The Annals of Statistics
2017, Vol. 45, No. 1, 386–414
DOI: 10.1214/16-AOS1456

FLEXIBLE RESULTS FOR QUADRATIC FORMS WITH APPLICATIONS TO VARIANCE COMPONENTS ESTIMATION

BY LEE H. DICKER[1] AND MURAT A. ERDOGDU

Rutgers University and Stanford University

We derive convenient uniform concentration bounds and finite sample multivariate normal approximation results for quadratic forms, then describe some applications involving variance components estimation in linear random-effects models. Random-effects models and variance components estimation are classical topics in statistics, with a corresponding well-established asymptotic theory. However, our finite sample results for quadratic forms provide additional flexibility for easily analyzing random-effects models in nonstandard settings, which are becoming more important in modern applications (e.g., genomics). For instance, in addition to deriving novel non-asymptotic bounds for variance components estimators in classical linear random-effects models, we provide a concentration bound for variance components estimators in linear models with correlated random-effects and discuss an application involving sparse random-effects models. Our general concentration bound is a uniform version of the Hanson–Wright inequality. The main normal approximation result in the paper is derived using Reinert and Röllin [*Ann. Probab.* (2009) **37** 2150–2173] embedding technique for Stein's method of exchangeable pairs.

1. Introduction. Suppose that $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_d)^\top \in \mathbb{R}^d$ is a random vector with independent components satisfying $\mathbb{E}(\zeta_j) = 0$, $j = 1, \ldots, d$. Additionally, let Q be a $d \times d$ positive semidefinite matrix with real (non-random) entries. Quadratic forms $\boldsymbol{\zeta}^\top Q \boldsymbol{\zeta}$ have been studied for decades in statistics [Sevastyanov (1961); Whittle (1964); Hanson and Wright (1971); Hall (1984); Götze and Tikhomirov (1999, 2002); Chatterjee (2008)]. This paper is largely motivated by recent applications involving random-effects models, which rely heavily on properties of quadratic forms [e.g., Jiang et al. (2016); de Los Campos, Sorensen and Gianola (2015); Dicker and Erdogdu (2016a)]. In the first part of the paper, we give two new finite sample bounds for quadratic forms—a uniform concentration inequality (Theorem 1) and a normal approximation result (Theorem 2)—which may be useful in a variety of statistical applications. The second part of the paper focuses on applications of Theorems 1–2 related to variance components estimation in linear random-effects models, including nonstandard models with corre-

Received September 2015; revised February 2016.

[1]Supported by NSF Grant DMS-12-08785 and NSF CAREER Award DMS-14-54817.

MSC2010 subject classifications. Primary 62F99; secondary 62E17, 62F12.

Key words and phrases. Hanson–Wright inequality, random-effects models, model misspecification, Stein's method, uniform concentration bounds.

lated random effects and a sparse random-effects model, which was previously studied in the context of genome wide association studies.

Theorems 1 and 2 are the main theoretical results in the paper. We rate the novelty of our normal approximation result Theorem 2, which is a multivariate normal approximation results proved via Stein's method for exchangeable pairs, higher than that of Theorem 1. However, the main emphasis of both results is convenience for use in applications.

Our concentration bound, Theorem 1, is a uniform version of the Hanson–Wright inequality for quadratic forms. The method of proof for Theorem 1 is relatively standard—combining a chaining argument from empirical process theory [e.g., Chapter 3 of Van de Geer (2000)] with the pointwise-bound of the original Hanson–Wright inequality—and it should be possible to generalize our result to larger classes of quadratic forms, similar to Adamczak (2015). However, we note that while Theorem 1 is restricted to relatively simple (Lipschitz) classes of quadratic forms, it is not a corollary of the uniform bounds in Adamczak (2015), which require a stronger condition on the distribution of $\boldsymbol{\zeta}$ (see the comments in Section 3.1 following the statement of Theorem 1).

Theorem 2 is a normal approximation result for vectors of quadratic forms. Most of the existing normal approximation results for quadratic forms are asymptotic results [Whittle (1964); Hall (1984); Jiang (1996)], require the random variables ζ_i to be i.i.d. [Hall (1984); Götze and Tikhomirov (1999, 2002); Chatterjee (2008)], or have other limitations [e.g., Gaussian ζ_i; Sevastyanov (1961)]. Theorem 2 gives a nonasymptotic normal approximation bound, which applies to $\boldsymbol{\zeta}$ with independent (but not necessarily identically distributed) sub-Gaussian components. Furthermore, in contrast with most existing results on quadratic forms, which are predominantly univariate, Theorem 2 is a multivariate result, which applies to vectors of quadratic forms $(\boldsymbol{\zeta}^\top Q_1 \boldsymbol{\zeta}, \ldots, \boldsymbol{\zeta}^\top Q_K \boldsymbol{\zeta})$, for positive semidefinite matrices $Q_1, \ldots, Q_K$ (the applications to random-effects models considered in Section 4 require $K = 2$). The proof of Theorem 2 relies on Stein's method of exchangeable pairs and follows the embedding approach of Reinert and Röllin (2009). Theorem 2 and its proof shares similarities with Proposition 3.1 of Chatterjee (2008). However, Proposition 3.1 of Chatterjee (2008) applies only to a single quadratic form $\boldsymbol{\zeta}^\top Q \boldsymbol{\zeta}$ in i.i.d. Rademacher random variables ζ_i satisfying $\mathbb{P}(\zeta_i = 1) = \mathbb{P}(\zeta_i = -1) = 1/2$.

Linear random-effects models are studied in Section 4. Asymptotic results for quadratic forms serve as the theoretical underpinning for many applications involving random-effects models [Hartley and Rao (1967); Jiang (1996, 1998)]. However, new applications of random-effects models in genomics have pushed the boundaries of existing theoretical results [Yang et al. (2010, 2014); Golan and Rosset (2011); Speed et al. (2012); Zaitlen and Kraft (2012); Jiang et al. (2016); de Los Campos, Sorensen and Gianola (2015)]. In Section 4, we present

new nonasymptotic bounds for variance components estimators in linear random-effects models. To our knowledge, these are the first finite sample results on the statistical properties of variance components estimators. Many now-classical asymptotic results for random-effects models [e.g., Jiang (1996)] follow as corollaries of our finite sample results in Section 4. More significantly, nonasymptotic bounds, like those in this paper, provide increased flexibility for use in applications. In particular, our results can be easily applied in nonstandard settings, where it is less clear how to adapt the existing asymptotic theory; see, for example, Proposition 2, which applies to random-effects models with correlated random-effects, and Section 4.4 for an application involving sparse random-effects models.

The rest of the paper proceeds as follows. Some basic notation is introduced in Section 2. The main theoretical results are stated in Section 3. Linear random-effects models are studied in Section 4. Theorems 1–2 are proved in the Appendices; other results are proved in the supplemental article [Dicker and Erdogdu (2016b)].

2. Notation. If $\mathbf{u} = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p$, then $\|\mathbf{u}\| = (u_1^2 + \cdots + u_p^2)^{1/2}$ is its Euclidean norm. For a $d \times m$ matrix $A = (a_{ij})$ with real entries, let $\|A\| = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ and $\|A\|_{\mathrm{HS}} = \{\sum_{i=1}^d \sum_{j=1}^m a_{ij}^2\}^{1/2}$ be the operator norm and the Hilbert–Schmidt (Frobenius) norm of A, respectively. If $f : \mathbb{R}^m \to \mathbb{R}$ is a function with kth order derivatives, define

$$|f|_j = \sup_{\substack{\mathbf{x}\in\mathbb{R}^m, \\ 1\le i_1,\ldots,i_j\le m}} \left|\frac{\partial^j}{\partial x_{i_1}\cdots\partial x_{i_j}} f(\mathbf{x})\right|, \qquad j = 1, \ldots, k,$$

and let $|f|_0 = \sup_{\mathbf{x}\in\mathbb{R}^m} |f(\mathbf{x})|$. Additionally, define $\mathcal{C}_b^k(\mathbb{R}^m) = \{f : \mathbb{R}^m \to \mathbb{R}; |f|_j < \infty, j = 0, 1, \ldots, k\}$ to be the class of real-valued functions on $\mathbb{R}^m$ with bounded derivatives up to order k. Finally, following Vershynin (2012), let $\|\zeta\|_{\psi_2} = \sup_{r\ge 1} r^{-1/2}\mathbb{E}(|\zeta|^r)^{1/r}$ be the sub-Gaussian norm of the real-valued random variable ζ.

3. Results for quadratic forms.

3.1. *Uniform concentration bound.* The Hanson–Wright inequality is a classical probabilistic bound for quadratic forms, which has been the subject of renewed attention recently in applications related to random matrix theory [e.g., Hsu, Kakade and Zhang (2012); Rudelson and Vershynin (2013); Adamczak (2015)]. Theorem 1 is a uniform version of the Hanson–Wright inequality, which applies to families of quadratic forms $\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta}$, where $Q(\mathbf{u})$ is a matrix function of $\mathbf{u} \in \mathbb{R}^K$. As illustrated in Section 4, Theorem 1 has applications in the analysis of random-effects models; more broadly, it has applications in M-estimation and maximum likelihood problems with non-i.i.d. data.

THEOREM 1. *Let $0 < R < \infty$ and let $t_1(\mathbf{u}), \dots, t_m(\mathbf{u})$ be real-valued Lipschitz functions on $[0, R]^K \subseteq \mathbb{R}^K$, satisfying*

$$\max_{i=1,\dots,m} |t_i(\mathbf{u}) - t_i(\mathbf{u}')| \le L\|\mathbf{u} - \mathbf{u}'\|, \qquad \mathbf{u}, \mathbf{u}' \in [0, R]^K, \tag{1}$$

for some constant $0 < L < \infty$. Let $T(\mathbf{u}) = \operatorname{diag}\{t_1(\mathbf{u}), \dots, t_m(\mathbf{u})\}$, let V be an $d \times m$ matrix, and define $Q(\mathbf{u}) = VT(\mathbf{u})V^\top$. Additionally, let $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_d)^\top \in \mathbb{R}^d$, where $\zeta_1, \dots, \zeta_d$ are independent mean 0 sub-Gaussian random variables satisfying

$$\max_{i=1,\dots,d} \|\zeta_i\|_{\psi_2} \le \gamma \tag{2}$$

for some constant $\gamma \in (0, \infty)$. Then there exists an absolute constant $C \in (0, \infty)$ such that

$$\begin{aligned}\mathbb{P}\Big[\sup_{\mathbf{u}\in[0,R]^K} |\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta} - \mathbb{E}\{\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta}\}| > r\Big] \\ \le C\exp\Big[-\frac{1}{C}\min\Big\{\frac{r^2}{\gamma^4\|V^\top V\|^2(\|T(0)\|_{\mathrm{HS}}^2 + KL^2R^2m)}, \\ \frac{r}{\gamma^2\|V^\top V\|(\|T(0)\| + K^{1/2}LR)}\Big\}\Big],\end{aligned}$$

whenever $r^2 \ge C\gamma^4\|V^\top V\|^2K^3L^2R^2m$.

Theorem 1 is proved in Appendix A. As mentioned in Section 1, the proof combines a chaining argument with the classical (nonuniform) Hanson–Wright inequality. In a typical application, the dimension K will be small (in Section 4, we use Theorem 1 with $K = 1$) and m, d may be large. A uniform Hanson–Wright inequality, with a similar upper bound, is also given in Adamczak (2015). However, Adamczak's result applies to random vectors satisfying a relatively strong concentration property and does not cover sub-Gaussian random vectors satisfying only (2); see Remark 2.9 following Theorem 2.5 in Adamczak (2015).

3.2. *Normal approximation.* The main result of this section is Theorem 2, a multivariate normal approximation result for vectors of quadratic forms $(\boldsymbol{\zeta}^\top Q_1\boldsymbol{\zeta}, \dots, \boldsymbol{\zeta}^\top Q_K\boldsymbol{\zeta})^\top \in \mathbb{R}^K$. Theorem 2 may be viewed as a generalization of Proposition 3.1 in Chatterjee (2008), which applies to a single quadratic for $\boldsymbol{\zeta}^\top Q\boldsymbol{\zeta}$ in Rademacher random variables ζ_j satisfying $\mathbb{P}(\zeta_j = \pm 1) = 1/2$ (though our bound in Theorem 2 is not as tight as Chatterjee's; see the discussion after the statement of the theorem).

THEOREM 2. *Let $\zeta_1, \dots, \zeta_d$ be independent sub-Gaussian random variables with mean 0 and variance 1, and assume that they satisfy (2). Let*

$\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_d)^\top \in \mathbb{R}^d$. *Additionally, for* $k = 1, \ldots, K$, *let* $Q_k = (q_{ij}^{(k)})$ *be an* $d \times d$ *positive semidefinite matrix and let* $\breve{Q}_k = \operatorname{diag}(q_{11}^{(k)}, \ldots, q_{dd}^{(k)})$. *Define* $w_k = \boldsymbol{\zeta}^\top Q_k \boldsymbol{\zeta} - \operatorname{tr}(Q_k)$, $\breve{w}_k = \boldsymbol{\zeta}^\top \breve{Q}_k \boldsymbol{\zeta} - \operatorname{tr}(Q_k)$, *and*

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \breve{w}_1 \\ \vdots \\ w_K \\ \breve{w}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{\zeta}^\top Q_1 \boldsymbol{\zeta} - \operatorname{tr}(Q_1) \\ \boldsymbol{\zeta}^\top \breve{Q}_1 \boldsymbol{\zeta} - \operatorname{tr}(Q_1) \\ \vdots \\ \boldsymbol{\zeta}^\top Q_K \boldsymbol{\zeta} - \operatorname{tr}(Q_K) \\ \boldsymbol{\zeta}^\top \breve{Q}_K \boldsymbol{\zeta} - \operatorname{tr}(Q_K) \end{bmatrix} \in \mathbb{R}^{2K}.$$

Finally, let $\mathbf{z} \sim N(0, I_{2K})$ *and* $V = \operatorname{Cov}(\mathbf{w})$. *There is an absolute constant* $0 < C < \infty$ *such that*

$$\begin{aligned} |\mathbb{E}\{f(\mathbf{w})\} - \mathbb{E}\{f(V^{1/2}\mathbf{z})\}| &\le C(\gamma+1)^8 K^{3/2} d^{1/2} |f|_2 \Big(\max_{k=1,\ldots,K} \|Q_k\| \Big)^2 \\ &\quad + C(\gamma+1)^8 K^3 d |f|_3 \Big(\max_{k=1,\ldots,K} \|Q_k\| \Big)^3, \end{aligned} \tag{3}$$

for all three-times differentiable functions $f : \mathbb{R}^{2K} \to \mathbb{R}$.

A proof of Theorem 2 may be found in Appendix B. The proof is based on Stein's method of exchangeable pairs and the embedding technique from Reinert and Röllin (2009). In more detail, the embedding technique leads us to consider the vector $\mathbf{w}$, formed from pairs $(w_k, \breve{w}_k)$, as opposed to just $(w_1, \ldots, w_k)^\top$, which might be more natural from an applications viewpoint. Indeed, let $\mathbf{w}' \in \mathbb{R}^{2K}$ be defined in the exact same way as $\mathbf{w}$, except that one of the random variables ζ_i is selected at random and replaced with an independent copy. Then we have the simple relationship (derived in detail in Appendix B)

$$\mathbb{E}(\mathbf{w}' - \mathbf{w} | \boldsymbol{\zeta}) = -\Lambda_K \mathbf{w}, \tag{4}$$

where

$$\Lambda_1 = \begin{bmatrix} \frac{2}{d} & -\frac{1}{d} \\ 0 & \frac{1}{d} \end{bmatrix} \in \mathbb{R}^{2\times 2}, \qquad \Lambda_K = \begin{bmatrix} \Lambda_1 & 0 & \cdots & 0 \\ 0 & \Lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Lambda_1 \end{bmatrix} \in \mathbb{R}^{2K\times 2K}.$$

The identity (4) forms the basis for applying the method of exchangeable pairs; the proof of the theorem is concluded by using an appropriate Stein identity and bounding various moments involving $\mathbf{w}, \mathbf{w}'$.

The upper bound (3) does not appear to be optimal; cf. Section 5 of Jiang (1996) and Section 3 of Chatterjee (2008), where conditions for convergence depend on the ratios $\|Q_k - \breve{Q}_k\|^2 / \operatorname{Var}(\boldsymbol{\zeta}^\top Q_k \boldsymbol{\zeta})$ and $\operatorname{tr}(Q_k^4) / \operatorname{Var}(\boldsymbol{\zeta}^\top Q_k \boldsymbol{\zeta})^2$, respectively. However, it is likely that (3) can be improved by carefully examining the proof in Appendix B, if one is willing to accept a more complex

(and potentially less user-friendly) bound. Moreover, we argue presently that the bound (3) is already effective in a range of practical settings. Assume that in addition to the conditions of Theorem 2, the ζ_i are i.i.d. with excess kurtosis $\gamma_2 = \mathbb{E}(\zeta_i) - 3 \geq -2$. Also, let $\sigma_k^2 = \operatorname{Var}(\boldsymbol{\zeta}^\top Q_k \boldsymbol{\zeta})$. By Lemma S8 from the supplement, $\sigma_k^2 = 2\operatorname{tr}(Q_k^2) + \gamma_2 \operatorname{tr}(\breve{Q}_k^2) \geq (2+\gamma_2)\operatorname{tr}(Q_k^2)$. Hence, the upper bound in Theorem 2 implies that $\boldsymbol{\zeta}^\top Q_k \boldsymbol{\zeta}/\sigma_k$ is asymptotically $N(0,1)$, if

$$\frac{d^{1/2}\|Q_k\|^2}{\sigma_k^2} + \frac{d\|Q_k\|^3}{\sigma_k^3} \leq \frac{d^{1/2}\|Q_k\|^2}{(2+\gamma_2)\operatorname{tr}(Q_k^2)} + \left\{\frac{d^{2/3}\|Q_k\|^2}{(2+\gamma_2)\operatorname{tr}(Q_k^2)}\right\}^{3/2} \to 0.$$

We conclude that if (i) $\liminf \gamma_2 > -2$ and (ii) $\|Q_k\|^2/\operatorname{tr}(Q_k^2) = o(d^{-2/3})$, then $\boldsymbol{\zeta}^\top Q_k \boldsymbol{\zeta}/\sigma_k$ is asymptotically $N(0,1)$. Regarding (i), note that $\gamma_2 > -2$ for all distributions except the Rademacher distribution; furthermore, (ii) holds if, for instance, all of the eigenvalues of Q_k are contained in a compact subset of $(0,\infty)$.

On the other hand, there are some fairly natural applications where the conditions on Q_k discussed in the previous paragraph do not hold. In the next section, we will consider applications where the nonzero eigenvalues of Q_k may be identified with those of a sample covariance matrix (up to a scalar multiple). If the data are from a "spiked covariance" model [Johnstone (2001)] with strong spikes, where $d_0 \asymp 1$ of Q_k's eigenvalues are $\asymp d$ and the rest are $\asymp 1$, then

$$\frac{d^{1/2}\|Q_k\|^2}{\sigma_k^2} + \frac{d\|Q_k\|^3}{\sigma_k^3} \asymp \frac{d^{3/2}}{d^2 d_0^2 + (d-d_0)} + \frac{d^4}{d^3 d_0^{3/2} + (d-d_0)^{3/2}} \to \infty$$

and the upper bound in Theorem 2 may diverge. However, depending on the application, in settings like this it may be feasible to use principal components analysis as a pre-processing step, to "project out" the large eigenvalues of Q_k; Theorem 2 may then possibly be applied to the projected data.

4. Linear random-effects models. In this section, we apply the results from Section 3 to the variance components estimation problem in a linear random-effects model. We assume that

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ is an observed n-dimensional outcome vector, $X = (x_{ij})$ is an observed $n \times p$ predictor matrix with $x_{ij} \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown p-dimensional vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ is an unobserved error vector. We further assume that $\beta_1, \ldots, \beta_p, \varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with

$$\begin{aligned} &\mathbb{E}(\varepsilon_i) = 0 \quad \text{and} \quad \operatorname{Var}(\varepsilon_i) = \sigma_0^2, \qquad i = 1, \ldots, n, \\ &\mathbb{E}(\beta_j) = 0 \quad \text{and} \quad \operatorname{Var}(\beta_j) = \frac{\sigma_0^2 \eta_0^2}{p}, \qquad j = 1, \ldots, p. \end{aligned} \tag{6}$$

Here, we assume that the β_j are all independent. In Section 4.3, we investigate a more general model with dependent random-effects and give a corresponding concentration bound. Throughout, we also assume that X is independent of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\beta}$. Overall, (5)–(6) is a linear random-effects model with variance components parameters $\boldsymbol{\theta}_0 = (\sigma_0^2, \eta_0^2)$. In this parametrization, σ_0^2 is the residual variance and η_0^2 is a measure of the signal-to-noise ratio; this parametrization is standard [e.g., Hartley and Rao (1967)].

Let $\boldsymbol{\theta} = (\sigma^2, \eta^2)$ and define the Gaussian data log-likelihood,

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

Note that $\ell(\boldsymbol{\theta})$ is the log-likelihood for $\boldsymbol{\theta}$, if $\boldsymbol{\beta} \sim N\{0, (\eta_0^2\sigma_0^2/p)I\}$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma_0^2 I)$ are Gaussian. In this section, we study properties of the maximum likelihood estimator (MLE),

$$\hat{\boldsymbol{\theta}} = (\hat{\sigma}^2, \hat{\eta}^2) = \operatorname*{argmax}_{\sigma^2, \eta^2 \geq 0} \ell(\boldsymbol{\theta}), \tag{7}$$

in settings where $\boldsymbol{\varepsilon}$ and $\boldsymbol{\beta}$ are not necessarily Gaussian. [N.B. if $\ell(\boldsymbol{\theta})$ in (7) has multiple maximizers, then use any pre-determined rule to select $\hat{\boldsymbol{\theta}}$.]

The estimator $\hat{\boldsymbol{\theta}}$ has already been widely studied in the literature, even in settings where $\boldsymbol{\varepsilon}$ and $\boldsymbol{\beta}$ are not Gaussian [e.g., Richardson and Welsh (1994); Jiang (1996)]. In practice, $\hat{\boldsymbol{\theta}}$ and other closely related estimators, such as REML estimators, are probably the most commonly used variance components estimators for linear random-effects models [Harville (1977); Searle, Casella and McCulloch (1992); Demidenko (2013)]. Jiang's (1996) work is especially relevant for the results in this section. Jiang studied models with independent random-effects and derived general consistency and asymptotic normality results for $\hat{\boldsymbol{\theta}}$ that are valid in some of the settings considered here. However, asymptotic results tend to have more limited flexibility for use in certain applications. This has become more notable recently, with the widespread use of random-effects models in genomic and other applications, as discussed in Section 1.

In Sections 4.2–4.5 below, we present finite sample concentration and normal approximation bounds for $\hat{\boldsymbol{\theta}}$, which follow from Theorems 1 and 2, respectively. These bounds have not been optimized and some of the quantities in the bounds can be extremely large for given values of $\boldsymbol{\theta}_0$ and $p^{-1}XX^\top$ [e.g., $\kappa(\eta_0^2, \sigma_0^2, \Lambda)^{-1}$ and $\nu(\eta_0^2, \sigma_0^2, \Lambda)$, defined in (9) and (22) below]. However, as described in the text below, Propositions 1 and 4 still yield the "correct" asymptotic conclusions, similar to Jiang (1996), which ensure consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$, if

$p/n \to \rho \in (0, \infty)$ and the model parameters are bounded. Though it may be of interest to further optimize Propositions 1–4 (and it is almost certainly possible), our main emphasis is that the nonasymptotic approach taken here provides additional flexibility for deriving and understanding results in less standard settings. For instance, while Propositions 1 and 4 parallel existing results in Jiang (1996), Proposition 2 is a concentration bound for linear models with correlated random-effects and appears to be more novel [an application of Proposition 2 may be found in Dicker and Erdogdu (2016a)].

4.1. *Additional notation.* It is convenient to introduce some notation relating to the spectrum of X. Let $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of $p^{-1}XX^\top$ and suppose that $p^{-1}XX^\top = U\Lambda U^\top$ is the eigen-decomposition of $p^{-1}XX^\top$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and U is an $n \times n$ orthogonal matrix. Let $n_0 = \max\{i; \lambda_i > 0\}$ be the rank of $p^{-1}XX^\top$ and define the empirical variance of the eigenvalues of $p^{-1}XX^\top$,

$$\begin{aligned}\mathfrak{v}(\Lambda) &= \frac{1}{n}\sum_{i=1}^n \lambda_i^2 - \left(\frac{1}{n}\sum_{i=1}^n \lambda_i\right)^2 \\ &= \frac{1}{n}\operatorname{tr}\left\{\left(\frac{1}{p}XX^\top\right)^2\right\} - \left\{\frac{1}{n}\operatorname{tr}\left(\frac{1}{p}XX^\top\right)\right\}^2.\end{aligned}$$

4.2. *Concentration bound.* To derive a concentration bound for $\hat{\boldsymbol{\theta}}$ (Proposition 1 below), we follow standard steps in the analysis of variance components estimators [Hartley and Rao (1967)]. In particular, we introduce the profile likelihood and other related objects, which essentially reduce the bivariate optimization problem (7) to a univariate problem. Basic calculus implies that if $\eta^2 \geq 0$, then

$$\max_{\sigma^2, \eta^2 \geq 0} \ell(\sigma^2, \eta^2) = \max_{\eta^2 \geq 0} \ell_*(\eta^2),$$

where $\ell_*(\eta^2) = \ell\{\sigma_*^2(\eta^2), \eta^2\}$ is called the profile likelihood and

$$\sigma_*^2(\eta^2) = \frac{1}{n}\mathbf{y}^\top\left(\frac{\eta^2}{p}XX^\top + I\right)^{-1}\mathbf{y}.$$

It follows that $\hat{\boldsymbol{\theta}} = (\sigma_*^2(\hat{\eta}^2), \hat{\eta}^2)$, where

$$\hat{\eta}^2 = \operatorname*{argmax}_{\eta^2 \geq 0} \ell_*(\eta^2).$$

The proof of Proposition 1 hinges on comparing the profile likelihood $\ell_*(\eta^2)$ to its population version,

$$\ell_0(\eta^2) = -\frac{1}{2}\log\{\sigma_0^2(\eta^2)\} - \frac{1}{2n}\log\det\left(\frac{\eta^2}{p}XX^\top + I\right) - \frac{1}{2},$$

where we have replaced $\sigma_*^2(\eta^2)$ in $\ell_*^2(\eta^2)$ with its expectation,

$$\sigma_0^2(\eta^2) = \mathbb{E}\{\sigma_*^2(\eta^2)|X\} = \frac{\sigma_0^2}{n}\operatorname{tr}\left\{\left(\frac{\eta_0^2}{p}XX^\top + I\right)\left(\frac{\eta^2}{p}XX^\top + I\right)^{-1}\right\}$$
$$= \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\frac{\eta_0^2\lambda_i + 1}{\eta^2\lambda_i + 1}.$$

Observe that $\sigma_0^2(\eta_0^2) = \sigma_0^2$.

Overall, our strategy for proving Proposition 1 mirrors the classical parametric theory for consistency of maximum likelihood and M-estimators [e.g., Chapter 5 of van der Vaart (1998)], except that we employ Theorem 1 at several key steps. As in the standard analysis, two important facts underlying Proposition 1 are (i) η_0^2 is the unique maximizer of $\ell_0(\eta^2)$ and (ii) $\ell_*(\eta^2) \approx \ell_0(\eta^2)$, when n, p are large. Theorem 1 is used to make the approximation $\ell_*(\eta^2) \approx \ell_0(\eta^2)$ more precise. It should not be surprising that quadratic forms play an important role in the analysis, given the dependence of $\ell_*(\eta^2) = \ell\{\sigma_*^2(\eta^2), \eta^2\}$ on the quadratic form $\sigma_*^2(\eta^2)$. We emphasize that to prove Proposition 1, we use Theorem 1 with $K = 1$ and $\mathbf{u} = \eta^2$; the general version of Theorem 1 with matrix functions defined on $\mathbb{R}^K$ may be useful for studying random-effects model with K-groups of random-effects, for example, the general linear random-effects model considered in Jiang (1996). Proposition 1 is proved in Section S1 of the supplement.

PROPOSITION 1. *Assume that the linear random-effects model* (5)–(6) *holds and that* $\beta_1, \ldots, \beta_p$, $\varepsilon_1, \ldots, \varepsilon_n$ *are independent sub-Gaussian random variables satisfying*

$$\max\{\|p^{1/2}\beta_j\|_{\psi_2}, \|\varepsilon_i\|_{\psi_2}; i = 1, \ldots, n, j = 1, \ldots, p\} \le \gamma \tag{8}$$

for some $0 < \gamma < \infty$. *Finally, define*

$$\kappa(\sigma_0^2, \eta_0^2, \Lambda) = \frac{\sigma_0^4\eta_0^8\mathfrak{v}(\Lambda)^2}{(\sigma_0^2+1)^5(\eta_0^2+1)^{12}(\lambda_1+1)^{18}(\lambda_{n_0}^{-1}+1)^8\{\mathfrak{v}(\Lambda)+1\}^2}. \tag{9}$$

(a) *Suppose that* $n_0 = n$. *There is an absolute constant* $0 < C < \infty$ *such that*

$$\mathbb{P}\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > r|X\} \le C\exp\left\{-\frac{n}{C}\cdot\frac{\kappa(\sigma_0^2, \eta_0^2, \Lambda)}{\gamma^2(\gamma+1)^2}\cdot\frac{r^2}{(r+1)^2}\right\}$$

for every $r \ge 0$.

(b) *Suppose that* $n_0 < n$. *There is an absolute constant* $0 < C < \infty$ *such that*

$$\mathbb{P}\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > r|X\}$$
$$\le C\exp\left\{-\frac{n}{C}\cdot\frac{\kappa(\sigma_0^2, \eta_0^2, \Lambda)}{\gamma^2(\gamma+1)^2}\cdot\left(1-\frac{n_0}{n}\right)^4\left(\frac{n_0}{n}\right)^2\cdot\frac{r^2}{(r+1)^2}\right\}$$

for every $r \ge 0$.

For given values of σ_0^2, η_0^2 and Λ, the quantity $\kappa(\sigma_0^2, \eta_0^2, \Lambda)$ in Proposition 1 may be extremely small [for instance, if $\sigma_0^2 = \eta_0^2 = \lambda_1 = \mathfrak{v}(\Lambda) = 1$ and $\lambda_{n_0} = 1/3$, then $\kappa(\sigma_0^2, \eta_0^2, \Lambda) = 2^{-53}$; if $\sigma_0^2 = \eta_0^2 = \lambda_1 = \mathfrak{v}(\Lambda) = 3$ and $\lambda_{n_0} = 1$, then $\kappa(\sigma_0^2, \eta_0^2, \Lambda) = 3^{14} \cdot 2^{-82}$]. However, we have not attempted to optimize $\kappa(\sigma_0^2, \eta_0^2, \Lambda)$, and the bounds in the proposition can almost certainly be improved at the expense of some additional calculations and a more complex bound. Furthermore, despite the magnitude of $\kappa(\sigma_0^2, \eta_0^2, \Lambda)$, the proposition yields very sensible asymptotic conclusions. Indeed, the key property of $\kappa(\sigma_0^2, \eta_0^2, \Lambda)$ is that if $\mathcal{U} \subseteq (0, \infty)$ is compact, then

$$0 < \inf\{\kappa(\sigma_0^2, \eta_0^2, \Lambda); \sigma_0^2, \eta_0^2, \lambda_1, \ldots, \lambda_n, \mathfrak{v}(\Lambda) \in \mathcal{U}\}. \tag{10}$$

An immediate consequence is that if $\sigma_0^2, \eta_0^2, \lambda_1, \ldots, \lambda_n, \mathfrak{v}(\Lambda)$ are contained in a compact subset of $(0, \infty)$, then Proposition 1 implies that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ at rate $n^{1/2}$ [at least when $n = n_0$; if $n_0 < n$, then part (b) of the proposition requires the additional condition that n_0/n stays away from 1—this is discussed further below].

The bounds in Proposition 1 are tighter [i.e., $\kappa(\sigma_0^2, \eta_0^2, \Lambda)$ is larger] when the eigenvalue variance $\mathfrak{v}(\Lambda)$ is large. This is related to identifiability: σ_0^2 and η_0^2 are not identifiable when $\mathfrak{v}(\Lambda) = 0$, and it is easier to distinguish between them when $\mathfrak{v}(\Lambda)$ is large.

The cases where $n_0 = n$ and $n_0 < n$ are considered separately in Proposition 1 because the large-η^2 asymptotic behavior of $\sigma_0^2(\eta^2) = \mathbb{E}\{\sigma_*^2(\eta^2)|X\}$ differs in these two settings. In particular, assuming that σ_0^2, n_0, n and $\lambda_1, \ldots, \lambda_n$ are fixed, if $n_0 = n$, then as $\eta^2 \to \infty$.

$$\sigma_0^2(\eta^2) = \frac{\sigma_0^2}{n} \sum_{i=1}^{n} \frac{\eta_0^2 \lambda_i + 1}{\eta^2 \lambda_i + 1} \asymp \frac{1}{\eta^2} \to 0,$$

as $\eta^2 \to \infty$; on the other hand, if $n_0 < n$, then

$$\sigma_0^2(\eta^2) = \frac{\sigma_0^2}{n} \sum_{i=1}^{n_0} \frac{\eta_0^2 \lambda_i + 1}{\eta^2 \lambda_i + 1} + \sigma_0^2 \left(1 - \frac{n_0}{n}\right) \asymp 1,$$

as $\eta^2 \to \infty$.

Note that Proposition 1(a) actually makes no explicit reference to p, or to the relative convergence rates of p and n. However, there are implicit conditions on p. For instance, since $n_0 = n$ in part (a), we must have $n \leq p$. Additionally, in order ensure that $\lambda_1, \ldots, \lambda_n$ are contained in a compact subset $\mathcal{U} \subseteq (0, \infty)$, so that (10) holds, it may be natural to enforce other conditions on p, for example, $p/n \to \rho \in (1, \infty)$.

Part (b) of Proposition 1 applies to settings where $p < n$. Note that the upper bound in part (b) contains an additional term $(1 - n_0/n)^4(n_0/n)^2$, as compared to Proposition 1(a). Thus, assuming that $\sigma_0^2, \eta_0^2, \lambda_1, \ldots, \lambda_n, \mathfrak{v}(\Lambda)$ are contained in a compact subset of $(0, \infty)$, we conclude that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ at rate $n^{1/2}$, if

$$\liminf\left(1 - \frac{n_0}{n}\right)\frac{n_0}{n} > 0. \tag{11}$$

Observe that (11) implies $p \to \infty$. Hence, we need $p \to \infty$ in order to ensure that $\hat{\boldsymbol{\theta}}$ is consistent. This is reasonable because information about $\eta_0^2 = p\mathbb{E}(\beta_j^2)/\sigma_0^2$ is accumulated through $\beta_1, \ldots, \beta_p$. The condition (11) also implies that if X is full rank, then we must have $p/n \to \rho < 1$ in order to ensure consistency. This condition seems less natural and can likely be relaxed with a more careful analysis; similar challenges arise frequently in random matrix theory when $p/n \to 1$ [e.g., Bai, Miao and Yao (2003)].

4.3. *A more general concentration bound.* In this section, we investigate the performance of $\hat{\boldsymbol{\theta}}$ in models where the random-effects might be dependent. Suppose that $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^\top \in \mathbb{R}^p$ is a random vector that is independent of $\boldsymbol{\varepsilon}$, X and let

$$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}. \tag{12}$$

We do not assume that $\tilde{\boldsymbol{\beta}}$ has independent components or that each of the components has the same variance. We define the variance components estimator based on the data $(\tilde{\mathbf{y}}, X)$,

$$\tilde{\boldsymbol{\theta}} = (\tilde{\sigma}^2, \tilde{\eta}^2) = \operatorname*{argmax}_{\sigma^2, \eta^2 \geq 0} \tilde{\ell}(\boldsymbol{\theta}), \tag{13}$$

where

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

The next proposition is a concentration bound for $\tilde{\boldsymbol{\theta}}$, which implies that the estimator may still perform reliably, if there is a good coupling for $\tilde{\boldsymbol{\beta}}$.

PROPOSITION 2. *Suppose that $\tilde{\mathbf{y}}, \tilde{\boldsymbol{\theta}}$ satisfy (12)–(13). Suppose further that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is a random vector with independent components, which is independent of $\boldsymbol{\varepsilon}$, X (but may be correlated with $\tilde{\boldsymbol{\beta}}$), such that the independent random-effects model (5)–(6) and (8) hold. Let $\kappa(\sigma_0^2, \eta_0^2, \Lambda)$ be as in (9).*

(a) *Suppose that* $n_0 = n$. *There is an absolute constant* $0 < C < \infty$ *such that*

$$\begin{aligned}&\mathbb{P}\{\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > r|X\}\\ &\quad\le C\exp\left\{-\frac{n}{C}\cdot\frac{\kappa(\sigma_0^2,\eta_0^2,\Lambda)}{\gamma^2(\gamma+1)^2}\cdot\frac{r^2}{(r+1)^2}\right\}\\ &\qquad+4\mathbb{P}\left\{\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > \frac{1}{C}\cdot\frac{\kappa(\sigma_0^2,\eta_0^2,\Lambda)}{(\gamma+1)^4}\cdot\frac{n}{p+n}\cdot\frac{r}{r+1}\Big|X\right\}.\end{aligned} \tag{14}$$

for every $r \ge 0$.

(b) *Suppose that* $n_0 < n$. *There is an absolute constant* $0 < C < \infty$ *such that*

$$\begin{aligned}&\mathbb{P}\{\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > r|X\}\\ &\quad\le C\exp\left\{-\frac{n}{C}\cdot\frac{\kappa(\sigma_0^2,\eta_0^2,\Lambda)}{\gamma^2(\gamma+1)^2}\cdot\left(1-\frac{n_0}{n}\right)^4\left(\frac{n_0}{n}\right)^2\cdot\frac{r^2}{(r+1)^2}\right\}\\ &\qquad+4\mathbb{P}\left\{\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > \frac{1}{C}\cdot\frac{\kappa(\sigma_0^2,\eta_0^2,\Lambda)}{(\gamma+1)^4}\cdot\frac{n}{p+n}\cdot\frac{r}{r+1}\Big|X\right\}\end{aligned} \tag{15}$$

for every $r \ge 0$.

The proof of Proposition 2 is similar to that of Proposition 1 and may be found in Section S2 of the supplement. Observe that the first term in each upper bound (14)–(15) is the exact same as in Proposition 1. The second term in each of the bounds is new; this term is small, if $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ is typically small. In other words, in a random-effects model where (6) does not hold, the Gaussian maximum likelihood estimator $\tilde{\boldsymbol{\theta}}$ may be a reliable estimator for the variance components parameter $\boldsymbol{\theta}_0 = (\sigma_0^2, \eta_0^2)$ from a corresponding random-effects model (5)–(6), if $\tilde{\boldsymbol{\beta}} \approx \boldsymbol{\beta}$.

Proposition 2 is useful for applications involving misspecified random-effects models. In the next section, we show how it can be used to recover some of Jiang et al. (2016) consistency results for sparse random-effects models in genome wide association studies. An application of Proposition 2 involving fixed-effects models can be found in Dicker and Erdogdu (2016a). Despite its evident usefulness in these applications, there are some relatively subtle issues related to the interpretation of Proposition 2. Indeed, assume that a model (12) is specified for some particular application, with potentially correlated or heteroscedastic random-effects. In such a setting, it may be natural to assume that the estimator $\tilde{\boldsymbol{\theta}}$ targets the residual variance and signal-to-noise ratio parameters from the model (12). However, Proposition 2 bounds the distance between $\tilde{\boldsymbol{\theta}}$ and the corresponding parameters $\boldsymbol{\theta}_0 = (\sigma_0^2, \eta_0^2)$ from the independent random-effects model (5)–(6). It is not clear a priori that $\boldsymbol{\theta}_0$ should be interpreted in the same way as the parameters from the original model (12); thus, one should take care when interpreting $\tilde{\boldsymbol{\theta}}$ in practice. On the other hand, if $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \approx 0$ and the second term in the upper bounds in Proposition 2 is small, then it often follows that $\mathbb{E}(\|\tilde{\boldsymbol{\beta}}\|^2) \approx \mathbb{E}(\|\boldsymbol{\beta}\|^2) = \sigma_0^2\eta_0^2$. This

gives a natural interpretation of $\boldsymbol{\theta}_0$ in the context of the original model (12), that is, $\eta_0^2 \approx \mathbb{E}(\|\tilde{\boldsymbol{\beta}}\|^2)/\sigma_0^2$ and $\sigma_0^2 = \mathrm{Var}(\varepsilon_i)$.

4.4. *Example*: *Sparse random-effects models for GWAS data*. Linear random-effects models like (5)–(6) (and, more generally, linear mixed models) have been used for applications in genetics since at least the 1950s [Henderson (1950)]. More recently, variants of (5)–(6) have become an important tool for analyzing data from genome wide association studies (GWAS) [Yang et al. (2010, 2014); Golan and Rosset (2011); Speed et al. (2012); Zaitlen and Kraft (2012); Jiang et al. (2016); de Los Campos, Sorensen and Gianola (2015)]. In GWAS datasets, the outcome $y_i \in \mathbb{R}$ often represents a measured quantitative trait from an individual (e.g., their height) and the corresponding predictors $x_{ij} \in \{0, 1, 2\}$, $1 \le j \le p$, indicate the genotype (minor allele frequency) for each single nucleotide polymorphism (SNP) assayed in the study. In large GWAS datasets, one may have n in the 10,000s–100,000s and p in the 100,000s–1,000,000s. An important objective in many GWAS analyses involving random-effects models is estimating the heritability parameter, which represents the overall proportion of variability in y_i that is explained by the genotype $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Heritability can be naturally formulated in terms of variance components parameters, as in (16) below.

Jiang et al. (2016) studied a sparse random-effects (SRE) model with random predictors for applications involving GWAS data; sparse models are popular in GWAS because it is often assumed that many of the SNPs have no effect on the outcome. While Jiang et al.'s model is highly idealized, their work appears to be one of the first in-depth theoretical analyses of mixed-models with a specific emphasis on GWAS. The SRE model studied in Jiang et al. (2016) satisfies (5) and:

(SRE1) For $1 \le i \le n$, $1 \le j \le p$, the predictors x_{ij} are i.i.d. random variables that have been standardized to have mean 0 and variance 1.

(SRE2) $\varepsilon_1, \ldots, \varepsilon_n \sim N(0, \sigma_0^2)$ are i.i.d.

(SRE3) There is a fixed, unknown subset $\mathcal{C} \subseteq \{1, \ldots, p\}$ with p_0 elements such that

$$\beta_j \sim N(0, \eta_g^2 \sigma_0^2/p) \qquad \text{if } j \in \mathcal{C},$$
$$\beta_j = 0 \qquad \text{if } j \in \{1, \ldots, p\} \setminus \mathcal{C},$$

where $\eta_g^2 > 0$ and the nonzero β_j are all independent.

(SRE4) X, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\beta}$ are all independent of each other.

Under this model, the heritability parameter is

$$h_0^2 = \frac{\mathrm{Var}(\mathbf{x}_i^\top \boldsymbol{\beta})}{\mathrm{Var}(y_i)} = \frac{(p_0/p)\eta_g^2}{(p_0/p)\eta_g^2 + 1}. \tag{16}$$

In the remainder of this section, we focus on the estimators $\hat{\boldsymbol{\theta}} = (\hat{\sigma}^2, \hat{\eta}^2)$ [defined in (7)] and $\hat{h}^2 = \hat{\eta}^2/(\hat{\eta}^2 + 1)$ under the SRE model.

Jiang et al. (2016) used asymptotic arguments to show that under the SRE model, if $p/n \to \rho \in [1, \infty)$ and $p_0/p \to \rho_0 \in (0, 1]$, then $\hat{h}^2 = \hat{\eta}^2/(\hat{\eta}^2+1) \to h_0^2$ in probability (they also derive asymptotic normality results for $\hat{h}^2$ under the SRE model). Here, we provide an alternative analysis and show that $\hat{h}^2$ is consistent using Proposition 2. It is important to note that Jiang et al.'s (2016) results average over the distribution of X, in addition to the distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$; in other words, they are not conditional on X and they take advantage of assumption (SRE1). In our analysis, we leverage (SRE1) as well, using an exchangeability argument before applying Proposition 2. We believe this argument provides some new insight into the relationship between the random-effects β_j and the predictors x_{ij}; this is discussed further after Proposition 3 below.

Before proceeding, observe that Proposition 1 cannot be applied to $\hat{\boldsymbol{\theta}}$ under the SRE model because $\operatorname{Var}(\beta_j) = 0 \neq \eta_g^2\sigma_0^2/p = \operatorname{Var}(\beta_{j'})$ for $j \notin \mathcal{C}$, $j' \in \mathcal{C}$, and $\mathcal{C}$ is unknown. Moreover, Proposition 2 cannot be applied directly (with $\boldsymbol{\beta}$ in place of $\tilde{\boldsymbol{\beta}}$), because the coordinates of $\boldsymbol{\beta}$ in the SRE model are not exchangeable and it is not possible to find a good coupling, as required in the proposition. However, we can *induce* exchangeability in $\boldsymbol{\beta}$ by using the exchangeability of x_{ij}; then we can apply Proposition 2.

To begin, assume that the SRE model holds and let Π be an independent uniformly distributed $p \times p$ permutation matrix. Next, define $\tilde{\boldsymbol{\beta}} = \Pi\boldsymbol{\beta}$, and let $\tilde{\mathbf{y}}$ be given in (12). The key observation is that (SRE1) implies that the MLEs $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y}, X)$ and $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\tilde{\mathbf{y}}, X)$, defined in (7) and (13), respectively, have the same distribution. Indeed, for any Borel set $B \subseteq \mathbb{R}^2$,

$$\begin{aligned}\mathbb{P}\{\hat{\boldsymbol{\theta}}(\mathbf{y}, X) \in B\} &= \mathbb{P}\{\hat{\boldsymbol{\theta}}(X\Pi^\top\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, X) \in B\} \\ &= \mathbb{P}\{\hat{\boldsymbol{\theta}}(X\Pi^\top\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, X\Pi^\top) \in B\} \\ &= \mathbb{P}\{\hat{\boldsymbol{\theta}}(X\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, X) \in B\} = \mathbb{P}\{\tilde{\boldsymbol{\theta}}(\tilde{\mathbf{y}}, X) \in B\},\end{aligned} \tag{17}$$

where the second equality holds because

$$\hat{\boldsymbol{\theta}}(\mathbf{y}, X) = \hat{\boldsymbol{\theta}}(\mathbf{y}, XU) \tag{18}$$

for any $p \times p$ orthogonal matrix U and the third equality follows from (SRE1) (in particular, the fact that the x_{ij} are exchangeable). Thus, in order to study distributional properties of $\hat{\boldsymbol{\theta}}$ under the SRE model (unconditional on X), it suffices to study $\tilde{\boldsymbol{\theta}}$. Moreover, $\tilde{\boldsymbol{\theta}}$ and the model (12) are relatively nice because the coordinates of $\tilde{\boldsymbol{\beta}}$ are exchangeable, unlike the coordinates of $\boldsymbol{\beta}$. In particular, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^\top \in \mathbb{R}^p$ is formed by selecting p_0 coordinates from $\{1, \ldots, p\}$ uniformly at random and setting them to be $N(0, \eta_g^2\sigma_0^2/p)$ random variables, then setting the remaining coordinates equal to 0.

In order to apply Proposition 2 with $\tilde{\boldsymbol{\beta}}$, we need to construct a coupling; that is, we need to find a vector $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top \in \mathbb{R}^p$ whose coordinates have mean 0, constant variance, and are independent, such that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$ is small ($\boldsymbol{\beta}^*$ will be

used in place of $\boldsymbol{\beta}$ in Proposition 2). Begin by taking $z_1, \ldots, z_p \sim N(0,1)$ to be independent and let $I_1, \ldots, I_p$ be independent Bernoulli random variables with $\mathbb{P}(I_j = 1) = p_0/p$. Now let $\mathcal{S} = \{j; I_j = 1\}$. If $p_{\mathcal{S}} = |\mathcal{S}| > p_0$, define $\tilde{\mathcal{S}}$ by selecting p_0 elements uniformly at random from $\mathcal{S}$; if $p_{\mathcal{S}} = p_0$, define $\tilde{\mathcal{S}} = \mathcal{S}$; and if $p_{\mathcal{S}} < p_0$, define $\tilde{\mathcal{S}}$ by selecting $p_0 - p_{\mathcal{S}}$ elements uniformly at random from $\{1, \ldots, p\} \setminus \mathcal{S}$ and adding them to $\mathcal{S}$. Then, in any event, $\tilde{\mathcal{S}}$ is uniformly distributed on subsets of $\{1, \ldots, p\}$ with size p_0. Consequently, we may assume without loss of generality that

$$\tilde{\beta}_j = \frac{\eta_g \sigma_0}{\sqrt{p}} z_j \qquad \text{if } j \in \tilde{\mathcal{S}},$$

$$\tilde{\beta}_j = 0 \qquad \text{if } j \in \{1, \ldots, p\} \setminus \tilde{\mathcal{S}}.$$

Next, define $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top = (\eta_g \sigma_0/\sqrt{p})(z_1 I_1, \ldots, z_p I_p)^\top$, so that

$$\beta_j^* = \frac{\eta_g \sigma_0}{\sqrt{p}} z_j \qquad \text{if } j \in \mathcal{S},$$

$$\beta_j^* = 0 \qquad \text{if } j \in \{1, \ldots, p\} \setminus \mathcal{S}.$$

Clearly, the coordinates of $\boldsymbol{\beta}^*$ are independent, $\mathbb{E}(\beta_j^*) = 0$, and $\mathrm{Var}(\beta_j^*) = \eta_g^2 \sigma_0^2 p_0/p^2$. Thus, we can now apply Proposition 2, with $\boldsymbol{\beta}^*$ in place of $\boldsymbol{\beta}$, to obtain bounds on $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$, where $\boldsymbol{\theta}^* = (\sigma_0^2, \eta_g^2 p_0/p)$.

The last step, in order to determine the effectiveness of the bounds given by Proposition 2, is to bound the difference $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$; this will allow us to control the second term in the upper bounds from Proposition 2. The difference $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$ can be bounded precisely by combining the fact that $\tilde{\beta}_j \neq \beta_j^*$ if and only if $j \in (\mathcal{S} \setminus \tilde{\mathcal{S}}) \cup (\tilde{\mathcal{S}} \setminus \mathcal{S})$ with some elementary tail bounds. Piecing everything together leads to finite-sample bounds on $\mathbb{P}\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > r\}$ and, consequently, $\mathbb{P}\{|\hat{\sigma}^2 - \sigma_0^2| > r\}$, $\mathbb{P}\{|\hat{\eta}^2 - (p_0/p)\eta_g^2| > r\}$, and $\mathbb{P}\{|\hat{h}^2 - h_0^2| > r\}$. However, rather than going through these details here, we emphasize the asymptotic implications, which are given in the following proposition. [The only additional work required to convert the finite sample bounds outlined above into an asymptotic result is to get probabilistic bounds the spectrum of $p^{-1}XX^\top$; this may be achieved using standard results from random matrix theory, for example, Chapter 3 of Bai and Silverstein (2010).]

PROPOSITION 3. *Assume that* (5) *and* (SRE1)–(SRE4) *hold, along with the following conditions*:

(i) $p_0/p \to \rho_0 \in (0, 1]$ *and* $p/n \to \rho \in (0, \infty) \setminus \{1\}$.
(ii) *There exists a compact set* $D \subseteq (0, \infty)$ *such that* $\sigma_0^2, \eta_g^2 \in D$.
(iii) *There is a common probability distribution* F, *with mean zero and finite fourth moment, such that* $x_{ij} \sim F$ *for all* $1 \le i \le n, 1 \le j \le p$.

Then $|\hat{\sigma}^2 - \sigma_0^2| \to 0$, $|\hat{\eta}^2 - (p_0/p)\eta_g^2| \to 0$, *and* $|\hat{h}^2 - h_0^2| \to 0$ *in probability.*

Proposition 3 is very similar to Theorem 3.1 from Jiang et al. (2016). However, beyond recapitulating an existing result, we believe that the approach outlined in this section provides additional insight into random-effects models—especially, into the role of the predictors X and how they interface with the random-effects $\boldsymbol{\beta}$.

Proposition 1, for the random-effects model (5)–(6), requires relatively weak conditions on the matrix of predictors X. In particular, if $p/n \to \rho \in (0, \infty) \setminus \{1\}$, then the upper bounds in Proposition 1 typically vanish when $\mathfrak{v}(\Lambda), \lambda_1, \lambda_n \asymp 1$. These conditions allow for relatively strong (though not arbitrary) correlation between the predictors x_{ij}. On the other hand, GWAS researchers have noted that correlation between the predictors can lead to bias in estimates of the heritability h_0^2 [Speed et al. (2012); de Los Campos, Sorensen and Gianola (2015)]. In many models, this bias arises from the combined effects of correlation between the predictors x_{ij} and a sparse random-effects model with nonexchangeable β_j, that is, a SRE model without the assumption (SRE1). Conversely, our results suggest that $\hat{h}^2$ may be a reliable estimator for the heritability parameter, as long as at least one of X or $\boldsymbol{\beta}$ has exchangeable entries. Indeed, this follows by combining Propositions 1–2 with the exchangeability argument outlined above in (17)–(18).

Exchangeability is a strong condition for X or $\boldsymbol{\beta}$, in order to ensure consistency of $\hat{h}^2$. Our approach in this paper provides some guidance for rigorously investigating the necessity of this condition and how to relax it. The two key facts underlying Proposition 3 for the SRE model are the distributional identity

$$X \overset{\mathscr{D}}{=} X\Pi^\top \tag{19}$$

and the invariance property for $\hat{\boldsymbol{\theta}}$ given in (18). It might be of interest to investigate the extent to which the identity (19) could be relaxed to an approximation, without significantly impacting the performance of $\hat{h}^2$. Additionally, for an SRE model satisfying (5) and (SRE2)–(SRE4) (without the condition that x_{ij} are i.i.d.), it might be possible to reformulate some of the bounds in this paper in terms of the empirical variance of the eigenvalues of *submatrices* of $n^{-1}X^\top X$, rather than $\mathfrak{v}(\Lambda)$ (the empirical variance of the eigenvalues of $p^{-1}XX^\top$). This approach might lead to more effective bounds for SRE models, conditional on the predictors X.

More broadly, the theoretical approach outlined in this paper may suggest other relevant models and effective methods for analyzing GWAS data. For instance, assuming that $\mathbb{E}(x_{ij}) = 0$ and $\mathrm{Var}(x_{ij}) = 1$, one might investigate the class of models (12) for which there is a coupling satisfying (5)–(6) and

$$\left|\mathrm{Var}(\mathbf{x}_i^\top \tilde{\boldsymbol{\beta}}) - \mathrm{Var}(\mathbf{x}_i^\top \boldsymbol{\beta})\right| = \left|\mathbb{E}\{(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top (\mathbf{x}_i \mathbf{x}_i^\top)(\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}})\}\right| \approx 0.$$

Proposition 2 indicates that within this class of models, $\hat{h}^2 = \hat{\eta}^2/(\hat{\eta}^2 + 1)$ may be a reasonable estimate of the heritability parameter

$$h_0^2 = \frac{\mathrm{Var}(\mathbf{x}_i^\top \tilde{\boldsymbol{\beta}})}{\mathrm{Var}(\tilde{y}_i)} = \frac{\mathrm{Var}(\mathbf{x}_i^\top \tilde{\boldsymbol{\beta}})}{\mathrm{Var}(\mathbf{x}_i^\top \tilde{\boldsymbol{\beta}}) + \mathrm{Var}(\varepsilon_i)} \approx \frac{\mathrm{Var}(\mathbf{x}_i^\top \boldsymbol{\beta})}{\mathrm{Var}(\mathbf{x}_i^\top \boldsymbol{\beta}) + \mathrm{Var}(\varepsilon_i)} = \frac{\eta_0^2}{\eta_0^2 + 1}.$$

Thus, it becomes a substantive question of whether or not this class of models can accurately describe interesting aspects of GWAS.

It is worth re-emphasizing that the analysis in this section is primarily unconditional, averaging over the distribution of X and $\boldsymbol{\beta}$. The unconditional approach is convenient mathematically, because we can leverage distributional assumptions on X and $\boldsymbol{\beta}$, such as exchangeability. On the other hand, a conditional analysis (more similar to what is found elsewhere in Section 4) often seems more natural in applications like GWAS. Indeed, the predictors X are observed, so it makes sense to condition on them; additionally, each parameter value β_j corresponds to the effect of a presumably well-defined predictor (e.g., SNP), so even averaging over the β_j may seem arguable. If one takes the conditional perspective, then it is less clear how to interpret exchangeability assumptions on X, $\boldsymbol{\beta}$ and, more generally, the results in this section. One approach to resolving these issues, which could be a topic for future research, involves converting the unconditional results in this section into conditional results, by replacing assumptions on the distribution of $(X, \boldsymbol{\beta})$ (e.g., exchangeability) with corresponding assumptions on their actual value.

4.5. *Normal approximation.* In this section, we shift our attention back to the independent random-effects model (5)–(6) and give a normal approximation result for $\hat{\boldsymbol{\theta}}$ (Proposition 4 below). One consequence of Proposition 4 is that under conditions similar to those described after Proposition 1, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal, when $p/n \to \rho \in [0, \infty)$. As with consistency (discussed in Section 4.2), asymptotic normality of $\hat{\boldsymbol{\theta}}$ has been studied previously in similar settings [Jiang (1996)]. However, the main significance of Theorem 4 is its flexible finite-sample nature, which makes it an easy-to-use tool for applications.

To derive Theorem 4, we again follow the standard strategy for parametric M-estimators. First, we introduce the score function:

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

Then $S(\hat{\boldsymbol{\theta}}) = 0$, provided $\hat{\eta}^2 > 0$. The main idea of the proof is to Taylor expand the score function about $\boldsymbol{\theta}_0$ so that

$$0 = S(\hat{\boldsymbol{\theta}}) = S(\boldsymbol{\theta}_0) + J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{r}, \tag{20}$$

where $J(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta})$ and $\mathbf{r}$ is a remainder term. Theorem 4 follows by solving for $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ above, then using three key intermediate results: (i) $S(\boldsymbol{\theta}_0)$ is approximately

normal, (ii) $J(\boldsymbol{\theta}_0) \approx J_0(\boldsymbol{\theta}_0)$, where

$$J_0(\boldsymbol{\theta}) = \mathbb{E}\{J(\boldsymbol{\theta})\} = \mathbb{E}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta})\right\}, \tag{21}$$

and (iii) the remainder term $\mathbf{r}$ is small. Approximate normality of $S(\boldsymbol{\theta}_0)$ follows from Theorem 2 in this paper. The approximation $J(\boldsymbol{\theta}_0) \approx J_0(\boldsymbol{\theta}_0)$ and the fact that $\mathbf{r}$ is small follow from concentration properties of quadratic forms.

PROPOSITION 4. *Assume that the linear random-effects model* (5)–(6) *holds and that* $\beta_1, \ldots, \beta_p$, $\varepsilon_1, \ldots, \varepsilon_n$ *are independent random variables satisfying* (8). *Define*

$$\nu(\sigma_0^2, \eta_0^2, \Lambda) = \frac{(\sigma_0^2+1)^9(\eta_0^2+1)^{16}(\lambda_1+1)^{24}}{\sigma_0^6\eta_0^2} \cdot \frac{\{\mathfrak{v}(\Lambda)+1\}^3}{\mathfrak{v}(\Lambda)^3}, \tag{22}$$

let $f \in \mathcal{C}_b^3(\mathbb{R}^2)$, *and let* $\mathbf{z}_2 \sim N(0, I)$ *be a two-dimensional standard normal random vector. Finally, let* $J_0(\boldsymbol{\theta}_0)$ *be as in* (21), *define* $\mathcal{I}(\boldsymbol{\theta}_0) = \operatorname{Var}\{S(\boldsymbol{\theta}_0)|X\}$, *and define*

$$\Psi = J_0(\boldsymbol{\theta}_0)^{-1}\mathcal{I}(\boldsymbol{\theta}_0)J_0(\boldsymbol{\theta}_0)^{-1}. \tag{23}$$

There is an absolute constant $0 < C < \infty$ *such that*

$$\begin{aligned} &|\mathbb{E}[f\{\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}|X] - \mathbb{E}\{f(\Psi^{1/2}\mathbf{z}_2)|X\}| \\ &\quad \le C(\gamma+1)^8 \nu(\sigma_0^2, \eta_0^2\Lambda)\left\{\prod_{k=1}^{3}(1+|f|_k)\right\} \cdot \frac{p+n}{n} \cdot \frac{\log(n)^2}{n^{1/2}} \\ &\qquad + 2|f|_0 \mathbb{P}\left\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > \frac{\sigma_0^2 \log(n)}{2\sqrt{n}} \middle| X\right\}. \end{aligned} \tag{24}$$

A detailed proof of Proposition 4 may be found in Section S3 of the supplement. The quantity $\nu(\sigma_0^2, \eta_0^2, \Lambda)$ in (24) is potentially extremely large, and plays a role similar to $\kappa(\sigma_0^2, \eta_0^2, \Lambda)$ in Propositions 1–2. As with the previous propositions, despite the potential magnitude of $\nu(\sigma_0^2, \eta_0^2, \Lambda)$, the asymptotic implications of Proposition 4 are very reasonable. Indeed, assume that the conditions of the proposition hold. If, additionally, $\sigma_0^2, \eta_0^2, \lambda_1, \ldots, \lambda_n, \mathfrak{v}(\Lambda)$ are contained in a compact subset of $(0, \infty)$ and $p/n \to \rho \in [0, \infty)$, then it is clear that the first term on the right-hand side of (24) converges to 0. Moreover, Theorem 1 implies that the second term on the right-hand side of (24) converges to 0, as long as we have the additional condition (11) when $n_0 < n$. Thus, under the specified conditions,

$$|\mathbb{E}[f\{\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}|X] - \mathbb{E}\{f(\Psi^{1/2}\mathbf{z})|X\}| \to 0 \tag{25}$$

for all $f \in \mathcal{C}_b^3(\mathbb{R}^2)$. This is an asymptotic normality result for $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. One apparent limitation of (25) is that it only applies for $f \in \mathcal{C}_b^3(\mathbb{R}^2)$. However, standard arguments [e.g., Section 3 of Reinert and Röllin (2009)] imply

that (25) is valid for broader classes of nonsmooth functions f, including indicator functions for measurable convex subsets of $\mathbb{R}^2$; thus, we may conclude that $\sqrt{n}\Psi^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightsquigarrow N(0, I)$ in distribution, where Ψ is defined in (23). We note additionally that if $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are Gaussian, then $\Psi = \mathcal{I}(\boldsymbol{\theta}_0)^{-1} = \mathcal{I}_N(\boldsymbol{\theta}_0)^{-1}$, where $\mathcal{I}_N(\boldsymbol{\theta}_0) = (\iota_{ij}(\boldsymbol{\theta}_0))_{i,j=1,2}$ is the Gaussian Fisher information matrix for $\boldsymbol{\theta}_0$ and

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

$k, l = 1, 2$. Moreover, standard likelihood theory [e.g., Chapter 6 of Lehmann and Casella (1998)] implies that $\hat{\boldsymbol{\theta}}$ is asymptotically efficient in the Gaussian random-effects model.

5. Discussion. We have presented new uniform concentration and normal approximation bounds for quadratic forms, and described some applications to variance components estimation in linear random-effects models. We expect that the general results for quadratic forms, found in Section 3, will be useful in a range of other applications, such as variance components estimation in nonstandard random- and fixed-effects linear models, which arise in genomics and other applications [Jiang et al. (2016); Dicker and Erdogdu (2016a)]; hypothesis testing for variance components parameters in high-dimensional models; and other hypothesis testing problems, where the test statistics involve quadratic forms in many random variables. As discussed in Sections 3.2 and 4, many of the bounds in the paper can be improved, at the expense of introducing some additional complexity into the results. Furthermore, all of our results require sub-Gaussian random variables. It may be of interest to sharpen the results in the paper and extend them to allow for heavier-tailed random variables with sufficiently many moments.

APPENDIX A: PROOF OF THEOREM 1

The proof begins with a chaining construction. Fix a positive integer M and define a regular grid on $[0, R]^K$ with $(2^M + 1)^K$ points, $\mathcal{U}_M^K = \mathcal{U}_M \times \cdots \times \mathcal{U}_M$, where $\mathcal{U}_M = \{i2^{-M}R\}_{i=0}^{2^M} \subseteq [0, R] \subseteq \mathbb{R}$. For each $\mathbf{u} = (u_1, \ldots, u_K)^\top \in [0, R]^K$ and $j = 1, \ldots, M$ define $\dot{\mathbf{u}}_j = (\dot{u}_{1j}, \ldots, \dot{u}_{Kj})^\top$ where $\dot{u}_{ij}$ is the smallest point in $\mathcal{U}_M$ that is at least as large as u_i; additionally, define $\dot{u}_0 = 0$.

Next, consider the decomposition $\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta} - \mathbb{E}\{\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta}\} = \Delta_1(\mathbf{u}) + \Delta_2(\mathbf{u}) + \Delta_3(\mathbf{u})$, where

$$\Delta_1(\mathbf{u}) = \boldsymbol{\zeta}^\top\{Q(\mathbf{u}) - Q(\dot{\mathbf{u}}_M)\}\boldsymbol{\zeta} - \mathbb{E}[\boldsymbol{\zeta}^\top\{Q(\mathbf{u}) - Q(\dot{\mathbf{u}}_M)\}\boldsymbol{\zeta}],$$

$$\Delta_2(\mathbf{u}) = \boldsymbol{\zeta}^\top Q(\dot{\mathbf{u}}_0)\boldsymbol{\zeta} - \mathbb{E}\{\boldsymbol{\zeta}^\top Q(\dot{\mathbf{u}}_0)\boldsymbol{\zeta}\},$$

$$\Delta_3(\mathbf{u}) = \sum_{j=1}^{M}\boldsymbol{\zeta}^\top\{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta} - \sum_{j=1}^{M}\mathbb{E}[\boldsymbol{\zeta}^\top\{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta}].$$

Let $r_1, r_2, r_3 > 0$ satisfy $r_1 + r_2 + r_3 = r$. Then

$$\begin{aligned}(26)\qquad &\mathbb{P}\Big[\sup_{\mathbf{u}\in[0,R]^K} |\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta} - \mathbb{E}\{\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta}\}| > r\Big] \\ &\quad\le \sum_{i=1}^{3} \mathbb{P}\Big\{\sup_{\mathbf{u}\in[0,R]^K} |\Delta_i(\mathbf{u})| > r_l\Big\}.\end{aligned}$$

To prove the theorem, we bound each term on the right-hand side of (26).

To bound the term in (26) involving $\Delta_1(\mathbf{u})$, observe that

$$\begin{aligned}|\Delta_1(\mathbf{u})| &\le \|\boldsymbol{\zeta}\|^2 \|Q(\mathbf{u}) - Q(\dot{\mathbf{u}}_M)\| + |\mathrm{tr}[\mathbb{E}(\boldsymbol{\zeta}\boldsymbol{\zeta}^\top)\{Q(\mathbf{u}) - Q(\dot{\mathbf{u}}_M)\}]| \\ &\le L\|\mathbf{u} - \dot{\mathbf{u}}_M\| \|V^\top V\| \left\{\|\boldsymbol{\zeta}\|^2 + \sum_{i=1}^{d} \mathbb{E}(\zeta_i^2)\right\} \\ &\le K^{1/2} L R 2^{-M} \|V^\top V\| (\|\boldsymbol{\zeta}\|^2 + 4d\gamma^2),\end{aligned}$$

where the second inequality follows from Von Neumann's trace inequality and the last inequality follows from (2). It follows that

$$\begin{aligned}(27)\qquad \mathbb{P}\Big\{\sup_{\mathbf{u}\in[0,R]^K} |\Delta_1(\mathbf{u})| > r_1\Big\} &\le \mathbb{P}\Big\{\frac{K^{1/2} L R}{2^M r_1} \|V^\top V\| (\|\boldsymbol{\zeta}\|^2 + 4d\gamma^2) > 1\Big\} \\ &\le \frac{K^{1/2} L R}{2^M r_1} \|V^\top V\| \mathbb{E}(\|\boldsymbol{\zeta}\|^2 + 4d\gamma^2) \\ &\le \frac{8K^{1/2} L R d\gamma^2}{2^M r_1} \|V^\top V\|.\end{aligned}$$

To bound the term in (26) that depends on $\Delta_2(\mathbf{u})$, we use the Hanson–Wright inequality [Theorem 1.1 of Rudelson and Vershynin (2013)], which implies that there is an absolute constant $c > 0$, such that

$$\begin{aligned}(28)\qquad &\mathbb{P}\Big\{\sup_{\mathbf{u}\in[0,R]^K} |\Delta_2(\mathbf{u})| > r_2\Big\} \\ &\quad= \mathbb{P}[|\boldsymbol{\zeta}^\top Q(0)\boldsymbol{\zeta} - \mathbb{E}\{\boldsymbol{\zeta}^\top Q(0)\boldsymbol{\zeta}\}| > r_2] \\ &\quad\le 2\exp\Big[-c\min\Big\{\frac{r_2^2}{\gamma^4 \|Q(0)\|_{\mathrm{HS}}^2}, \frac{r_2}{\gamma^2 \|Q(0)\|}\Big\}\Big] \\ &\quad\le 2\exp\Big[-c\min\Big\{\frac{r_2^2}{\gamma^4 \|V^\top V\|^2 \|T(0)\|_{\mathrm{HS}}^2}, \frac{r_2}{\gamma^2 \|V^\top V\| \|T(0)\|}\Big\}\Big]\end{aligned}$$

(specifically, the first inequality above follows from the Hanson–Wright inequality; the second inequality follows from basic bounds on matrix norms).

Finally, we bound the term in involving $\Delta_3(\mathbf{u})$ in (26). Let $s_1, \dots, s_K \geq 0$ satisfy $s_1 + \cdots + s_K = 1$. Then

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{\mathbf{u}\in[0,R]^K} |\Delta_3(\mathbf{u})| > r_3 \Big\} \\
&\quad \leq \sum_{j=1}^{M} \mathbb{P}\Big\{ \sup_{\mathbf{u}\in[0,R]^K} |\boldsymbol{\zeta}^\top \{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta} \\
&\qquad - \mathbb{E}[\boldsymbol{\zeta}^\top \{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta}]| > s_j r_3 \Big\}.
\end{aligned}$$

By construction, for each $j = 1, \dots, M$ and $i = 1, \dots, K$, there is a $k = 1, \dots, 2^J$ such that $\dot{u}_{ij} = k2^{-j}R$. Thus, for each $j = 1, \dots, M$ and $i = 1, \dots, K$, there are 2^{jK} possible pairs $(\dot{\mathbf{u}}_j, \dot{\mathbf{u}}_{j-1})$ and it follows that

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{\mathbf{u}\in[0,R]^K} |\boldsymbol{\zeta}^\top \{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta} - \mathbb{E}[\boldsymbol{\zeta}^\top \{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta}]| > s_j r_3 \Big\} \\
&\quad \leq 2^{jK} \max_{(\dot{\mathbf{u}}_j, \dot{\mathbf{u}}_{j-1})} \mathbb{P}\{|\boldsymbol{\zeta}^\top \{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta} \\
&\qquad - \mathbb{E}[\boldsymbol{\zeta}^\top \{Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\}\boldsymbol{\zeta}]| > s_k r_3\} \\
&\quad \leq 2^{jK+1} \max_{(\dot{\mathbf{u}}_j, \dot{\mathbf{u}}_{j-1})} \exp\Big[-c \min\Big\{ \frac{s_j^2 r_3^2}{\gamma^4 \|Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\|_{\mathrm{HS}}^2}, \\
&\qquad \frac{s_j r_3}{\gamma^2 \|Q(\dot{\mathbf{u}}_j) - Q(\dot{\mathbf{u}}_{j-1})\|} \Big\} \Big] \\
&\quad \leq 2^{jK+1} \exp\Big[-c \min\Big\{ \frac{4^j s_j^2 r_3^2}{\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{2^j s_j r_3}{\gamma^2 \|V^\top V\| K^{1/2} L R} \Big\} \Big],
\end{aligned}$$

where we have used the Hanson–Wright inequality again in the third line above. We conclude that

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{\mathbf{u}\in[0,R]^K} |\Delta_3(\mathbf{u})| > r_3 \Big\} \\
&\quad \leq \sum_{j=1}^{M} 2^{jK+1} \exp\Big[-c \min\Big\{ \frac{4^j s_j^2 r_3^2}{\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{2^j s_j r_3}{\gamma^2 \|V^\top V\| K^{1/2} L R} \Big\} \Big].
\end{aligned}$$

Now take $s_j = (1/3) \cdot (3/4)^j$, for $j = 1, \dots, M-1$, and $s_M = 1 - (s_1 + \cdots + s_{M-1}) > (1/3) \cdot (3/4)^M$. Then

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{\mathbf{u}\in[0,R]^K} |\Delta_3(\mathbf{u})| > r_3 \Big\} \\
&\quad \leq \sum_{j=1}^{M} 2^{jK+1} \exp\Big[-c \min\Big\{ \frac{(9/4)^j r_3^2}{9\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{(3/2)^j r_3}{3\gamma^2 \|V^\top V\| K^{1/2} L R} \Big\} \Big]
\end{aligned}$$

$$\leq \sum_{k=1}^{K} \exp\Bigg[(jK+1)\log(2) - c\min\left\{ \frac{(9/4)^j r_3^2}{9\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{(3/2)^j r_3}{3\gamma^2 \|V^\top V\| K^{1/2} L R} \right\} \Bigg].$$

If

$$\frac{r_3^2}{\gamma^4 \|V^\top V\|^2 K L^2 R^2 m} \geq \frac{225K^2}{\min\{c, c^2\}}, \tag{29}$$

then

$$(jK+1)\log(2) - c\min\left\{ \frac{(9/4)^j r_3^2}{9\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{(3/2)^j r_3}{3\gamma^2 \|V^\top V\| K^{1/2} L R} \right\}$$
$$\leq -j\log(2) - c\min\left\{ \frac{r_3^2}{9\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{r_3}{3\gamma^2 \|V^\top V\| K^{1/2} L R} \right\}.$$

Hence, if (29) holds,

$$\mathbb{P}\Big\{ \sup_{\mathbf{u}\in[0,R]^K} |\Delta_3(\mathbf{u})| > r_3 \Big\} \leq \exp\Bigg[-c\min\left\{ \frac{r_3^2}{9\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{r_3}{3\gamma^2 \|V^\top V\| K^{1/2} L R} \right\} \Bigg]. \tag{30}$$

To complete the proof, we combine (26)–(28) and (30), and let $K \to \infty$, to obtain

$$\mathbb{P}\Big[\sup_{\mathbf{u}\in[0,R]^K} |\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta} - \mathbb{E}\{\boldsymbol{\zeta}^\top Q(\mathbf{u})\boldsymbol{\zeta}\}| > r \Big]$$
$$\leq 2\exp\Bigg[-c\min\left\{ \frac{r_2^2}{\gamma^4 \|V^\top V\|^2 \|T(0)\|_{\mathrm{HS}}^2}, \frac{r_2}{\gamma^2 \|V^\top V\| \|T(0)\|} \right\} \Bigg]$$
$$+ \exp\Bigg[-c\min\left\{ \frac{r_3^2}{9\gamma^4 \|V^\top V\|^2 K L^2 R^2 m}, \frac{r_3}{3\gamma^2 \|V^\top V\| K^{1/2} L R} \right\} \Bigg],$$

whenever (29) holds. The theorem follows by taking, say, $r_1 = r_2 = r_3 = r/3$.

APPENDIX B: PROOF OF THEOREM 2

We follow the proof of Theorem 2.1 in Reinert and Röllin (2009), and use Stein's method with exchangeable pairs. Let $f : \mathbb{R}^{2K} \to \mathbb{R}$ be a three-times differentiable function. By Lemma 2.6 in Reinert and Röllin (2009), there is a 3-times differentiable function $g : \mathbb{R}^{2K} \to \mathbb{R}$ satisfying the Stein identity

$$\mathbb{E}\{f(\mathbf{w})\} - \mathbb{E}\{f(V^{1/2}\mathbf{w})\} = \mathbb{E}\{\nabla^\top V \nabla g(\mathbf{w}) - \mathbf{w}^\top \nabla g(\mathbf{w})\} \tag{31}$$

and $|\frac{\partial^k g(\mathbf{x})}{\prod_{j=1}^k \partial x_{i_j}}| \le \frac{1}{k}|\frac{\partial^k f(\mathbf{x})}{\prod_{j=1}^k \partial x_{i_j}}|$ for all $\mathbf{x} = (x_1, \ldots, x_{2K})^\top \in \mathbb{R}^{2K}$, $k = 1, 2, 3$, and $i_j \in \{1, \ldots, 2K\}$. To prove the theorem, we bound

$$S = \mathbb{E}\{\nabla^\top V \nabla g(\mathbf{w}) - \mathbf{w}^\top \nabla g(\mathbf{w})\}. \tag{32}$$

Next, we use exchangeability. Let $\boldsymbol{\zeta}' = (\zeta_1', \ldots, \zeta_d')^\top$ be an independent copy of $\boldsymbol{\zeta}$, and let $\underline{i} \in \{1, \ldots, d\}$ be an independent and uniformly distributed random index. Define the vector $\mathbf{w}' \in \mathbb{R}^{2K}$ exactly as we defined $\mathbf{w}$, except that $\zeta_{\underline{i}}$ is replaced with $\zeta_{\underline{i}}'$ throughout. More precisely, let $\mathbf{e}_i \in \mathbb{R}^d$ be the ith standard basis vector in $\mathbb{R}^d$ and define

$$\begin{aligned} w_k' &= \{\boldsymbol{\zeta} + (\zeta_{\underline{i}}' - \zeta_{\underline{i}})\mathbf{e}_{\underline{i}}\}^\top Q_k \{\boldsymbol{\zeta} + (\zeta_{\underline{i}}' - \zeta_{\underline{i}})\mathbf{e}_{\underline{i}}\} \\ &= w_k + 2(\zeta_{\underline{i}}' - \zeta_{\underline{i}})\mathbf{e}_{\underline{i}}^\top Q_k \boldsymbol{\zeta} + \mathbf{e}_{\underline{i}}^\top Q_k \mathbf{e}_{\underline{i}} (\zeta_{\underline{i}}' - \zeta_{\underline{i}})^2, \end{aligned} \tag{33}$$

$$\begin{aligned} \breve{w}_k' &= \{\boldsymbol{\zeta} + (\zeta_{\underline{i}}' - \zeta_{\underline{i}})\mathbf{e}_{\underline{i}}\}^\top \breve{Q}_k \{\boldsymbol{\zeta} + (\zeta_{\underline{i}}' - \zeta_{\underline{i}})\mathbf{e}_{\underline{i}}\} - \operatorname{tr}(Q_k) \\ &= \breve{w}_k + \mathbf{e}_{\underline{i}}^\top Q_k \mathbf{e}_{\underline{i}} \{(\zeta_{\underline{i}}')^2 - \zeta_{\underline{i}}^2\}, \end{aligned} \tag{34}$$

for $k = 1, \ldots, K$. Then $\mathbf{w}' = (w_1', \breve{w}_1', \ldots, w_K', \breve{w}_K')^\top \in \mathbb{R}^{2K}$.

Let us compute $\mathbb{E}(w_k' - w_k|\boldsymbol{\zeta})$ and $\mathbb{E}(\breve{w}_k' - \breve{w}_k|\boldsymbol{\zeta})$. By (33)–(34),

$$\begin{aligned} \mathbb{E}(w_k' - w_k|\boldsymbol{\zeta}) &= \mathbb{E}\left\{2(\zeta_{\underline{i}}' - \zeta_{\underline{i}}) \sum_{j=1}^d q_{\underline{i}j}^{(k)} \zeta_j + q_{\underline{i}\underline{i}}^{(k)} (\zeta_{\underline{i}}' - \zeta_{\underline{i}})^2 \middle| \boldsymbol{\zeta}\right\} \\ &= -\frac{2}{d} \sum_{i,j=1}^d q_{ij}^{(k)} \zeta_i \zeta_j + \frac{1}{d} \sum_{i=1}^N q_{ii}^{(k)} \zeta_i^2 + \frac{1}{N} \operatorname{tr}(Q_k) \\ &= -\frac{2}{d} w_k + \frac{1}{d} \breve{w}_k \end{aligned}$$

and $\mathbb{E}(\breve{w}_k' - \breve{w}_k|\boldsymbol{\zeta}) = \mathbb{E}[q_{\underline{i}\underline{i}}^{(k)}\{(\zeta_{\underline{i}}')^2 - \zeta_{\underline{i}}^2\}|\boldsymbol{\zeta}] = -\frac{1}{d}\breve{w}_k$. The identity (4), which was stated in the main text, follows.

Next, we will work our way back to the Stein identity (32) and take advantage of (4). Define $G(\mathbf{x}', \mathbf{x}) = \frac{1}{2}(\mathbf{x}' - \mathbf{x})^\top \Lambda_K^{-\top}\{\nabla g(\mathbf{x}') + \nabla g(\mathbf{x})\}$, for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{2K}$. By exchangeability, $\mathbb{E}\{G(\mathbf{w}', \mathbf{w})\} = 0$. Thus,

$$\begin{aligned} 0 &= \frac{1}{2}\mathbb{E}[(\mathbf{w}' - \mathbf{w})^\top \Lambda_K^{-\top}\{\nabla g(\mathbf{w}') + \nabla g(\mathbf{w})\}] \\ &= \mathbb{E}\{(\mathbf{w}' - \mathbf{w})^\top \Lambda_K^{-\top} \nabla g(\mathbf{w})\} \\ &\quad + \frac{1}{2}\mathbb{E}[(\mathbf{w}' - \mathbf{w})^\top \Lambda_K^{-\top}\{\nabla g(\mathbf{w}') - \nabla g(\mathbf{w})\}] \\ &= -\mathbb{E}\{\mathbf{w}^\top \nabla g(\mathbf{w})\} + \frac{1}{2}\mathbb{E}[(\mathbf{w}' - \mathbf{w})^\top \Lambda_K^{-\top}\{\nabla g(\mathbf{w}') - \nabla g(\mathbf{w})\}], \end{aligned} \tag{35}$$

where we used (4) in the last step. Now we Taylor expand and use some other basic manipulations to get a direct connection between (32) and (35). Indeed, by Taylor's theorem,

$$\begin{aligned}
&(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\{\nabla g(\mathbf{w}')-\nabla g(\mathbf{w})\}\\
&\quad=(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\nabla^2 g(\mathbf{w})(\mathbf{w}'-\mathbf{w})+(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\mathbf{r}^{(2)}\\
&\quad=\operatorname{tr}\big[(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\nabla^2 g(\mathbf{w})\big]+(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\mathbf{r}^{(2)},
\end{aligned}$$

where $\mathbf{r}^{(2)}=(r_1^{(2)},\dots,r_{2K}^{(2)})^\top$, $r_k^{(2)}=(\mathbf{w}'-\mathbf{w})^\top R_k^{(2)}(\mathbf{w}'-\mathbf{w})$, and each $R_k^{(2)}=(R_{ijk}^{(2)})$ is a $2K\times 2K$ matrix with $|R_{ijk}^{(2)}|\le (1/2)|f|_3$. Thus, by (35),

$$\begin{aligned}
\mathbb{E}\{\mathbf{w}^\top\nabla g(\mathbf{w})\}&=\frac{1}{2}\mathbb{E}\operatorname{tr}\big[(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\nabla^2 g(\mathbf{w})\big]\\
&\quad+\frac{1}{2}\mathbb{E}\{(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\mathbf{r}^{(2)}\}.
\end{aligned}\tag{36}$$

Since

$$\begin{aligned}
\mathbb{E}\{(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top\}&=2\mathbb{E}\{\mathbf{w}(\mathbf{w}-\mathbf{w}')^\top\}\\
&=2\mathbb{E}(\mathbf{w}\mathbf{w}^\top\Lambda_K^\top)=2V\Lambda_K^\top,
\end{aligned}\tag{37}$$

it follows that

$$\begin{aligned}
\mathbb{E}\{\nabla^\top V\nabla g(\mathbf{w})\}&=\frac{1}{2}\mathbb{E}\big[\nabla^\top\mathbb{E}\{(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top\}\Lambda_K^{-\top}\nabla g(\mathbf{w})\big]\\
&=\frac{1}{2}\mathbb{E}\operatorname{tr}\big[\mathbb{E}\{(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top\}\Lambda_K^{-\top}\nabla^2 g(\mathbf{w})\big].
\end{aligned}\tag{38}$$

Combining (32) with (36) and (38) yields

$$\begin{aligned}
S&=\mathbb{E}\{\nabla^\top V\nabla g(\mathbf{w})-\mathbf{w}^\top\nabla g(\mathbf{w})\}\\
&=\frac{1}{2}\mathbb{E}\operatorname{tr}\big[\mathbb{E}\{(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top\}\Lambda_K^{-\top}\nabla^2 g(\mathbf{w})\big]\\
&\quad-\frac{1}{2}\mathbb{E}\operatorname{tr}\{(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\nabla^2 g(\mathbf{w})\}\\
&\quad-\frac{1}{2}\mathbb{E}\{(\mathbf{w}'-\mathbf{w})^\top \Lambda_K^{-\top}\mathbf{r}^{(2)}\}\\
&=\frac{1}{2}S_1+\frac{1}{2}S_2,
\end{aligned}\tag{39}$$

where $S_1=-(1/2)\mathbb{E}\operatorname{tr}\{T\Lambda_K^{-\top}\nabla^2 g(\mathbf{w})\}$, $S_2=-(1/2)\mathbb{E}\{(\mathbf{w}'-\mathbf{w})^\top\Lambda_K^{-\top}\mathbf{r}^{(2)}\}$, and $T=\mathbb{E}\{(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top|\boldsymbol{\zeta}\}-\mathbb{E}\{(\mathbf{w}'-\mathbf{w})(\mathbf{w}'-\mathbf{w})^\top\}$. Thus, in order to bound S it suffices to bound S_1, S_2.

First, we work with S_1. Notice that

$$\begin{aligned}|\mathbb{E}\operatorname{tr}\{T\Lambda_K^{-\top}\nabla^2 g(\mathbf{w})\}| &\le \frac{|f|_2}{2}\|\Lambda_K^{-\top}\|_{\mathrm{HS}}\mathbb{E}(\|T\|_{\mathrm{HS}}) \\ &= \frac{|f|_2 K^{1/2}}{2}\operatorname{tr}\{(\Lambda_1^\top\Lambda_1)^{-1}\}^{1/2}\mathbb{E}(\|T\|_{\mathrm{HS}}) \\ &\le \frac{3}{5}K^{1/2}d|f|_2\mathbb{E}(\|T\|_{\mathrm{HS}}),\end{aligned}$$

where we have used the fact that $\operatorname{tr}\{(\Lambda_1^\top\Lambda_1)^{-1} = (3/2)d^2$. Thus,

$$|S_1| \le \frac{3}{10}K^{1/2}d|f|_2\mathbb{E}(\|T\|_{\mathrm{HS}}). \tag{40}$$

It requires a bit more work to bound $\mathbb{E}(\|T\|_{\mathrm{HS}})$ in (40).

The matrix T can be written as

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1K} \\ T_{21} & T_{22} & \cdots & T_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ T_{K1} & T_{K2} & \cdots & T_{KK}\end{bmatrix} \qquad \text{where } T_{kl} = \begin{bmatrix} t_{11}^{kl} & t_{12}^{kl} \\ t_{21}^{kl} & t_{22}^{kl}\end{bmatrix},$$

and

$$t_{11}^{kl} = \mathbb{E}\{(w_k' - w_k)(w_l' - w_l)|\boldsymbol{\zeta}\} - \mathbb{E}\{(w_k' - w_k)(w_l' - w_l)\}, \tag{41}$$

$$t_{12}^{kl} = \mathbb{E}\{(w_k' - w_k)(\breve{w}_l' - \breve{w}_l)|\boldsymbol{\zeta}\} - \mathbb{E}\{(w_k' - w_k)(\breve{w}_l' - \breve{w}_l)\}, \tag{42}$$

$$t_{21}^{kl} = \mathbb{E}\{(\breve{w}_k' - \breve{w}_k)(w_l' - w_l)|\boldsymbol{\zeta}\} - \mathbb{E}\{(\breve{w}_k' - \breve{w}_k)(w_l' - w_l)\}, \tag{43}$$

$$t_{22}^{kl} = \mathbb{E}\{(\breve{w}_k' - \breve{w}_k)(\breve{w}_l' - \breve{w}_l)|\boldsymbol{\zeta}\} - \mathbb{E}\{(\breve{w}_k' - \breve{w}_k)(\breve{w}_l' - \breve{w}_l)\}. \tag{44}$$

We conclude that $\mathbb{E}(\|T\|_F) \le [\sum_{k,l=1}^K \sum_{i,j=1}^2 \mathbb{E}\{(t_{ij}^{kl})^2\}]^{1/2}$ and, furthermore, if we can control each of the terms $\mathbb{E}\{(t_{ij}^{kl})^2\}$, then a bound on $\mathbb{E}(\|T\|_F)$ will follow. Fortunately, Lemma S2 from the supplemental article gives bounds for on these moments. Indeed, let $c(\gamma) = 4096(\gamma+1)^8$ and $q_{\max} = \max_{k=1,\ldots,K}\|Q\|$. It follows from Lemma S2 that

$$\begin{aligned}\mathbb{E}(\|T\|_F) &\le \frac{Kq_{\max}^2}{d^{1/2}}\big[8\{108c(\gamma)^2 + 763c(\gamma) + 930\} \\ &\qquad + 4\{24c(\gamma)^2 + 69c(\gamma) + 1\} + c(\gamma) + 4\big]^{1/2} \\ &\le \frac{Kq_{\max}^2}{d^{1/2}}\{65c(\gamma) + 104\}.\end{aligned}$$

Combining this bound on $\mathbb{E}(\|T\|_F)$ with (40) yields

$$|S_1| \le 4\{5c(\gamma) + 8\}K^{3/2}d^{1/2}|f|_2 q_{\max}^2. \tag{45}$$

Next, we bound S_2. First, consider the basic inequalities

$$
\begin{aligned}
|S_2| &\le \frac{1}{2}\|\Lambda_K^{-\top}\|\mathbb{E}(\|\mathbf{w}'-\mathbf{w}\|\|\mathbf{r}^{(2)}\|) \\
&\le \frac{d}{4}\left\|\begin{bmatrix}1 & 0\\ 1 & 2\end{bmatrix}\right\|\mathbb{E}\left\{\|\mathbf{w}'-\mathbf{w}\|^3\left(\sum_{k=1}^{2K}\|R_k^{(2)}\|^2\right)^{1/2}\right\} \\
&\le \frac{5\cdot 2^{1/2}}{8}K^{3/2}d|f|_3\mathbb{E}(\|\mathbf{w}'-\mathbf{w}\|^3).
\end{aligned}
\tag{46}
$$

Now focus on bounding $\mathbb{E}(\|\mathbf{w}'-\mathbf{w}\|^3)$. Each inequality in the following chain is elementary:

$$
\begin{aligned}
&\mathbb{E}(\|\mathbf{w}'-\mathbf{w}\|^3) \\
&\quad= \mathbb{E}\left[\left\{\sum_{k=1}^{K}(w_k'-w_k)^2+\sum_{k=1}^{K}(\breve{w}_k'-\breve{w}_k)^2\right\}^{3/2}\right] \\
&\quad= \mathbb{E}\left[\left\{\sum_{k=1}^{K}(2(\zeta_{\underline{i}}'-\zeta_{\underline{i}})\mathbf{e}_{\underline{i}}^\top Q_k\boldsymbol{\zeta}+\mathbf{e}_{\underline{i}}^\top Q_k\mathbf{e}_I(\zeta_{\underline{i}}'-\zeta_{\underline{i}})^2)^2\right.\right. \\
&\qquad\left.\left.+\sum_{k=1}^{K}(\mathbf{e}_{\underline{i}}^\top Q_k\mathbf{e}_I)^2\{(\zeta_{\underline{i}}')^2-\zeta_{\underline{i}}^2\}^2\right\}^{3/2}\right] \\
&\quad\le 2^{3/2}\mathbb{E}\left[\left\{8\sum_{k=1}^{K}\{(\zeta_{\underline{i}}')^2+\zeta_{\underline{i}}^2\}(\mathbf{e}_I^\top Q_k\boldsymbol{\zeta})^2+9\sum_{k=1}^{K}\|Q_k\|^2\{(\zeta_{\underline{i}}')^4+\zeta_{\underline{i}}^4\}\right\}^{3/2}\right] \\
&\quad= \frac{2^{3/2}}{d}\sum_{i=1}^{d}\mathbb{E}\left[\left\{8\sum_{k=1}^{K}\{(\zeta_i')^2+\zeta_i^2\}(\mathbf{e}_i^\top Q_k\boldsymbol{\zeta})^2\right.\right. \\
&\qquad\left.\left.+9\sum_{k=1}^{K}\|Q_k\|^2\{(\zeta_i')^4+\zeta_i^4\}\right\}^{3/2}\right] \\
&\quad\le \frac{150K^{1/2}}{d}\sum_{i=1}^{d}\sum_{k=1}^{K}\mathbb{E}\{(|\zeta_i'|^3+|\zeta_i|^3)|\mathbf{e}_i^\top Q_k\boldsymbol{\zeta}|^3\} \\
&\qquad+\frac{300K^{1/2}}{d}\sum_{i=1}^{d}\sum_{k=1}^{K}\|Q_k\|^3\mathbb{E}(\zeta_i^6) \\
&\quad\le \frac{300K^{1/2}c(\gamma)^{1/2}}{d}\sum_{i=1}^{d}\sum_{k=1}^{K}\sqrt{\mathbb{E}\{(\mathbf{e}_i^\top Q_k\boldsymbol{\zeta})^6\}}+300K^{3/2}c(\gamma)q_{\max}^3.
\end{aligned}
$$

It remains to bound $\mathbb{E}\{(\mathbf{e}_i^\top Q_k \boldsymbol{\zeta})^6\}$. This is accomplished by a version of Khintchine's inequality, given in Corollary 5.12 of Vershynin (2012). It implies that there is an absolute constant $C_1 > 0$ such that $\mathbb{E}\{(\mathbf{e}_i^\top Q_k \boldsymbol{\zeta})^6\} \le C_1^2(\gamma + 1)^6\{\sum_{j=1}^d (q_{ij}^{(k)})^2\}^3$. Thus,

$$\begin{aligned}\mathbb{E}\big(\|\mathbf{w}' - \mathbf{w}\|^3\big) &\le \frac{300CK^{1/2}c(\gamma)^{1/2}(\gamma+1)^3}{d}\sum_{k=1}^K\sum_{i=1}^d\Bigg\{\sum_{j=1}^d (q_{ij}^{(k)})^2\Bigg\}^{3/2} \\ &\quad + 300K^{3/2}c(\gamma)q_{\max}^3 \\ &\le 300K^{3/2}\big\{C_1c(\gamma)^{1/2}(\gamma+1)^3 + c(\gamma)\big\}q_{\max}^3.\end{aligned}$$

Combining this with (46) yields

$$|S_2| \le 266K^3q_{\max}^3\big\{C_1c(\gamma)^{1/2}(\gamma+1)^3 + c(\gamma)\big\}|f|_3 d. \tag{47}$$

Finally, combining (31)–(32), (39), (45) and (47), we obtain

$$\big|\mathbb{E}\big\{f(\mathbf{w})\big\} - \mathbb{E}\big\{f\big(V^{1/2}\mathbf{z}\big)\big\}\big| \le C(\gamma+1)^8\big(K^{3/2}d^{1/2}|f|_2q_{\max}^2 + K^3d|f|_3q_{\max}^3\big)$$

for some absolute constant $C > 0$, which proves the theorem.

SUPPLEMENTARY MATERIAL

Supplement to "Flexible results for quadratic forms with applications to variance components estimation" (DOI: 10.1214/16-AOS1456SUPP; .pdf). The supplementary document Dicker and Erdogdu (2016b) contains proofs of Propositions 1–4, along with statements and proofs of additional auxiliary results.

REFERENCES

ADAMCZAK, R. (2015). A note on the Hanson–Wright inequality for random vectors with dependencies. *Electron. Commun. Probab.* **20** no. 72, 13. MR3407216

BAI, Z. D., MIAO, B. and YAO, J.-F. (2003). Convergence rates of spectral distributions of large sample covariance matrices. *SIAM J. Matrix Anal. Appl.* **25** 105–127. MR2002902

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. MR2567175

CHATTERJEE, S. (2008). A new method of normal approximation. *Ann. Probab.* **36** 1584–1610. MR2435859

DE LOS CAMPOS, G., SORENSEN, D. and GIANOLA, D. (2015). Genomic heritability: What is it? *PLoS Genet.* **11** e1005048.

DEMIDENKO, E. (2013). *Mixed Models*: *Theory and Applications with R*, 2nd ed. Wiley, Hoboken, NJ. MR3235905

DICKER, L. H. and ERDOGDU, M. A. (2016a). Maximum likelihood for variance estimation in high-dimensional linear models. In *Proceedings of the* 19*th International Conference on Artificial Intelligence and Statistics* 159–167. JMLR Workshop & Conference Proceedings.

DICKER, L. H. and ERDOGDU, M. A. (2016b). Supplement to "Flexible results for quadratic forms with applications to variance components estimation." DOI:10.1214/16-AOS1456SUPP.

GOLAN, D. and ROSSET, S. (2011). Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics* **27** i317–i323.

GÖTZE, F. and TIKHOMIROV, A. N. (1999). Asymptotic distribution of quadratic forms. *Ann. Probab.* **27** 1072–1098. MR1699003

GÖTZE, F. and TIKHOMIROV, A. (2002). Asymptotic distribution of quadratic forms and applications. *J. Theoret. Probab.* **15** 423–475. MR1898815

HALL, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14** 1–16. MR0734096

HANSON, D. L. and WRIGHT, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.* **42** 1079–1083. MR0279864

HARTLEY, H. O. and RAO, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* **54** 93–108. MR0216684

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72** 320–340. MR0451550

HENDERSON, C. R. (1950). Abstract: Estimation of genetic parameters. *Ann. Math. Stat.* **21** 309–310.

HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17** no. 52, 6. MR2994877

JIANG, J. (1996). REML estimation: Asymptotic behavior and related topics. *Ann. Statist.* **24** 255–286. MR1389890

JIANG, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statist. Sinica* **8** 861–885. MR1651513

JIANG, J., LI, C., PAUL, D., YANG, C. and ZHAO, H. (2016). On high-dimensional genome-wide association study and misspecified mixed model analysis. *Ann. Statist.* **44** 2127–2160. MR3546446

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

REINERT, G. and RÖLLIN, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.* **37** 2150–2173. MR2573554

RICHARDSON, A. M. and WELSH, A. H. (1994). Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Aust. J. Stat.* **36** 31–43. MR1309503

RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York. MR1190470

SEVASTYANOV, B. A. (1961). A class of limit distributions for quadratic forms of normal stochastic variables. *Theory Probab. Appl.* **6** 337–340.

SPEED, D., HEMANI, G., JOHNSON, M. R. and BALDING, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91** 1011–1021.

VAN DE GEER, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

WHITTLE, P. (1964). On the convergence to normality of quadratic forms in independent variables. *Teor. Veroyatn. Primen.* **9** 113–118. MR0161429

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.

YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. and PRICE, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46** 100–106.

ZAITLEN, N. A. and KRAFT, P. (2012). Heritability in the genome-wide association era. *Hum. Genet.* **131** 1655–1664.

DEPARTMENT OF STATISTICS AND BIOSTATISTICS
RUTGERS UNIVERSITY
110 FRELINGHUYSEN ROAD
501 HILL CENTER
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: ldicker@stat.rutgers.edu

STATISTICS DEPARTMENT
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: erdogdu@stanford.edu