \beta^{\mathrm{T}} U^{\mathrm{T}} y + u^{\mathrm{T}} y \right)^2 + \frac{4}{n^4} \left( y^{\mathrm{T}} U X^{\mathrm{T}} y + \beta^{\mathrm{T}} U^{\mathrm{T}} X X^{\mathrm{T}} y + u^{\mathrm{T}} X X^{\mathrm{T}} y \right)^2 \\ &\leq \frac{8}{n^2} (\tau^2 + 1) ||y||^2 + \frac{12}{n^4} ||X^{\mathrm{T}} X|| \left( ||y||^2 + ||X^{\mathrm{T}} X|| \tau^2 + ||X^{\mathrm{T}} X|| \right) ||y||^2 \\ &\leq \frac{16}{n} (\tau^2 + 1) \left( \lambda_1 \tau^2 + \frac{1}{n} ||\epsilon||^2 \right) \\ &\quad + \frac{168}{n} \lambda_1 \left\{ \lambda_1^2 \tau^2 (\tau^2 + 1) + \frac{1}{n} ||\epsilon||^2 \left( \lambda_1 + \frac{1}{n} ||\epsilon||^2 \right) \right\} \\ &= O \left[ \frac{1}{n} \left\{ (\lambda_1^3 + \lambda_1) \tau^2 (\tau^2 + 1) + \frac{1}{n} ||\epsilon||^2 \left( \lambda_1 + \frac{1}{n} ||\epsilon||^2 + 1 \right) \right\} \right] \end{aligned} \tag{27}$$

and

$$\begin{aligned} ||D^2_{\tilde{U}} S(X, \epsilon)|| &\leq \frac{2}{n} (\tau + 1)^2 + \frac{2}{n^2} \left\{ ||y||^2 + 4 ||X|| (\tau + 1) ||y|| + ||X^{\mathrm{T}} X|| (\tau + 1)^2 \right\} \\ &= O \left[ \frac{1}{n} \left\{ (\lambda_1 + 1)(\tau + 1)^2 + \frac{1}{n} ||\epsilon||^2 \right\} \right]. \end{aligned} \tag{28}$$

Combining (26)–(28), we obtain

$$\begin{aligned} \kappa_2 &= \left[ E \left\{ ||\nabla^2 g(X, \epsilon)||^4 \right\} \right]^{1/4} \\ &= O \left[ \frac{1}{n} \left\{ \eta_8^{1/4} + \eta_4^{1/4} + \eta_0^{1/4} \tau^2 (\tau^2 + 1) + \gamma_4^{1/4} + \gamma_0^{1/4} (\tau^2 + 1) \right\} \right], \end{aligned} \tag{29}$$

where

$$\eta_k = E \left\{ ||\nabla^2 h(S)||^4 (\lambda_1 + 1)^{12} \left( \frac{1}{n} ||\epsilon||^2 \right)^k \right\}.$$

Appealing to Theorem S1, the bounds (25) and (29) imply

$$d_{TV}\left\{g(X,\epsilon),w\right\} = O\left(\frac{\xi\nu}{n^{3/2}\psi^2}\right),$$

where

$$\xi = \xi(\sigma^2,\tau^2,\Sigma,d,n) = \gamma_4^{1/4} + \gamma_2^{1/4} + \gamma_0^{1/4}\tau(\tau+1)$$

and

$$\nu = \nu(\sigma^2,\tau^2,\Sigma,d,n) = \eta_8^{1/4} + \eta_4^{1/4} + \eta_0^{1/4}\tau^2(\tau^2+1) + \gamma_4^{1/4} + \gamma_0^{1/4}(\tau^2+1).$$

This completes the proof of Theorem 1.

### *Proof outline for Proposition 2*

Assume that the conditions of Proposition 2 are satisfied and let

$$\begin{aligned}\tilde{\sigma}^2(\hat{m}) = \tilde{\sigma}^2 &= \left\{1 + \frac{d\hat{m}_1^2}{(n+1)\hat{m}_2}\right\}\frac{1}{n}||y||^2 - \frac{\hat{m}_1}{n(n+1)\hat{m}_2}||X^{\mathrm{T}}y||^2,\\ \tilde{\tau}^2(\hat{m}) = \tilde{\tau}^2 &= -\frac{d\hat{m}_1^2}{n(n+1)\hat{m}_2}||y||^2 + \frac{\hat{m}_1}{n(n+1)\hat{m}_2}||X^{\mathrm{T}}y||^2,\end{aligned}$$

where $\hat{m} = (\hat{m}_1,\hat{m}_2)^{\mathrm{T}}$. Let $m = (m_1,m_2)^{\mathrm{T}} = (d^{-1}\mathrm{tr}(\Sigma), d^{-1}\mathrm{tr}(\Sigma^2))^{\mathrm{T}}$ and consider the estimators $\tilde{\sigma}^2(m)$, $\tilde{\tau}^2(m)$. Properties of the Wishart distribution and Proposition S1 imply that $\hat{m}_k = m_k + O\{(nd)^{-1/2}\}$, $k = 1,2$. It follows that

$$\tilde{\sigma}^2(\hat{m}) = \tilde{\sigma}^2(m) + O\{(nd)^{-1/2}\}, \quad \tilde{\tau}^2(\hat{m}) = \tilde{\tau}^2(m) + O\{(nd)^{-1/2}\}. \tag{30}$$

Using Proposition S1 again, one can further prove that

$$\tilde{\sigma}^2(m) = \sigma^2 + O(n^{-1/2} + \tilde{\Delta}_2), \quad \tilde{\tau}^2(m) = \tau^2 + O(n^{-1/2} + \tilde{\Delta}_2). \tag{31}$$

Part (i) of Proposition 2 follows from (30)–(31).

By Proposition S1,

$$\begin{gathered}E\{\tilde{\sigma}^2(m)\} = \sigma^2 + O(\tilde{\Delta}_2), \quad E\{\tilde{\tau}^2(m)\} = \tau^2 + O(\tilde{\Delta}_2),\\ n\mathrm{var}\{\tilde{\sigma}^2(m)\} = \tilde{\psi}_1^2 + O\{(nd)^{-1/2} + \tilde{\Delta}_3\}, \quad n\mathrm{var}\{\tilde{\tau}^2(m)\} = \tilde{\psi}_2^2 + O\{(nd)^{-1/2} + \tilde{\Delta}_3\},\\ n\mathrm{cov}\{\tilde{\sigma}^2(m),\tilde{\tau}^2(m)\} = \tilde{\psi}_{12} + O\{(nd)^{-1/2} + \tilde{\Delta}_3\},\end{gathered} \tag{32}$$

where

$$\tilde{\psi}_{12} = -2\left\{\frac{dm_1^2}{nm_2}(\sigma^2+\tau^2)^2 + \frac{2m_1m_3}{m_2^2}\tau^4 + 2\left(\frac{m_1m_3}{m_2^2} - 1\right)\sigma^2\tau^2\right\}.$$

Part (ii) of Proposition 2 follows from Theorem 1 (as in Corollary 1), (30) and (32).

## S3. Moment calculations for the Wishart distribution

Suppose that $X = (x_1, \ldots, x_n)^{\mathrm{T}}$ is an $n \times d$ matrix with iid rows $x_1, \ldots, x_n \sim N(0, \Sigma)$ and that $\Sigma$ is a $d \times d$ positive definite matrix. Then $W = X^{\mathrm{T}}X$ is a Wishart$(n, \Sigma)$ random matrix. Let $\beta \in \mathbb{R}^d$. In this Supplemental Text we provide formulas for various moments involving $W$ that are used in the paper. Letac and Massam (2004) and Graczyk et al. (2005) provide techniques for computing all such moments. These techniques are utilized here.

Let $S_k$ denote the symmetric group on $k$ elements. Then each permutation $\pi \in S_k$ can be uniquely as a product of disjoint cycles $\pi = C_1 \cdots C_{m(\pi)}$, where $C_j = (c_{1j} \cdots c_{k_j j})$, $k_1 + \cdots + k_{m(\pi)} = k$, and all of the $c_{ij} \in \{1, \ldots, k\}$ are distinct.

Let $H_1, \ldots, H_k$ be $d \times d$ symmetric matrices and define the polynomial

$$r_\pi(\Sigma)(H_1, \ldots, H_k) = \prod_{j=1}^{m(\pi)} \mathrm{tr}\left(\prod_{i=1}^{k_j} \Sigma H_{c_{ij}}\right).$$

Theorem 1 in Letac and Massam (2004) and Proposition 1 in Graczyk et al. (2005) give the following formula:

$$E\left\{\mathrm{tr}(WH_1) \cdots \mathrm{tr}(WH_k)\right\} = \sum_{\pi \in S_k} 2^{k-m(\pi)} n^{m(\pi)} r_\pi(\Sigma)(H_1, \ldots, H_k). \tag{33}$$

This is our main tool for deriving the explicit formulas in the following proposition. For non-negative integers $k$, define $\tau_k^2 = \beta^{\mathrm{T}}\Sigma^k\beta$ and $m_k = d^{-1}\mathrm{tr}(\Sigma^k)$.

**Proposition S1.** *We have*

$$E\left\{\mathrm{tr}(W)\right\} = dnm_1 \tag{34}$$

$$E\left\{\mathrm{tr}(W)^2\right\} = d^2n^2m_1^2 + 2dnm_2 \tag{35}$$

$$E\left\{\mathrm{tr}(W^2)\right\} = d^2nm_1^2 + dn(n+1)m_2 \tag{36}$$

$$E\left(\beta^{\mathrm{T}}W\beta\right) = n\tau_1^2 \tag{37}$$

$$E\left(\beta^{\mathrm{T}}W^2\beta\right) = dnm_1\tau_1^2 + n(n+1)\tau_2^2 \tag{38}$$

$$E\left\{\mathrm{tr}(W)\beta^{\mathrm{T}}W\beta\right\} = dn^2m_1\tau_1^2 + 2n\tau_2^2 \tag{39}$$

$$\begin{aligned} E\left\{\mathrm{tr}(W)\beta^{\mathrm{T}}W^2\beta\right\} &= d^2n^2m_1^2\tau_1^2 + dn(n^2+n+2)m_1\tau_2^2 \\ &\quad + 2dnm_2\tau_1^2 + 4n(n+1)\tau_3^2 \end{aligned} \tag{40}$$

$$E(\beta^{\mathrm{T}}W\beta\beta^{\mathrm{T}}W^2\beta) = dn(n+2)m_1\tau_1^4 + n(n+2)(n+3)\tau_1^2\tau_2^2 \tag{41}$$

$$\begin{aligned} E\left(\beta^{\mathrm{T}}W^3\beta\right) &= d^2nm_1^2\tau_1^2 + 2dn(n+1)m_1\tau_2^2 \\ &\quad + dn(n+1)m_2\tau_1^2 + n(n^2+3n+4)\tau_3^2 \end{aligned} \tag{42}$$

$$E\left\{(\beta^{\mathrm{T}}W^2\beta)^2\right\} = d^2n(n+2)m_1^2\tau_1^4 + 2dn(n+2)(n+3)m_1\tau_1^2\tau_2^2 + 2dn(n+2)m_2\tau_1^4 + 4n(n+2)(n+3)\tau_1^2\tau_3^2 + n(n+1)(n+2)(n+3)\tau_2^4. \tag{43}$$

*Proof.* Formulas (34) and (37) are trivial (notice that $\beta^{\mathrm{T}}W\beta \sim \tau_1^2\chi_n^2$). Formulas (35)–(36) may be found in (Letac and Massam, 2004).

Now let $u_1, \ldots, u_d \in \mathbb{R}^d$ be an orthonormal basis of $\mathbb{R}^d$, with $\beta = ||\beta||u_1$. Define the $d \times d$ symmetric matrices $H_{ij} = (u_iu_j^{\mathrm{T}} + u_ju_i^{\mathrm{T}})/2$ and $H_j = H_{1j}$, $i, j = 1, \ldots, d$. Then

$$\beta^{\mathrm{T}}W^2\beta = \tau\sum_{j=1}^d \mathrm{tr}(WH_j)^2. \tag{44}$$

Since $S_2 = \{(1\ 2), (1)(2)\}$, the formula (33) and Lemma S2 below imply

$$\begin{aligned} E\left\{\mathrm{tr}(WH_j)^2\right\} &= 2^{2-m((1\ 2))}n^{m((1\ 2))}\mathrm{tr}(\Sigma H_j\Sigma H_j) + 2^{2-m((1)(2))}n^{m((1)(2))}\mathrm{tr}(\Sigma H_j)^2 \\ &= n\left\{(u_1^{\mathrm{T}}\Sigma u_j)^2 + u_1^{\mathrm{T}}\Sigma u_1u_j^{\mathrm{T}}\Sigma u_j\right\} + n^2(u_1^{\mathrm{T}}\Sigma u_j)^2. \end{aligned}$$

To prove (38), observe that

$$\begin{aligned} E\left(\beta^{\mathrm{T}}W^2\beta\right) &= \tau_0^2\sum_{j=1}^d E\left\{\mathrm{tr}(WH_j)^2\right\} \\ &= n(n+1)\sum_{j=1}^d \tau_0^2(u_1^{\mathrm{T}}\Sigma u_j)^2 + n\sum_{j=1}^d \tau_0^2u_1^{\mathrm{T}}\Sigma u_1u_j^{\mathrm{T}}\Sigma u_j \\ &= n(n+1)\tau_2^2 + dnm_1\tau_1^2. \end{aligned}$$

For (39), equation (33) implies

$$\begin{aligned} E\left\{\mathrm{tr}(W)\beta^{\mathrm{T}}W\beta\right\} &= \tau_0^2E\left\{\mathrm{tr}(W)\mathrm{tr}(WH_1)\right\} \\ &= 2n\tau_0^2\mathrm{tr}(\Sigma^2H_1) + n^2\tau_0^2\mathrm{tr}(\Sigma)\mathrm{tr}(\Sigma H_1) \\ &= 2n\tau_2^2 + dn^2m_1\tau_1^2. \end{aligned}$$

To prove (40), first notice that

$$E\left\{\mathrm{tr}(W)\beta^{\mathrm{T}}W^2\beta\right\} = \tau_0^2\sum_{j=1}^d E\left\{\mathrm{tr}(W)\mathrm{tr}(WH_j)^2\right\} \tag{45}$$

and that (33) implies

$$E\left\{\mathrm{tr}(W)\mathrm{tr}(WH_j)^2\right\} = \sum_{\pi \in S_3} 2^{3-m(\pi)} n^{m(\pi)} r_\pi(\Sigma)(I, H_j, H_j).$$

It is clear that

$$r_{(1\ 2\ 3)}(\Sigma)(I, H_j, H_j) = r_{(1\ 3\ 2)}(\Sigma)(I, H_j, H_j)$$
$$r_{(1\ 2)(3)}(\Sigma)(I, H_j, H_j) = r_{(1\ 3)(2)}(\Sigma)(I, H_j, H_j).$$

Thus, by Lemma S2,

$$\begin{aligned}
E\left\{\mathrm{tr}(W)\mathrm{tr}(WH_j)^2\right\} &= 8nr_{(1\ 2\ 3)}(\Sigma)(I, H_j, H_j) + 4n^2 r_{(1\ 2)(3)}(\Sigma)(I, H_j, H_j) \\
&\quad + 2n^2 r_{(1)(2\ 3)}(\Sigma)(I, H_j, H_j) + n^3 r_{(1)(2)(3)}(\Sigma)(I, H_j, H_j) \\
&= 8n\mathrm{tr}(\Sigma^2 H_j \Sigma H_j) + 4n^2 \mathrm{tr}(\Sigma^2 H_j)\mathrm{tr}(\Sigma^2 H_j) \\
&\quad + 2n^2 \mathrm{tr}(\Sigma)\mathrm{tr}(\Sigma H_j \Sigma H_j) + n^3 \mathrm{tr}(\Sigma)\mathrm{tr}(\Sigma H_j)^2 \\
&= 2n(u_1^{\mathrm{T}}\Sigma^2 u_1 u_j^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma^2 u_j + 2u_1^{\mathrm{T}}\Sigma^2 u_j u_1^{\mathrm{T}}\Sigma u_j) \\
&\quad + 4n^2 u_1^{\mathrm{T}}\Sigma^2 u_j u_1^{\mathrm{T}}\Sigma u_j + n^2 \mathrm{tr}(\Sigma)\left\{(u_1^{\mathrm{T}}\Sigma u_j)^2 + u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_j\right\} \\
&\quad + n^3 \mathrm{tr}(\Sigma)(u_1^{\mathrm{T}}\Sigma u_j)^2
\end{aligned}$$

Combining this with (45) yields

$$\begin{aligned}
E\left\{\mathrm{tr}(W)\beta^{\mathrm{T}}W^2\beta\right\} &= 2n\tau_0^2 \sum_{j=1}^d u_1^{\mathrm{T}}\Sigma^2 u_1 u_j^{\mathrm{T}}\Sigma u_j + 2n\tau_0^2 \sum_{j=1}^d u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma^2 u_j \\
&\quad + 4n(n+1)\tau_0^2 \sum_{j=1}^d u_1^{\mathrm{T}}\Sigma^2 u_j u_1^{\mathrm{T}}\Sigma u_j + n^2 \mathrm{tr}(\Sigma)\tau_0^2 \sum_{j=1}^d u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_j \\
&\quad + n^2(n+1)\mathrm{tr}(\Sigma)\tau_0^2 \sum_{j=1}^d (u_1^{\mathrm{T}}\Sigma u_j)^2 \\
&= dn(n^2+n+2)m_1\tau_2^2 + 2dnm_2\tau_1^2 + 4d^2 n(n+1)\tau_3^2 + d^2 n^2 m_1^2 \tau_1^2.
\end{aligned}$$

The proof of (41) is similar to the proof of (40). By (33) and Lemma S2,

$$\begin{aligned}
E\left\{\mathrm{tr}(WH_1)\mathrm{tr}(WH_j)^2\right\} &= 8nr_{(1\ 2\ 3)}(\Sigma)(H_1, H_j, H_j) + 4n^2 r_{(1\ 2)(3)}(\Sigma)(H_1, H_j, H_j) \\
&\quad + 2n^2 r_{(1)(2\ 3)}(\Sigma)(H_1, H_j, H_j) + n^3 r_{(1)(2)(3)}(\Sigma)(H_1, H_j, H_j) \\
&= 8n\mathrm{tr}(\Sigma H_1 \Sigma H_j \Sigma H_j) + 4n^2 \mathrm{tr}(\Sigma H_1 \Sigma H_j)\mathrm{tr}(\Sigma H_j) \\
&\quad + 2n^2 \mathrm{tr}(\Sigma H_1)\mathrm{tr}(\Sigma H_j \Sigma H_j) + n^3 \mathrm{tr}(\Sigma H_1)\mathrm{tr}(\Sigma H_j)^2
\end{aligned}$$

$$\begin{aligned}
&= 2n\left\{(u_1^{\mathrm{T}}\Sigma u_1)^2 u_j^{\mathrm{T}}\Sigma u_j + 3u_1^{\mathrm{T}}\Sigma u_1 (u_1^{\mathrm{T}}\Sigma u_j)^2\right\} \\
&\quad + 4n^2 u_1^{\mathrm{T}}\Sigma u_1 (u_1^{\mathrm{T}}\Sigma u_j)^2 + n^3 u_1^{\mathrm{T}}\Sigma u_1 (u_1^{\mathrm{T}}\Sigma u_j)^2 \\
&\quad + n^2\left\{(u_1^{\mathrm{T}}\Sigma u_1)^2 u_j^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 (u_1^{\mathrm{T}}\Sigma u_j)^2\right\} \\
&= n(n+2)(u_1^{\mathrm{T}}\Sigma u_1)^2 u_j^{\mathrm{T}}\Sigma u_j + n(n^2+5n+6) u_1^{\mathrm{T}}\Sigma u_1 (u_1^{\mathrm{T}}\Sigma u_j)^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
E(\beta^{\mathrm{T}} W\beta\beta^{\mathrm{T}} W^2\beta) &= \tau_0^4 \sum_{j=1}^d \mathrm{tr}(WH_1)\mathrm{tr}(WH_j)^2 \\
&= n(n+2)\sum_{j=1}^d \tau_0^4 (u_1^{\mathrm{T}}\Sigma u_1)^2 u_j^{\mathrm{T}}\Sigma u_j \\
&\quad + n(n^2+5n+6)\sum_{j=1}^d \tau_0^4 u_1^{\mathrm{T}}\Sigma u_1 (u_1^{\mathrm{T}}\Sigma u_j)^2 \\
&= dn(n+2)m_1\tau_1^4 + n(n^2+5n+6)\tau_1^2\tau_2^2.
\end{aligned}$$

To prove (42), consider the decomposition

$$\beta^{\mathrm{T}} W^3\beta = \tau_0^2 \sum_{i,j=1}^d \mathrm{tr}(WH_i)\mathrm{tr}(WH_j)\mathrm{tr}(WH_{ij}).$$

Equation (33) implies that

$$E\left\{\mathrm{tr}(WH_i)\mathrm{tr}(WH_j)\mathrm{tr}(WH_{ij})\right\} = \sum_{\pi\in S_3} 2^{3-m(\pi)} n^{m(\pi)} r_\pi(\Sigma)(H_i, H_j, H_{ij}).$$

Since

$$\begin{aligned}
\sum_{i,j=1}^d r_{(1\ 2\ 3)}(\Sigma)(H_i, H_j, H_{ij}) &= \sum_{i,j=1}^d r_{(1\ 3\ 2)}(\Sigma)(H_i, H_j, H_{ij}) \\
\sum_{i,j=1}^d r_{(1)(2\ 3)}(\Sigma)(H_i, H_j, H_{ij}) &= \sum_{i,j=1}^d r_{(1\ 3)(2)}(\Sigma)(H_i, H_j, H_{ij}),
\end{aligned}$$

it follows that

$$E\left(\beta^{\mathrm{T}} W^3\beta\right) = 8n\tau_0^2 \sum_{i,j=1}^d r_{(1\ 2\ 3)}(\Sigma)(H_i, H_j, H_{ij}) + 4n^2\tau_0^2 \sum_{i,j=1}^d r_{(1)(2\ 3)}(\Sigma)(H_i, H_j, H_{ij}) \quad (46)$$

$$+ 2n^2\tau_0^2 \sum_{i,j=1}^d r_{(1\ 2)(3)}(\Sigma)(H_i, H_j, H_{ij}) + n^3\tau_0^2 \sum_{i,j=1}^d r_{(1)(2)(3)}(\Sigma)(H_i, H_j, H_{ij}).$$

By Lemma S2,

$$\begin{aligned}
\tau_0^2 \sum_{i,j=1}^d r_{(1\ 2\ 3)}(\Sigma)(H_i, H_j, H_{ij}) &= \tau_0^2 \sum_{i,j=1}^d \mathrm{tr}(\Sigma H_i \Sigma H_j \Sigma H_{ij}) \\
&= \frac{\tau_0^2}{8} \sum_{i,j=1}^d \big\{ u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 (u_i^{\mathrm{T}}\Sigma u_j)^2 \\
&\quad + (u_1^{\mathrm{T}}\Sigma u_i)^2 u_j^{\mathrm{T}}\Sigma u_j + (u_1^{\mathrm{T}}\Sigma u_j)^2 u_i^{\mathrm{T}}\Sigma u_i \\
&\quad + 4u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j \big\} \\
&= \frac{1}{8}\left(d^2 m_1^2\tau_1^2 + d m_2\tau_1^2 + 2d m_1\tau_2^2 + 4\tau_3^2\right) \\
\tau_0^2 \sum_{i,j=1}^d r_{(1)(2\ 3)}(\Sigma)(H_i, H_j, H_{ij}) &= \tau_0^2 \sum_{i,j=1}^d \mathrm{tr}(\Sigma H_i)\mathrm{tr}(\Sigma H_j \Sigma H_{ij}) \\
&= \frac{\tau_0^2}{2} \sum_{i,j=1}^d u_1^{\mathrm{T}}\Sigma u_i \left(u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}}\Sigma u_j\right) \\
&= \frac{1}{2}\left(\tau_3^2 + d m_1\tau_2^2\right) \\
\tau_0^2 \sum_{i,j=1}^d r_{(1\ 2)(3)}(\Sigma)(H_i, H_j, H_{ij}) &= \tau_0^2 \sum_{i,j=1}^d \mathrm{tr}(\Sigma H_i \Sigma H_j)\mathrm{tr}(\Sigma H_{ij}) \\
&= \frac{\tau_0^2}{2} \sum_{i,j=1}^d \left(u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j\right) u_i^{\mathrm{T}}\Sigma u_j \\
&= \frac{1}{2}\left(\tau_3^2 + d m_2\tau_1^2\right) \\
\tau_0^2 \sum_{i,j=1}^d r_{(1)(2)(3)}(\Sigma)(H_i, H_j, H_{ij}) &= \tau_0^2 \sum_{i,j=1}^d \mathrm{tr}(\Sigma H_i)\mathrm{tr}(\Sigma H_j)\mathrm{tr}(\Sigma H_{ij}) \\
&= \tau_0^2 \sum_{i,j=1}^d u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j \\
&= \tau_3^2.
\end{aligned}$$

Using these results with (46) we obtain

$$\begin{aligned} E\left(\beta^{\mathrm{T}}W^3\beta\right) &= n\left(d^2m_1^2\tau_1^2 + dm_2\tau_1^2 + 2dm_1\tau_2^2 + 4\tau_3^2\right) + 2n^2\left(\tau_3^2 + dm_1\tau_2^2\right) \\ &\quad + n^2\left(\tau_3^2 + dm_2\tau_1^2\right) + n^3\tau_3^2 \\ &= d^2nm_1^2\tau_1^2 + 2dn(n+1)m_1\tau_2^2 + dn(n+1)m_2\tau_1^2 + (n^3+3n^2+4n)\tau_3^2. \end{aligned}$$

Finally, we prove (43). Similar to the proof of (41)–(42), we have the decomposition

$$(\beta^{\mathrm{T}}W^2\beta)^2 = \tau_0^4\sum_{i,j=1}^{d}\mathrm{tr}(WH_i)^2\mathrm{tr}(WH_j)^2.$$

By (33),

$$E\left\{\mathrm{tr}(WH_i)^2\mathrm{tr}(WH_j)^2\right\} = \sum_{\pi\in S_4}2^{4-m(\pi)}n^{m(\pi)}r_\pi(\Sigma)(H_i,H_i,H_j,H_j).$$

It follows that

$$E\left\{(\beta^{\mathrm{T}}W^2\beta)^2\right\} = \sum_{\pi\in S_4}2^{4-m(\pi)}n^{m(\pi)}\tilde{r}_\pi,$$

where

$$\tilde{r}_\pi = \sum_{i,j=1}^{d}\tau_0^4 r_\pi(\Sigma)(H_i,H_i,H_j,H_j).$$

One can easily see that

$$\begin{aligned} \tilde{r}_{(1\ 2\ 3\ 4)} &= \tilde{r}_{(1\ 2\ 4\ 3)} = \tilde{r}_{(1\ 3\ 4\ 2)} = \tilde{r}_{(1\ 4\ 3\ 2)} \\ \tilde{r}_{(1\ 3\ 2\ 4)} &= \tilde{r}_{(1\ 4\ 2\ 3)} \\ \tilde{r}_{(1)(2\ 3\ 4)} &= \tilde{r}_{(1)(2\ 4\ 3)} = \tilde{r}_{(1\ 3\ 4)(2)} = \tilde{r}_{(1\ 4\ 3)(2)} = \tilde{r}_{(1\ 2\ 3)(4)} \\ &= \tilde{r}_{(1\ 3\ 2)(4)} = \tilde{r}_{(1\ 2\ 4)(3)} = \tilde{r}_{(1\ 4\ 2)(3)} \\ \tilde{r}_{(1\ 3)(2\ 4)} &= \tilde{r}_{(1\ 4)(2\ 3)} \\ \tilde{r}_{(1\ 2)(3)(4)} &= \tilde{r}_{(1)(2)(3\ 4)} \\ \tilde{r}_{(1\ 3)(2)(4)} &= \tilde{r}_{(1\ 4)(2)(3)} = \tilde{r}_{(1)(3)(2\ 4)} = \tilde{r}_{(1)(4)(2\ 3)}. \end{aligned}$$

Thus,

$$\begin{aligned} E\left\{(\beta^{\mathrm{T}}W^2\beta)^2\right\} &= 32n\tilde{r}_{(1\ 2\ 3\ 4)} + 16n\tilde{r}_{(1\ 3\ 2\ 4)} + 32n^2\tilde{r}_{(1)(2\ 3\ 4)} + 8n^2\tilde{r}_{(1\ 3)(2\ 4)} \\ &\quad + 4n^2\tilde{r}_{(1\ 2)(3\ 4)} + 4n^3\tilde{r}_{(1\ 2)(3)(4)} + 8n^3\tilde{r}_{(1\ 3)(2)(4)} + n^4\tilde{r}_{(1)(2)(3)(4)}. \end{aligned} \tag{47}$$

It only remains to evaluate the $\tilde{r}_\pi$. It follows from Lemma S2 that

$$\begin{aligned}
\tilde{r}_{(1\ 2\ 3\ 4)} &= \sum_{i,j=1}^d \tau_0^4 \mathrm{tr}(\Sigma H_i \Sigma H_i \Sigma H_j \Sigma H_j) \\
&= \sum_{i,j=1}^d \frac{\tau_0^4}{16} \big\{ 2(u_1^{\mathrm{T}} \Sigma u_i)^2 (u_1^{\mathrm{T}} \Sigma u_j)^2 + 3 u_1^{\mathrm{T}} \Sigma u_1 (u_1^{\mathrm{T}} \Sigma u_i)^2 u_j^{\mathrm{T}} \Sigma u_j \\
&\quad + 6 u_1^{\mathrm{T}} \Sigma u_1 u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \Sigma u_j + 3 u_1^{\mathrm{T}} \Sigma u_1 (u_1^{\mathrm{T}} \Sigma u_j)^2 u_i^{\mathrm{T}} \Sigma u_i \\
&\quad + (u_1^{\mathrm{T}} \Sigma u_1)^2 u_i^{\mathrm{T}} \Sigma u_i u_j \Sigma u_j + (u_1^{\mathrm{T}} \Sigma u_1)^2 (u_i^{\mathrm{T}} \Sigma u_j)^2 \big\} \\
&= \frac{1}{16} \big( 2\tau_2^4 + 6 d m_1 \tau_1^2 \tau_2^2 + 6 \tau_1^2 \tau_3^2 + d^2 m_1^2 \tau_1^4 + d m_2 \tau_1^4 \big) \\
\tilde{r}_{(1\ 3\ 2\ 4)} &= \sum_{i,j=1}^d \tau_0^4 \mathrm{tr}(\Sigma H_i \Sigma H_j \Sigma H_i \Sigma H_j) \\
&= \sum_{i,j=1}^d \frac{\tau_0^4}{8} \big\{ (u_1^{\mathrm{T}} \Sigma u_i)^2 (u_1^{\mathrm{T}} \Sigma u_j)^2 \\
&\quad + 6 u_1^{\mathrm{T}} \Sigma u_1 u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \Sigma u_j + (u_1^{\mathrm{T}} \Sigma u_1)^2 (u_i^{\mathrm{T}} \Sigma u_j)^2 \big\} \\
&= \frac{1}{8} (\tau_2^4 + 6 \tau_1^2 \tau_3^2 + d m_2 \tau_1^4) \\
\tilde{r}_{(1)(2\ 3\ 4)} &= \sum_{i,j=1}^d \tau_0^4 \mathrm{tr}(\Sigma H_i) \mathrm{tr}(\Sigma H_i \Sigma H_j \Sigma H_j) \\
&= \sum_{i,j=1}^d \frac{\tau_0^4}{4} \big\{ (u_1^{\mathrm{T}} \Sigma u_i)^2 (u_1^{\mathrm{T}} \Sigma u_j)^2 + u_1^{\mathrm{T}} \Sigma u_1 (u_1^{\mathrm{T}} \Sigma u_i)^2 u_j^{\mathrm{T}} \Sigma u_j \\
&\quad + 2 u_1^{\mathrm{T}} \Sigma u_1 u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \Sigma u_j \big\} \\
&= \frac{1}{4} (\tau_2^4 + d m_1 \tau_1^2 \tau_2^2 + 2 \tau_1^2 \tau_3^2) \\
\tilde{r}_{(1\ 3)(2\ 4)} &= \sum_{i,j=1}^d \tau_0^4 \mathrm{tr}(\Sigma H_i \Sigma H_j)^2 \\
&= \sum_{i,j=1}^d \frac{\tau_0^4}{4} \big\{ u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j + u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j \big\}^2 \\
&= \frac{1}{4} (\tau_2^4 + d m_2 \tau_1^4 + 2 \tau_1^2 \tau_3^2)
\end{aligned}$$

$$\begin{aligned}
\tilde{r}_{(1\ 2)(3\ 4)} &= \sum_{i,j=1}^{d} \tau_0^4 \mathrm{tr}(\Sigma H_i \Sigma H_i)\mathrm{tr}(\Sigma H_j \Sigma H_j) \\
&= \sum_{i,j=1}^{d} \frac{\tau_0^4}{4} \left\{(u_1^{\mathrm{T}}\Sigma u_i)^2 + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_i\right\}\left\{(u_1^{\mathrm{T}}\Sigma u_j)^2 + u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_j\right\} \\
&= \frac{1}{4}(\tau_2^4 + 2dm_1\tau_1^2\tau_2^2 + d^2 m_1^2 \tau_1^4) \\
\tilde{r}_{(1\ 2)(3)(4)} &= \sum_{i,j=1}^{d} \tau_0^4 \mathrm{tr}(\Sigma H_i \Sigma H_i)\mathrm{tr}(\Sigma H_j)^2 \\
&= \sum_{i,j=1}^{d} \frac{\tau_0^4}{2} \left\{(u_1^{\mathrm{T}}\Sigma u_i)^2 + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_i\right\}(u_1^{\mathrm{T}}\Sigma u_j)^2 \\
&= \frac{1}{2}(\tau_2^4 + dm_1\tau_1^2\tau_2^2) \\
\tilde{r}_{(1\ 3)(2)(4)} &= \sum_{i,j=1}^{d} \tau_0^4 \mathrm{tr}(\Sigma H_i \Sigma H_j)\mathrm{tr}(\Sigma H_i)\mathrm{tr}(\Sigma H_j) \\
&= \sum_{i,j=1}^{d} \frac{\tau_0^4}{2} \left\{u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j\right\} u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j \\
&= \frac{1}{2}(\tau_2^4 + \tau_1^2\tau_3^2) \\
\tilde{r}_{(1)(2)(3)(4)} &= \sum_{i,j=1}^{d} \tau_0^4 \mathrm{tr}(\Sigma H_i)^2\mathrm{tr}(\Sigma H_j)^2 \\
&= \sum_{i,j=1}^{d} \tau_0^4 (u_1^{\mathrm{T}}\Sigma u_i)^2 (u_1^{\mathrm{T}}\Sigma u_j)^2 \\
&= \tau_2^4.
\end{aligned}$$

Combining this with (47), we conclude that

$$\begin{aligned}
E\left\{(\beta^{\mathrm{T}}W^2\beta)^2\right\} &= 32n\tilde{r}_{(1\ 2\ 3\ 4)} + 16n\tilde{r}_{(1\ 3\ 2\ 4)} + 32n^2\tilde{r}_{(1)(2\ 3\ 4)} + 8n^2\tilde{r}_{(1\ 3)(2\ 4)} \\
&\quad + 4n^2\tilde{r}_{(1\ 2)(3\ 4)} + 4n^3\tilde{r}_{(1\ 2)(3)(4)} + 8n^3\tilde{r}_{(1\ 3)(2)(4)} + n^4\tilde{r}_{(1)(2)(3)(4)} \\
&= 2n(2\tau_2^4 + 6dm_1\tau_1^2\tau_2^2 + 6\tau_1^2\tau_3^2 + d^2m_1^2\tau_1^4 + dm_2\tau_1^4) \\
&\quad + 2n(\tau_2^4 + 6\tau_1^2\tau_3^2 + dm_2\tau_1^4) + 8n^2(\tau_2^4 + dm_1\tau_1^2\tau_2^2 + 2\tau_1^2\tau_3^2) \\
&\quad + 2n^2(\tau_2^4 + dm_2\tau_1^4 + 2\tau_1^2\tau_3^2) + n^2(\tau_2^4 + 2dm_1\tau_1^2\tau_2^2 + d^2m_1^2\tau_1^4)
\end{aligned}$$

$$
\begin{aligned}
&+ 2n^3(\tau_2^4 + dm_1\tau_1^2\tau_2^2) + 4n^3(\tau_2^4 + \tau_1^2\tau_3^2) + n^4\tau_2^4 \\
&= (n^4 + 6n^3 + 11n^2 + 6n)\tau_2^4 + d(2n^3 + 10n^2 + 12n)m_1\tau_1^2\tau_2^2 \\
&\quad + (4n^3 + 20n^2 + 24n)\tau_1^2\tau_3^2 + d^2(n^2 + 2n)m_1^2\tau_1^4 + d(2n^2 + 4n)m_2\tau_1^4 \\
&= d^2n(n+2)m_1^2\tau_1^4 + 2dn(n+2)(n+3)m_1\tau_1^2\tau_2^2 + 2dn(n+2)m_2\tau_1^4 \\
&\quad + 4n(n+2)(n+3)\tau_1^2\tau_3^2 + n(n+1)(n+2)(n+3)\tau_2^4.
\end{aligned}
$$

This concludes the proof of Proposition S1. □

**Lemma S2.** *Let $u_1, \ldots, u_d \in \mathbb{R}^d$ and define $H_j = (u_1u_j^{\mathrm{T}} + u_ju_1^{\mathrm{T}})/2$. For integers $1 \leq i, j \leq d$, we have*

$$
\begin{aligned}
\mathrm{tr}(\Sigma H_{ij}) &= u_i^{\mathrm{T}}\Sigma u_j && (48)\\
\mathrm{tr}(\Sigma H_i\Sigma H_j) &= \frac{1}{2}\left(u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j\right) && (49)\\
\mathrm{tr}(\Sigma H_i\Sigma H_{ij}) &= \frac{1}{2}\left(u_1^{\mathrm{T}}\Sigma u_i u_i\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_i\right) && (50)\\
\mathrm{tr}(\Sigma^2 H_i\Sigma H_j) &= \frac{1}{4}\Big(u_1^{\mathrm{T}}\Sigma^2 u_1 u_i^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma^2 u_j \\
&\quad + u_1^{\mathrm{T}}\Sigma^2 u_i u_1^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma^2 u_j\Big) && (51)\\
\mathrm{tr}(\Sigma H_i\Sigma H_j\Sigma H_j) &= \frac{1}{4}\left\{u_1^{\mathrm{T}}\Sigma u_i(u_1^{\mathrm{T}}\Sigma u_j)^2 + u_1^{\mathrm{T}}\Sigma u_1 u_1^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}}\Sigma u_j\right. \\
&\quad \left. + 2u_1^{\mathrm{T}}\Sigma u_1 u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j\right\} && (52)\\
\mathrm{tr}(\Sigma H_i\Sigma H_j\Sigma H_{ij}) &= \frac{1}{8}\left\{u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1(u_i^{\mathrm{T}}\Sigma u_j)^2\right. \\
&\quad + (u_1^{\mathrm{T}}\Sigma u_i)^2 u_j^{\mathrm{T}}\Sigma u_j + (u_1^{\mathrm{T}}\Sigma u_j)^2 u_i^{\mathrm{T}}\Sigma u_i \\
&\quad \left. + 4u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j\right\} && (53)\\
\mathrm{tr}(\Sigma H_i\Sigma H_i\Sigma H_j\Sigma H_j) &= \frac{1}{16}\left\{2(u_1^{\mathrm{T}}\Sigma u_i)^2(u_1^{\mathrm{T}}\Sigma u_j)^2 3u_1^{\mathrm{T}}\Sigma u_1(u_1^{\mathrm{T}}\Sigma u_i)^2 u_j^{\mathrm{T}}\Sigma u_j\right. \\
&\quad + 6u_1^{\mathrm{T}}\Sigma u_1 u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j + 3u_1^{\mathrm{T}}\Sigma u_1(u_1^{\mathrm{T}}\Sigma u_j)^2 u_i^{\mathrm{T}}\Sigma u_i \\
&\quad \left. + (u_1^{\mathrm{T}}\Sigma u_1)^2 u_i^{\mathrm{T}}\Sigma u_i u_j\Sigma u_j + (u_1^{\mathrm{T}}\Sigma u_1)^2(u_i^{\mathrm{T}}\Sigma u_j)^2\right\} && (54)\\
\mathrm{tr}(\Sigma H_i\Sigma H_j\Sigma H_i\Sigma H_j) &= \frac{1}{8}\left\{(u_1^{\mathrm{T}}\Sigma u_i)^2(u_1^{\mathrm{T}}\Sigma u_j)^2 + 6u_1^{\mathrm{T}}\Sigma u_1 u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j\right. \\
&\quad \left. + (u_1^{\mathrm{T}}\Sigma u_1)^2(u_i^{\mathrm{T}}\Sigma u_j)^2\right\} && (55)
\end{aligned}
$$

*Proof.* The identity (48) is trivial. To prove (49), we have

$$
\mathrm{tr}(\Sigma H_i\Sigma H_j) = \frac{1}{4}\mathrm{tr}\left\{\Sigma(u_1u_i^{\mathrm{T}} + u_iu_1^{\mathrm{T}})\Sigma(u_1u_j^{\mathrm{T}} + u_ju_1^{\mathrm{T}})\right\}
$$

$$= \frac{1}{4}\mathrm{tr}\left(\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\right)$$
$$= \frac{1}{2}\left(u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j\right).$$

Equation (50) follows from

$$\mathrm{tr}(\Sigma H_i \Sigma H_{ij}) = \frac{1}{4}\mathrm{tr}\left\{\Sigma(u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}})\Sigma(u_i u_j^{\mathrm{T}} + u_j u_i^{\mathrm{T}})\right\}$$
$$= \frac{1}{4}\mathrm{tr}\left(\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\right)$$
$$= \frac{1}{2}\left(u_1^{\mathrm{T}}\Sigma u_i u_i^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_i\right).$$

For (51), we have

$$\mathrm{tr}(\Sigma^2 H_i \Sigma H_j) = \frac{1}{4}\mathrm{tr}\left\{\Sigma^2(u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}})\Sigma(u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}})\right\}$$
$$= \frac{1}{4}\mathrm{tr}\left(\Sigma^2 u_1 u_i^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma^2 u_1 u_i^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}} + \Sigma^2 u_i u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma^2 u_i u_1^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\right)$$
$$= \frac{1}{4}\left(u_1^{\mathrm{T}}\Sigma u_i u_1^{\mathrm{T}}\Sigma^2 u_j + u_1^{\mathrm{T}}\Sigma^2 u_1 u_i^{\mathrm{T}}\Sigma u_j + u_1^{\mathrm{T}}\Sigma u_1 u_i^{\mathrm{T}}\Sigma^2 u_j + u_1^{\mathrm{T}}\Sigma^2 u_i u_1 \Sigma u_j\right).$$

To prove (52)–(53), observe that

$$\mathrm{tr}(\Sigma H_i \Sigma H_j \Sigma H_j) = \frac{1}{8}\mathrm{tr}\left\{\Sigma(u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}})\Sigma(u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}})\Sigma(u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}})\right\}$$
$$= \frac{1}{8}\mathrm{tr}\left(\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\right.$$
$$+ \Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}$$
$$+ \Sigma u_i u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}$$
$$\left.+\Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\Sigma u_j u_1^{\mathrm{T}}\right)$$
$$= \frac{1}{4}\left\{u_1^{\mathrm{T}}\Sigma u_i (u_1^{\mathrm{T}}\Sigma u_j)^2 + u_1^{\mathrm{T}}\Sigma u_1 u_1^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}}\Sigma u_j + 2u_1^{\mathrm{T}}\Sigma u_1 u_1^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\Sigma u_j\right\}$$

and

$$\mathrm{tr}(\Sigma H_i \Sigma H_j \Sigma H_{ij}) = \frac{1}{8}\mathrm{tr}\left\{\Sigma(u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}})\Sigma(u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}})\Sigma(u_i u_j^{\mathrm{T}} + u_j u_i^{\mathrm{T}})\right\}$$
$$= \frac{1}{8}\mathrm{tr}\left(\Sigma u_1 u_i^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_i u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}}\Sigma u_1 u_j^{\mathrm{T}}\Sigma u_j u_i^{\mathrm{T}}\right.$$

$$
\begin{aligned}
&+ \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_i u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \\
&+ \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_i u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \\
&+ \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_i u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \big) \\
&= \frac{1}{8} \left\{ u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_i u_j^{\mathrm{T}} \Sigma u_j + u_1^{\mathrm{T}} \Sigma u_1 (u_i^{\mathrm{T}} \Sigma u_j)^2 + (u_1^{\mathrm{T}} \Sigma u_i)^2 u_j^{\mathrm{T}} \Sigma u_j \right. \\
&\left. + (u_1^{\mathrm{T}} \Sigma u_j)^2 u_i^{\mathrm{T}} \Sigma u_i + 4 u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \Sigma u_j \right\}.
\end{aligned}
$$

Finally, to prove (54)–(55), we have

$$
\begin{aligned}
&\mathrm{tr}(\Sigma H_i \Sigma H_i \Sigma H_j \Sigma H_j) \\
&\quad = \frac{1}{16} \mathrm{tr} \left\{ \Sigma (u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}}) \Sigma (u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}}) \Sigma (u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}}) \Sigma (u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}}) \right\} \\
&\quad = \frac{1}{16} \mathrm{tr} \big( \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \big) \\
&\quad = \frac{1}{16} \left\{ 2 (u_1^{\mathrm{T}} \Sigma u_i)^2 (u_1^{\mathrm{T}} \Sigma u_j)^2 + 3 u_1^{\mathrm{T}} \Sigma u_1 (u_1^{\mathrm{T}} \Sigma u_i)^2 u_j^{\mathrm{T}} \Sigma u_j \right. \\
&\qquad + 6 u_1^{\mathrm{T}} \Sigma u_1 u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \Sigma u_j + 3 u_1^{\mathrm{T}} \Sigma u_1 (u_1^{\mathrm{T}} \Sigma u_j)^2 u_i^{\mathrm{T}} \Sigma u_i \\
&\qquad \left. + (u_1^{\mathrm{T}} \Sigma u_1)^2 u_i^{\mathrm{T}} \Sigma u_i u_j \Sigma u_j + (u_1^{\mathrm{T}} \Sigma u_1)^2 (u_i^{\mathrm{T}} \Sigma u_j)^2 \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{tr}(\Sigma H_i \Sigma H_j \Sigma H_i \Sigma H_j) \\
&\quad = \frac{1}{16} \mathrm{tr} \left\{ \Sigma (u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}}) \Sigma (u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}}) \Sigma (u_1 u_i^{\mathrm{T}} + u_i u_1^{\mathrm{T}}) \Sigma (u_1 u_j^{\mathrm{T}} + u_j u_1^{\mathrm{T}}) \right\} \\
&\quad = \frac{1}{16} \mathrm{tr} \big( \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&\qquad + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}}
\end{aligned}
$$

$$
\begin{aligned}
&+ \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&+ \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&+ \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&+ \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_1 u_i^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \\
&+ \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_1 u_j^{\mathrm{T}} + \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_1^{\mathrm{T}} \big) \\
&= \frac{1}{8} \left\{ (u_1^{\mathrm{T}} \Sigma u_i)^2 (u_1^{\mathrm{T}} \Sigma u_j)^2 + 6 u_1^{\mathrm{T}} \Sigma u_1 u_1^{\mathrm{T}} \Sigma u_i u_1^{\mathrm{T}} \Sigma u_j u_i^{\mathrm{T}} \Sigma u_j + (u_1^{\mathrm{T}} \Sigma u_1)^2 (u_i^{\mathrm{T}} \Sigma u_j)^2 \right\}.
\end{aligned}
$$

□