Supplementary materials for "Fast algorithms for conducting large-scale GWAS of age-at-onset traits using Cox mixed-effects models"

Liang He[1*], Alexander M. Kulminski[1*]

[1] Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, NC, USA

* Corresponding authors: Liang He, Alexander Kulminski
Email: lh235@duke.edu, alexander.kulminski@duke.edu

A. Supplementary text

1. Proof that $\tilde{V}_{22}^{-1}$ is the first-order approximation of V_{22}^{-1} in COXMEG-sparse

Here, we show that $\tilde{V}_{22}^{-1} = S^{-1} + S^{-1}QQ^TS^{-1}$ is a first-order approximation of $\boldsymbol{V}_{22}^{-1}$, where $S^{-1} = \left(WB + \frac{\Sigma^{-1}}{\tau}\right)^{-1}$. We follow the notations used in the main text. One useful observation about QQ^T is that the columns corresponding to the censored subjects in $\boldsymbol{M}$ do not contribute to QQ^T because their corresponding elements in $\boldsymbol{A}$ are zero. Denote by $\tilde{\boldsymbol{M}} \in \{0,1\}^{N \times N_1}$ the matrix after removing the columns corresponding to the censored subjects from $\boldsymbol{M}$, where N_1 is the number of subjects experiencing the event of interest. We rewrite H as

$$\begin{aligned} H &= WB - QQ^T \\ &= WB - WMA^2M^{\mathrm{T}}W \\ &= WB - W\tilde{M}\tilde{A}^2\tilde{M}^TW \\ &= WB - \tilde{Q}\tilde{Q}^{\mathrm{T}} \end{aligned}, \quad \text{(S1)}$$

where

$$\tilde{Q} = W\tilde{M}\tilde{A}$$
$$\tilde{A} = \mathrm{diag}^{-1}\{\tilde{M}^TW\mathbf{1}\}, \quad \text{(S2)}$$
$$B = \mathrm{diag}\{MA\mathbf{1}\} = \mathrm{diag}\{\tilde{M}\tilde{A}\mathbf{1}\}. \quad \text{(S3)}$$

Because S^{-1} is always positive definite after removing those subjects censored before the first failure, we then rewrite $\boldsymbol{V}_{22}^{-1}$ in the following way

$$
\begin{aligned}
&V_{22}^{-1} \\
&= \left(WB - QQ^T + \frac{\Sigma^{-1}}{\tau} \right)^{-1} \\
&= \left(\left(WB + \frac{\Sigma^{-1}}{\tau} \right) \left(I - \left(WB + \frac{\Sigma^{-1}}{\tau} \right)^{-1} QQ^T \right) \right)^{-1} \\
&= \left(I - S^{-1} QQ^T \right)^{-1} S^{-1}
\end{aligned}
\quad , \text{(S4)}
$$

Denote by $\|A\|$ the spectral norm of A, which is the largest eigenvalue of A when A is SPD (i.e., $\|A\|$ is $\lambda_{max}(A)$). Now we show that the spectral norm of $S^{-1}QQ^T$ in the last line in e.q. (S4) is strictly less than one, that is,

$$\left\| S^{-1} QQ^T \right\| < 1.$$

Note that both V_{22} and S^{-1} are SPD, so their product

$$
\begin{aligned}
&S^{-1} V_{22} \\
&= S^{-1} (S - QQ^T) \\
&= I - S^{-1} QQ^T
\end{aligned}
$$

has only positive eigenvalues. This implies that all eigenvalues of $S^{-1}QQ^T$ are strictly less than one. Therefore, we can rewrite V_{22}^{-1} by expanding the first inverse term in e.q. (S4) using a Neumann series, which gives

$$
\begin{aligned}
&V_{22}^{-1} \\
&= \left(I - S^{-1} QQ^T \right)^{-1} S^{-1} \\
&= \left(I + S^{-1} QQ^T + S^{-1} QQ^T S^{-1} QQ^T + \cdots \right) S^{-1} \\
&= \sum_{k=0}^{\infty} S^{-1} \left(QQ^T S^{-1} \right)^k
\end{aligned}
\quad ,
$$

from which we obtain $\tilde{V}_{22}^{-1}$ by taking the first two terms in the series.

2. Local convergence of COXMEG-sparse

We first show that COXMEG-sparse is locally convergent, and then discuss the factors affecting its convergence rate. There are multiple ways to investigate the local convergence of COXMEG-sparse. Our approach is to show that COXMEG-sparse belongs to a class of inexact Newton methods (Dembo *et al.* 1982) with the following property

$$\left\| -V(\theta_c)\, e + s(\theta_c) \right\| \le \eta_c \left\| s(\theta_c) \right\|, \text{(S5)}$$

where e is the step change in each iteration in COXMEG-sparse, and $s(\theta_c)$ and $V(\theta_c)$ are the score function and negative Hessian evaluated at the current step θ_c, respectively. The goal is to show that the step e chosen in COXMEG-sparse satisfies inequality (S5) with the forcing term $\eta_c < 1$. We first consider the convergence of using the zero-order approximation of V_{22}^{-1} (i.e., replacing V_{22}^{-1} by S^{-1} in V^{-1}). Note that the zero-order approximation of V_{22}^{-1} amounts to adding QQ^T to the bottom-right corner of V. So, the iteration step in this case is

$$e = \tilde{V}^{-1}(\theta_c)\, s(\theta_c) = \left(V(\theta_c) + Q_0\right)^{-1} s(\theta_c) \quad \text{, (S6)}$$

where we denote $Q_0 = \begin{pmatrix} 0 & 0 \\ 0 & QQ^T \end{pmatrix}$. Plugging e.q. (S6) into the left-hand side of (S5) gives

$$\begin{aligned}
&\left\| -V(\theta_c)\, e + s(\theta_c) \right\| = \left\| -V(\theta_c)\left(V(\theta_c) + Q_0\right)^{-1} s(\theta_c) + s(\theta_c) \right\| \\
&= \left\| \left(I - V(\theta_c)\left(V(\theta_c) + Q_0\right)^{-1}\right) s(\theta_c) \right\| \\
&\le \left\| \left(I - V(\theta_c)\left(V(\theta_c) + Q_0\right)^{-1}\right) \right\| \left\| s(\theta_c) \right\| \\
&= \left\| Q_0\left(V(\theta_c) + Q_0\right)^{-1} \right\| \left\| s(\theta_c) \right\| \\
&= \eta_c \left\| s(\theta_c) \right\|
\end{aligned} \quad .$$

It remains to show that the forcing term η_c is uniformly less than one. Under regularity conditions (e.g., Assumption 2.2.1 in (Kelley 1999)), $V(\theta_c)$ is Lipschitz continuous within a neighbourhood of θ_* and is nonsingular at θ_* for which $s(\theta_*) = 0$. This suggests that we can always find a neighbourhood of θ_* so that the smallest eigenvalue of $V(\theta_c)$ is bounded away from zero. Because the largest eigenvalue of Q_0 is one, we can find a neighbourhood of θ_*, within which $V(\theta_c)\left(V(\theta_c) + Q_0\right)^{-1}$ has all positive eigenvalues bounded away from zero, that is,

$$\lambda_{min}\left(V(\theta_c)\left(V(\theta_c) + Q_0\right)^{-1}\right) \ge \varepsilon > 0, \quad \theta_c \in \delta(\theta_*)$$

where $\lambda_{min}(\cdot)$ denotes the smallest eigenvalue of a matrix. This implies

$$\begin{aligned}
&\lambda_{min}\left(V(\theta_c)\left(V(\theta_c) + Q_0\right)^{-1}\right) \\
&= \lambda_{min}\left(\left(V(\theta_c) + Q_0 - Q_0\right)\left(V(\theta_c) + Q_0\right)^{-1}\right) \\
&= \lambda_{min}\left(I - Q_0\left(V(\theta_c) + Q_0\right)^{-1}\right) \ge \varepsilon > 0 \\
&\Leftrightarrow 1 - \lambda_{max}\left(Q_0\left(V(\theta_c) + Q_0\right)^{-1}\right) \ge \varepsilon \\
&\Leftrightarrow \left\| Q_0\left(V(\theta_c) + Q_0\right)^{-1} \right\| \le 1 - \varepsilon < 1
\end{aligned} \quad .$$

Therefore, according to Theorem 2.3 in (Dembo *et al.* 1982), the algorithm is locally convergent at least linearly in the norm $\|\theta\|_* = \left\|V^{-1}\theta\right\|$ with asymptotic rate constant no greater than $1 - \varepsilon$. When a higher-order (e.g., *K*th-order) approximation is used in COXMEG-sparse, the difference is to replace QQ^T in e.q. (S6) by the positive semidefinite matrix $\left(V_{22}^{-1} - \sum_{k=K+1}^{\infty} S^{-1}(QQ^TS^{-1})^k\right)^{-1} - V_{22}$ with a smaller matrix norm. Therefore, the proof of the local convergence still follows.

The convergence rate is controlled by ε. If ε is equal to 1, in which case no approximation of V_{22}^{-1} is used, the convergence rate becomes quadratic. Therefore, the convergence rate depends on how close the approximation $\tilde{V}_{22}^{-1}$ is to V_{22}^{-1}. We then assess which factors affect the convergence rate practically. Note that V_{22}^{-1} in e.q. (S4) can be written as

$$
\begin{aligned}
&\boldsymbol{V}_{22}{}^{-1} \\
&=\left(\boldsymbol{I}-\boldsymbol{S}^{-1}\boldsymbol{Q}\boldsymbol{Q}^{T}\right)^{-1}\boldsymbol{S}^{-1} \\
&=\left(\boldsymbol{I}-\left(\boldsymbol{W}\boldsymbol{B}\left(\boldsymbol{I}+\left(\boldsymbol{W}\boldsymbol{B}\right)^{-1}\frac{\boldsymbol{\Sigma}^{-1}}{\tau}\right)\right)^{-1}\boldsymbol{Q}\boldsymbol{Q}^{T}\right)^{-1}\boldsymbol{S}^{-1} \\
&=\left(\boldsymbol{I}-\left(\boldsymbol{I}+\left(\boldsymbol{W}\boldsymbol{B}\right)^{-1}\frac{\boldsymbol{\Sigma}^{-1}}{\tau}\right)^{-1}\boldsymbol{B}^{-1}\boldsymbol{W}^{-1}\boldsymbol{Q}\boldsymbol{Q}^{T}\right)^{-1}\boldsymbol{S}^{-1}
\end{aligned}
\quad \text{, (S7)}
$$

where the inverses are valid because of SPD of W, B and WB after removing all censored samples before the first occurrence of the event of interest. In COXMEG-sparse, we use a Neumann series to approximate the first inverse in (S7), so a lower-order approximation would have worse performance, at least in some direction, if the spectral norm of

$$
\left(\boldsymbol{I}+\left(\boldsymbol{W}\boldsymbol{B}\right)^{-1}\frac{\boldsymbol{\Sigma}^{-1}}{\tau}\right)^{-1}\boldsymbol{B}^{-1}\boldsymbol{W}^{-1}\boldsymbol{Q}\boldsymbol{Q}^{T} \quad \text{(S8)}
$$

is close to one. To investigate the spectral norm of (S8), we substitute e.q. (S1), (S2) and (S3) into $\boldsymbol{Q}\boldsymbol{Q}^{T}$ in (S8), which gives

$$
\begin{aligned}
&\left(\boldsymbol{I}+\left(\boldsymbol{W}\boldsymbol{B}\right)^{-1}\frac{\boldsymbol{\Sigma}^{-1}}{\tau}\right)^{-1}\boldsymbol{B}^{-1}\boldsymbol{W}^{-1}\boldsymbol{Q}\boldsymbol{Q}^{T} \\
&=\left(\boldsymbol{I}+\left(\boldsymbol{W}\boldsymbol{B}\right)^{-1}\frac{\boldsymbol{\Sigma}^{-1}}{\tau}\right)^{-1}\boldsymbol{B}^{-1}\tilde{\mathbf{M}}\tilde{\mathbf{A}}\tilde{\mathbf{Q}}^{T} \\
&=\left(\boldsymbol{I}+\left(\boldsymbol{W}\boldsymbol{B}\right)^{-1}\frac{\boldsymbol{\Sigma}^{-1}}{\tau}\right)^{-1}\underbrace{\mathrm{diag}^{-1}\{\tilde{\mathbf{M}}\tilde{\mathbf{A}}1\}\tilde{\mathbf{M}}\tilde{\mathbf{A}}}_{S_I}\underbrace{\mathrm{diag}^{-1}\{\tilde{\mathbf{M}}^{\mathrm{T}}W1\}\tilde{\mathbf{M}}^{T}W}_{S_{II}}
\end{aligned}
$$

It is clear that S_I and S_{II} are recognized as two rectangular row stochastic (also called Markov) matrices. We show that $\left\|S_I S_{II}\right\|=1$ (i.e., the spectral norm of $S_I S_{II}$ is one), and all its eigenvalues are between zero and one. First, we can easily verify that 1 is an eigenvector of $S_I S_{II}$ with eigenvalue one. Then, suppose that $\boldsymbol{\nu}$ is an eigenvector of $S_I S_{II}$, and $\boldsymbol{\nu}_{\mathrm{k}}$ is the element that has the largest absolute value of the non-zero elements in $\boldsymbol{\nu}$, we have

$$
\begin{aligned}
&\left|\boldsymbol{\lambda}\boldsymbol{\nu}_{\mathrm{k}}\right|=\left|\left(S_I S_{II}\boldsymbol{\nu}\right)_k\right| \\
&=\left|\nu_1\sum_i s_{Iki}s_{IIi1}+\nu_2\sum_i s_{Iki}s_{IIi2}+\ldots+\nu_N\sum_i s_{Iki}s_{IIiN}\right| \\
&\le\left|\sum_i s_{Iki}s_{IIi1}+\sum_i s_{Iki}s_{IIi2}+\ldots+\sum_i s_{Iki}s_{IIiN}\right|\left|\nu_k\right| \\
&=\left|\sum_j\sum_i s_{Iki}s_{IIij}\right|\left|\nu_k\right| \\
&=\left|\sum_i s_{Iki}\sum_j s_{IIij}\right|\left|\nu_k\right| \\
&=\left|\sum_i s_{Iki}\right|\left|\nu_k\right| \\
&=\left|\nu_k\right|
\end{aligned}
\quad ,
$$

from which we obtain $|\lambda| \leq 1$ (i.e., all eigenvalues have its absolute value no larger than one). Because $\boldsymbol{WB}$ is positive definite, $S_I S_{II}$ is similar to $(WB)^{-\frac{1}{2}} Q Q^T (WB)^{-\frac{1}{2}}$, which implies that all eigenvalues of $S_I S_{II}$ are non-negative. Summarizing all evidence above implies that the eigenvalues of $S_I S_{II}$ are between zero and one. Next, we show that all eigenvalues of $\left(I + (WB)^{-1} \frac{\Sigma^{-1}}{\tau}\right)^{-1}$ are less than one. In fact, because $\boldsymbol{WB}$ and Σ^{-1} are SPD, $(WB)^{-1}\Sigma^{-1}$ is similar to $(WB)^{-\frac{1}{2}} \Sigma^{-1} (WB)^{-\frac{1}{2}}$, which is a quadratic form and has all positive eigenvalues denoted by $\lambda_1, \cdots, \lambda_N > 0$. Therefore, the eigenvalues of $\left(I + (WB)^{-1} \frac{\Sigma^{-1}}{\tau}\right)^{-1}$ are $\left(1 + \frac{\lambda_1}{\tau}\right)^{-1}, \cdots, \left(1 + \frac{\lambda_N}{\tau}\right)^{-1}$, which are strictly less than one.

In summary, we have

$$\begin{aligned} &\left\| \left(I + (WB)^{-1} \frac{\Sigma^{-1}}{\tau}\right)^{-1} B^{-1} W^{-1} Q Q^T \right\| \\ &\leq \left\| \left(I + (WB)^{-1} \frac{\Sigma^{-1}}{\tau}\right)^{-1} \right\| \left\| B^{-1} W^{-1} Q Q^T \right\| \\ &\leq \left(1 + \frac{\lambda_N}{\tau}\right)^{-1} < 1 \end{aligned},$$

which suggests that the spectral norm of (S8) is bounded by $\left(1 + \frac{\lambda_N}{\tau}\right)^{-1}$. From this bound, we can see that a larger τ would generally drop the convergence rate. This is also confirmed by our simulation study (Figure S4), which shows that a higher-order approximation has much better performance (in terms of steps of convergence) for a larger τ (e.g., τ>0.2). In addition, the spectral density of Σ^{-1} also affects the convergence rate through λ_N. A large condition number of Σ^{-1} would slow down the convergence, which is also corroborated by the results of our simulation study (Figure S4). Higher-order approximation converged much faster under larger block sizes and stronger correlation, in which cases the condition number of the relatedness matrix is larger. This simulation study (Figure S4) also suggests that the first-order approximation is near-optimal for a common family-based design in which the average family size is 5 and most correlation coefficients are below 0.5.

3. Proof of the validity of the NR method for positive semidefinite covariance matrices

Here, we prove that using the GPPL $\tilde{l}_1$, the NR method is still valid for Σ being SPSD except that the sum of elements in each row of Σ is zero (i.e., Σ has eigenvector 1 with eigenvalue 0). We first show that when Σ is SPSD and has eigenvector 1 with eigenvalue 0, V_{22}, and thus V

become always non-invertible, which violates the regularity condition of the NR method. It is shown in A.2 that $WB - QQ^T$ is always positive semidefinite and has eigenvector 1 with eigenvalue 0. Combined with the fact that $\Sigma^- 1 = 0$ if $\Sigma 1 = 0$, we have

$$\begin{aligned} V_{22}\mathbf{1} &= \left(\mathbf{WB} - \mathbf{QQ}^{\mathrm{T}} + \frac{\boldsymbol{\Sigma}^-}{\tau}\right)\mathbf{1} \\ &= \left(\mathbf{WB} - \mathbf{QQ}^{\mathrm{T}}\right)\mathbf{1} + \frac{\boldsymbol{\Sigma}^-}{\tau}\mathbf{1} \\ &= \mathbf{0} \end{aligned},$$

suggesting that in such a case, V_{22} has a zero eigenvalue, and thus is non-invertible.
Next, we show that V_{22} is always invertible when Σ is SPSD and 1 is not one of its eigenvectors corresponding to the eigenvalue 0. We have shown in A.1 that

$$\begin{aligned} & WB - QQ^T \\ &= WB\left(I - B^{-1}\tilde{M}\tilde{A}^2\tilde{M}^T W\right) \\ &= WB(I - S_I S_{II}) \end{aligned},$$

and $S_I S_{II}$ is the product of two rectangular row stochastic matrices, and always has the largest eigenvalue one with eigenvector 1. Because W and $\tilde{A}$ are diagonal matrices with only positive elements, and $\tilde{\mathbf{M}}$ is similar through permutation to a lower-triangular matrix when assuming no ties, it is easy to verify that the elements of $S_I S_{II}$ are all positive. Using Breslow's approximation for ties does not change this conclusion. According to the Perron–Frobenius theorem, $S_I S_{II}$ has a unique largest eigenvalue, which is one. This means that all the other vectors have its eigenvalue strictly less than one. Suppose that ν is an eigenvector of V_{22}. Consider its eigenvalue

$$\begin{aligned} & V_{22}\boldsymbol{\nu} \\ &= WB(I - S_I S_{II})\,\boldsymbol{\nu} + \frac{\boldsymbol{\Sigma}^{-1}}{\tau}\boldsymbol{\nu} \end{aligned}.$$

If ν is 1, then the second term must be positive, and otherwise, the first term must be positive. Thus, the eigenvalues of V_{22} are all positive, which suggests that the NR method is still valid.

4. Approximation of the log-determinant using the SLQ method

We describe the details of estimating the variance component τ when the relatedness matrix Σ is large and fully dense. We estimate τ using the marginal likelihood

$$l_2 = l_1(\widehat{\theta}) - \frac{1}{2}\log\left|\frac{\boldsymbol{J}\left(l_1(\widehat{\theta}), \tau\right)}{2\pi}\right| = l_1(\widehat{\theta}) - \frac{1}{2}\log\left|\frac{\boldsymbol{W}(\widehat{\theta})\,\boldsymbol{B}(\widehat{\theta}) - \boldsymbol{Q}(\widehat{\theta})\,\boldsymbol{Q}(\widehat{\theta})^T + \frac{\boldsymbol{\Sigma}^{-1}}{\tau}}{2\pi}\right|, \quad \text{(S8)}$$

where $\widehat{\theta}$ is obtained by optimizing the PPL. Once l_1 is optimized, the addition step for estimating τ is the evaluation of the log-determinant in e.q. (S8), the time complexity of which is cubic when using a standard Cholesky decomposition. When Σ^{-1} is dense and very large, this evaluation is computationally intensive. Therefore, we resort to a randomized method based on SLQ to

reduce the time complexity to quadratic. We selected the SLQ method for approximating a log-determinant because our preliminary results suggested that it is more accurate than other randomized methods such as Chebyshev orthogonal polynomials (Pace and LeSage 2004; Han *et al.* 2016), and Martin's Taylor expansion (Martin 1992; Barry and Kelley Pace 1999) under the same computational burden. The log-determinant to be approximated is

$$\log\left|\boldsymbol{W}(\widehat{\boldsymbol{\theta}})\,\boldsymbol{B}(\widehat{\boldsymbol{\theta}})-\boldsymbol{Q}(\widehat{\boldsymbol{\theta}})\,\boldsymbol{Q}(\widehat{\boldsymbol{\theta}})^{T}+\frac{\boldsymbol{\Sigma}^{-1}}{\tau}\right|. \quad \text{(S9)}$$

If $\boldsymbol{\Sigma}$ is sparse and $\boldsymbol{\Sigma}^{-1}$ is dense, $\boldsymbol{\Sigma}$ might never be inverted when the sample size is large. In this case, we instead approximate the log-determinant $\frac{1}{2}\log|\tilde{\boldsymbol{J}}\tilde{\boldsymbol{J}}^{T}|$ using SLQ, where

$$\tilde{\boldsymbol{J}}=\boldsymbol{\Sigma}\left(\boldsymbol{W}(\widehat{\boldsymbol{\theta}})\boldsymbol{B}(\widehat{\boldsymbol{\theta}})-\operatorname{diag}\left(\boldsymbol{Q}(\widehat{\boldsymbol{\theta}})\boldsymbol{Q}(\widehat{\boldsymbol{\theta}})^{T}\right)\right)+\tau^{-1}\boldsymbol{I}.$$

The SLQ method works as follows. Given a certain matrix $\boldsymbol{P}$, we first approximate its log-determinant using a Monte Carlo trace estimator

$$\log|\boldsymbol{P}|=\operatorname{tr}(\log(\boldsymbol{P}))\approx\frac{1}{\mathrm{n_m}}\sum_{\mathrm{i}=1}^{\mathrm{n_m}}\mathbf{r}_{\mathrm{i}}^{\mathrm{T}}\log(\boldsymbol{P})\,\mathbf{r}_{\mathrm{i}},$$

where $\mathrm{n_m}$ is the number of Monte Carlo samples, and $\boldsymbol{r}_i$ is an *i.i.d* sample from the Rademacher distribution as proposed in (Hutchinson 1990). Since direct evaluation of $\log(\boldsymbol{P})$ is difficult, we rewrite it as

$$\begin{aligned}
&\frac{1}{\mathrm{n_m}}\sum_{\mathrm{i}=1}^{\mathrm{n_m}}\mathbf{r}_{\mathrm{i}}^{\mathrm{T}}\log(\boldsymbol{P})\,\mathbf{r}_{\mathrm{i}}\\
&=\frac{\mathbf{n}}{\mathbf{n_m}}\sum_{\mathrm{i}=1}^{\mathrm{n_m}}\frac{\mathbf{r}_{\mathrm{i}}}{\sqrt{\mathbf{n}}}^{\mathrm{T}}\mathbf{U}\log(\Lambda)\,\mathbf{U}^{\mathrm{T}}\frac{\mathbf{r}_{\mathrm{i}}}{\sqrt{\mathbf{n}}}\\
&=\frac{\mathbf{n}}{\mathbf{n_m}}\sum_{\mathrm{i}=1}^{\mathrm{n_m}}\tilde{\mathbf{r}}_{\mathrm{i}}^{\mathrm{T}}\log(\Lambda)\,\tilde{\mathbf{r}}_{\mathrm{i}}\\
&=\frac{\mathbf{n}}{\mathbf{n_m}}\sum_{\mathrm{i}=1}^{\mathrm{n_m}}\sum_{\mathrm{j}}\log\left(\lambda_{\mathrm{j}}\right)\tilde{\mathbf{r}}_{\mathrm{ij}}^{2},
\end{aligned}$$

where $\tilde{r}_i=U^T\frac{r_i}{\sqrt{n}}$, and $\tilde{r}_{ij}$ is the j^{th} element in $\tilde{r}_i$. The second summation in the last line is recognized as a Riemann-Stieltjes integral, which can then be approximated using the Gauss quadrature rule

$$\sum_{\mathrm{j}}\log\left(\lambda_{\mathrm{j}}\right)\tilde{\mathbf{r}}_{\mathrm{ij}}^{2}=\int_{\lambda_{min}}^{\lambda_{max}}\log(\lambda)\,\mathrm{d}\tilde{\mathbf{r}}_{\mathrm{i}}^{2}(\lambda)\approx\sum_{k=1}^{n_q}\omega_{ik}\log\left(\varphi_{ik}\right),$$

where n_q is the number of points in the Gauss quadrature rule, and ω_{ik} and φ_{ik} are the weights and nodes in the Gauss quadrature rule to be determined. It is nicely shown in e.g., Theorem 4.1 in (Golub and Meurant 2009), that the nodes φ_{ik} are the zeros in the Lanczos orthogonal polynomial of $\boldsymbol{P}$, which are the eigenvalues of the tridiagonal matrix $T_{Lanczos}$ corresponding to the orthogonal polynomial (as shown in Theorem 6.2 in (Golub and Meurant 2009)), which is the output of the Lanczos algorithm. The weights ω_{ik} are the square of the first element of the

eigenvector k of $T_{Lanczos}$. One potential issue is that the Lanczos algorithm is numerically unstable due to round-off errors. Even for a modest n_q, it is possible that the vectors produced by the algorithm become dependent. To prevent this issue, we stopped the algorithm once an off-diagonal element in $T_{Lanczos}$ was smaller than a small number (e.g., 1e-10).

The accuracy of the SLQ approximation highly depends on n_q and $\mathrm{n_m}$. Although theoretical error bounds of the approximation of the log-determinant are given in (Ubaru *et al.* 2017), we investigated its empirical performance in COXMEG. Our results showed that $n_m = 100$ and $n_q = 10$ yielded highly accurate estimate of the variance component τ when the condition number of the relatedness matrix is not too large (Figure S7). We observed that the approximation was poor only in the scenario where the block size was large (500) and also the correlations were high (0.9) (Figure S7), which is relatively rarely encountered in a real data analysis.

5. Simulation study for statistical power and empirical size

We investigated FPR and statistical power of the four methods, COXMEG-score, COXMEG-sparse, coxph with a shared frailty and coxme. Because dense matrices are too time consuming for coxme, we investigated the statistical power using block-diagonal relatedness matrices, and expected that the conclusion should be generalized to dense relatedness matrices. We first assessed the FPR under settings of different variance components, and correlation structures. We found that overall all methods except coxph controlled the type I error rate well. We found that COXMEG-score, COXMEG-sparse, and coxme had almost the same FPR as expected (Figure S2). In contrast, our results showed that coxph with a shared frailty had inflated FPR when the correlation coefficients were 0.5, while the power was slightly diminished when the correlation coefficients were 0.1 (Figure S2). When the correlation coefficients were 0.9, coxph with a shared frailty controlled the FPR as well as the other methods (Figure S2), which makes sense because under this correlation, subjects within a block had almost the same random effect.

We next evaluated the statistical power for detecting the effect of a predictor using a simulation study. We considered multiple settings of different sample sizes, proportion of censoring, variance components, and correlation structure. More than 5000 subjects were needed to detect a log(HR) of 0.1. We observed that COXMEG-sparse shared almost the same statistical power as coxme, and COXMEG-score in all settings (Figure S3). We also noted that coxph with a shared frailty had very similar power compared to the other methods (Figure S3).

References

Barry R. P., and R. Kelley Pace, 1999 Monte Carlo estimates of the log determinant of large sparse matrices. Linear Algebra Its Appl. 289: 41–54. https://doi.org/10.1016/S0024-3795(97)10009-X

Dembo R. S., S. C. Eisenstat, and Trond. Steihaug, 1982 Inexact Newton methods. SIAM J. Numer. Anal. 19: 400–408. https://doi.org/10.1137/0719025

Golub G. H., and G. Meurant, 2009 *Matrices, moments and quadrature with applications*. Princeton University Press.

Han I., D. Malioutov, H. Avron, and J. Shin, 2016 Approximating the spectral sums of large-scale matrices using Chebyshev approximations. ArXiv160600942 Cs.

Hutchinson M. F., 1990 A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. Commun. Stat. - Simul. Comput. 19: 433–450. https://doi.org/10.1080/03610919008812866

Kelley C. T., 1999 *Iterative methods for optimization*. SIAM.

Martin R. J., 1992 Approximations to the determinant term in gaussian maximum likelihood estimation of some spatial models. Commun. Stat. - Theory Methods 22: 189–205. https://doi.org/10.1080/03610929308831013

Pace R. K., and J. P. LeSage, 2004 Chebyshev approximation of log-determinants of spatial weight matrices. Comput. Stat. Data Anal. 45: 179–196. https://doi.org/10.1016/S0167-9473(02)00321-3

Ubaru S., J. Chen, and Y. Saad, 2017 Fast estimation of tr(f(A)) via stochastic Lanczos quadrature. SIAM J. Matrix Anal. Appl. 38: 1075–1099.

B. Supplementary Tables

CHR	POS	SNP	Gene	P-value
1	161155392	rs4575098	*ADAMTS4*	9.420E-01
1	207786828	rs2093760	*CR1*	8.568E-03
2	127891427	rs4663105	*BIN1*	2.961E-04
2	233981912	rs10933431	*INPP5D*	1.154E-01
4	11026028	rs6448453	*CLNK*	2.367E-02
6	32583357	rs6931277	*HLA-DRB1*	4.983E-02
6	47432637	rs9381563	*CD2AP*	9.214E-01
7	99971834	rs1859788	*ZCWPW1*	3.128E-01
7	143108158	rs7810606	*EPHA1*	6.735E-01

8	27464929	rs4236673	*CLU*	5.996E-02
10	11717397	rs11257238	*ECHDC3*	6.853E-01
11	59958380	rs2081545	*MS4A6A*	3.141E-02
11	85776544	rs867611	*PICALM*	2.537E-03
14	92938855	rs12590654	*SLC24A4*	1.468E-01
15	59022615	rs442495	*ADAM10*	3.412E-01
15	63569902	rs117618017	*APH1B*	6.624E-01
16	31133100	rs59735493	*KAT8*	6.414E-01
17	5138980	rs113260531	*SCIMP*	7.153E-02
17	47450775	rs28394864	*ABI3*	1.861E-02
19	1039323	rs111278892	*ABCA7*	3.925E-01
19	51727962	rs3865444	*CD33*	8.028E-01

20	54998544	rs6014724	*CASS4*	2.500E-02

Table S1. Results of 23 common significant SNPs identified by a recent meta-analysis of AD. The p-values of HRs were obtained from COXMEG-score. The positions of the SNPs are based on hg19.

C. Supplementary Figures

Figure S1: Comparison of computational time of estimating the variance component for a block-diagonal relatedness matrix under different sample sizes between coxme, COXMEG-sparse, and coxph with a shared frailty. The family (block) size is 5. The evaluation was performed on a Windows Core i5-6300HQ machine.

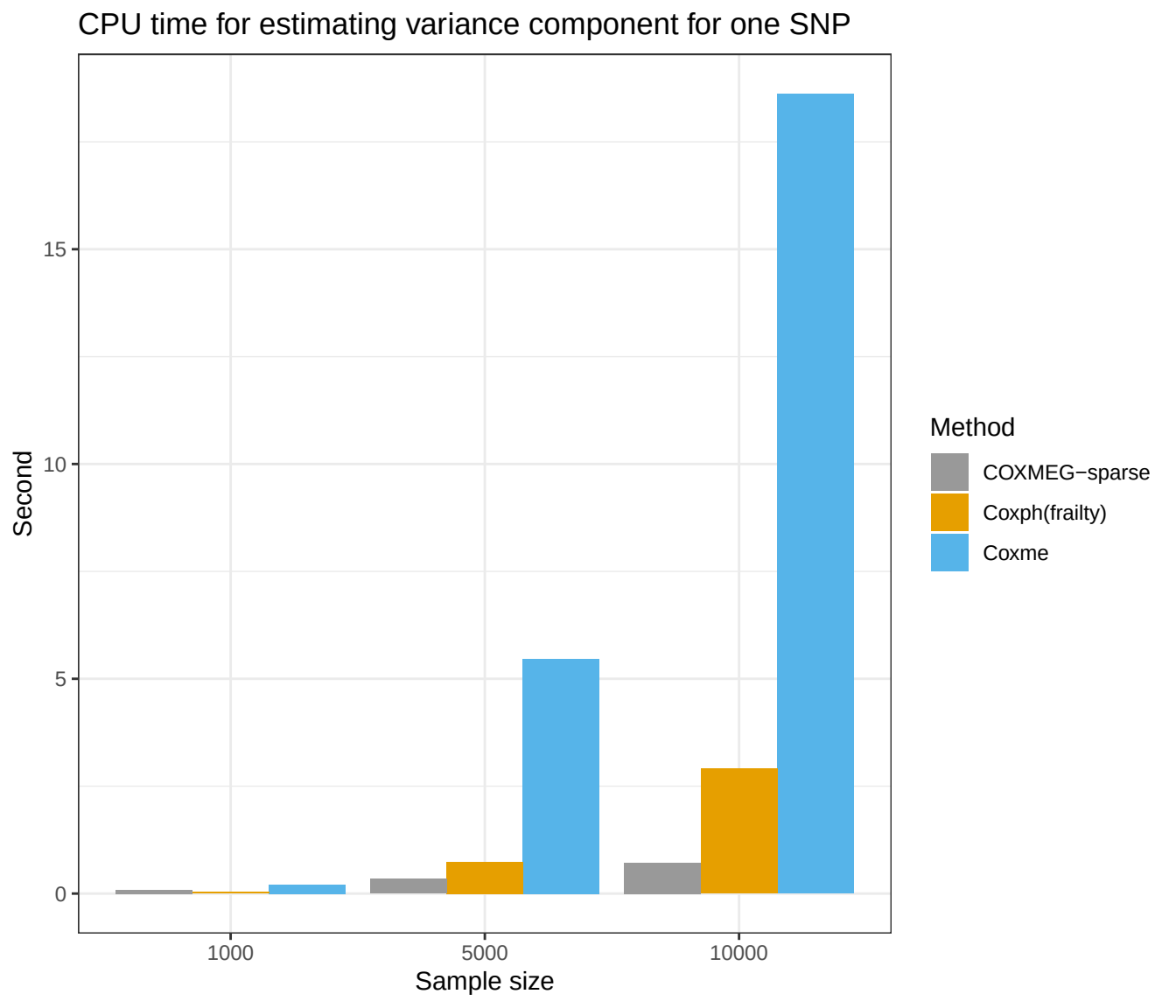

CPU time for estimating variance component for one SNP
Second
15
10
5
0
1000
5000
10000
Sample size
Method
COXMEG−sparse
Coxph(frailty)
Coxme

Figure S2. Comparison of empirical FPR of coxme, COXMEG-sparse, COXMEG-score, and coxph with a shared frailty. The relatedness matrices are block-diagonal correlation matrices with the block size ranging between 5-100. We evaluated the FPR for the correlation ρ in each block being 0.1, 0.5, and 0.9, and the variance component τ between 0.02 and 0.5. The red dots are the mean FPR.

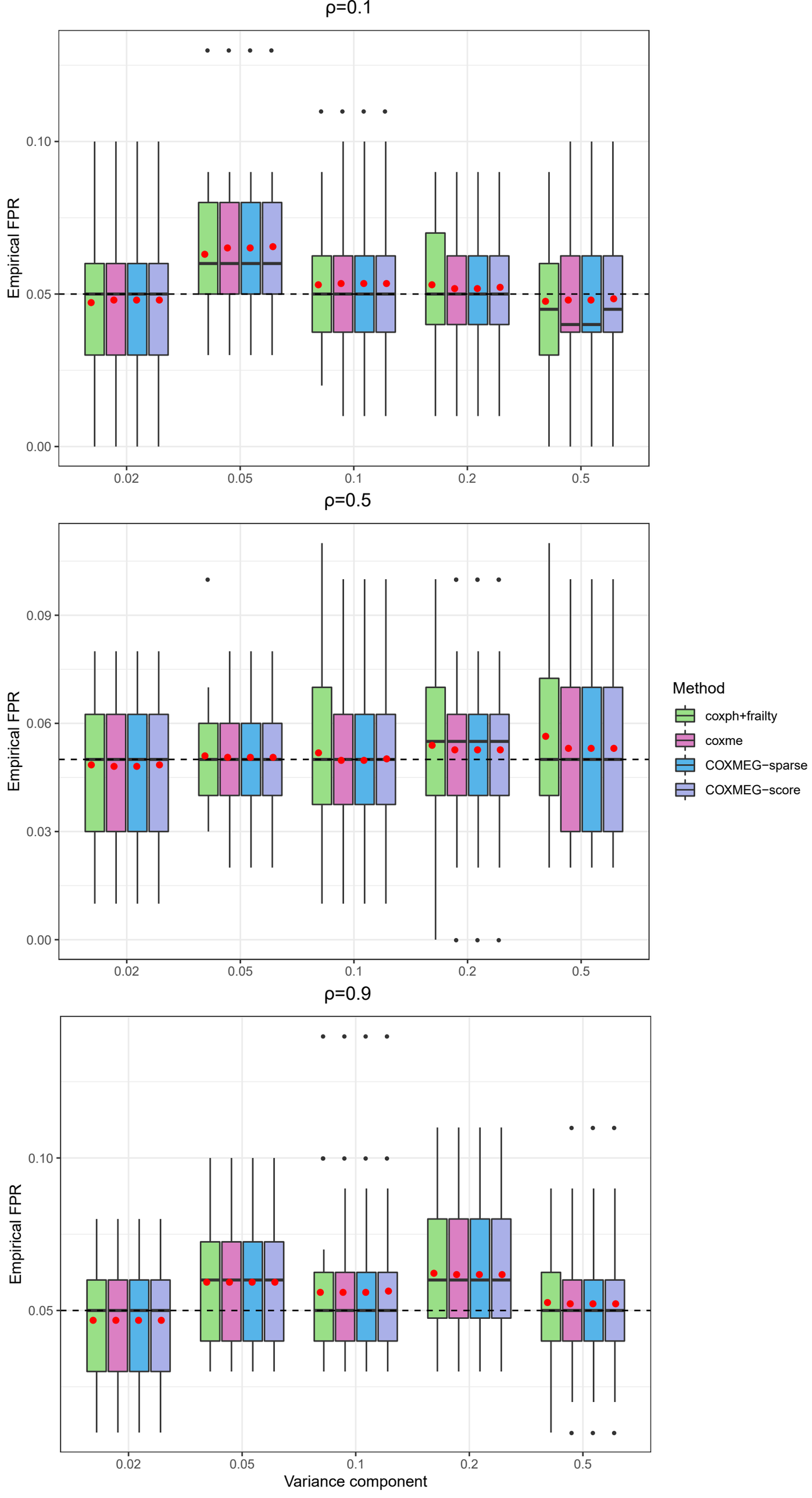

ρ=0.1
ρ=0.5
ρ=0.9
Empirical FPR
0.10
0.05
0.00
0.09
0.06
0.03
0.02
0.05
0.1
0.2
0.5
Variance component
Method
coxph+frailty
coxme
COXMEG−sparse
COXMEG−score

Figure S3. Comparison of empirical statistical power of coxme, COXMEG-sparse, COXMEG-score, and coxph with a shared frailty. The relatedness matrices are block-diagonal correlation matrices with the block size ranging between 5-100 and the correlation ρ ranging between 0.1 and 0.9. We evaluated the power for the HRs being 0.01, 0.05, and 0.1, and the sample size between 1000 and 5000. We also assess the power under no censoring, moderate censoring and heavy censoring.

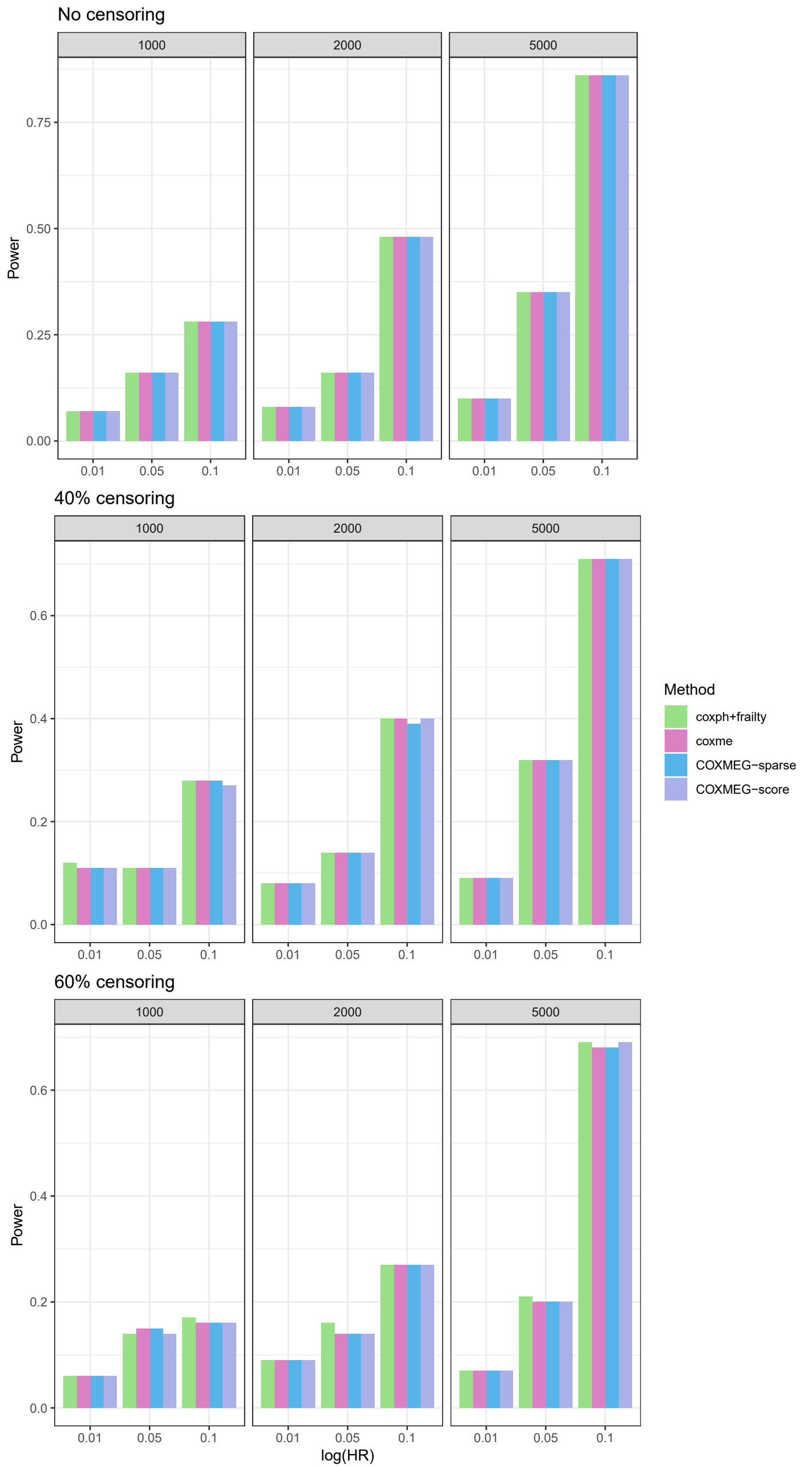

No censoring
40% censoring
60% censoring
1000
2000
5000
Power
0.00
0.25
0.50
0.75
0.0
0.2
0.4
0.6
0.01
0.05
0.1
log(HR)
Method
coxph+frailty
coxme
COXMEG−sparse
COXMEG−score

Figure S4. Evaluation of the convergence rate of higher-order approximations in COXMEG-sparse. The relatedness matrix used in the simulation is a block-diagonal correlation matrix with the block size ranging between 5-100 and the correlation ρ ranging between 0.1 and 0.9. For each setting, the convergence rate was measured by the time for estimating the HRs of one predictor given a variance component τ ranging from 0.02 to 0.5.

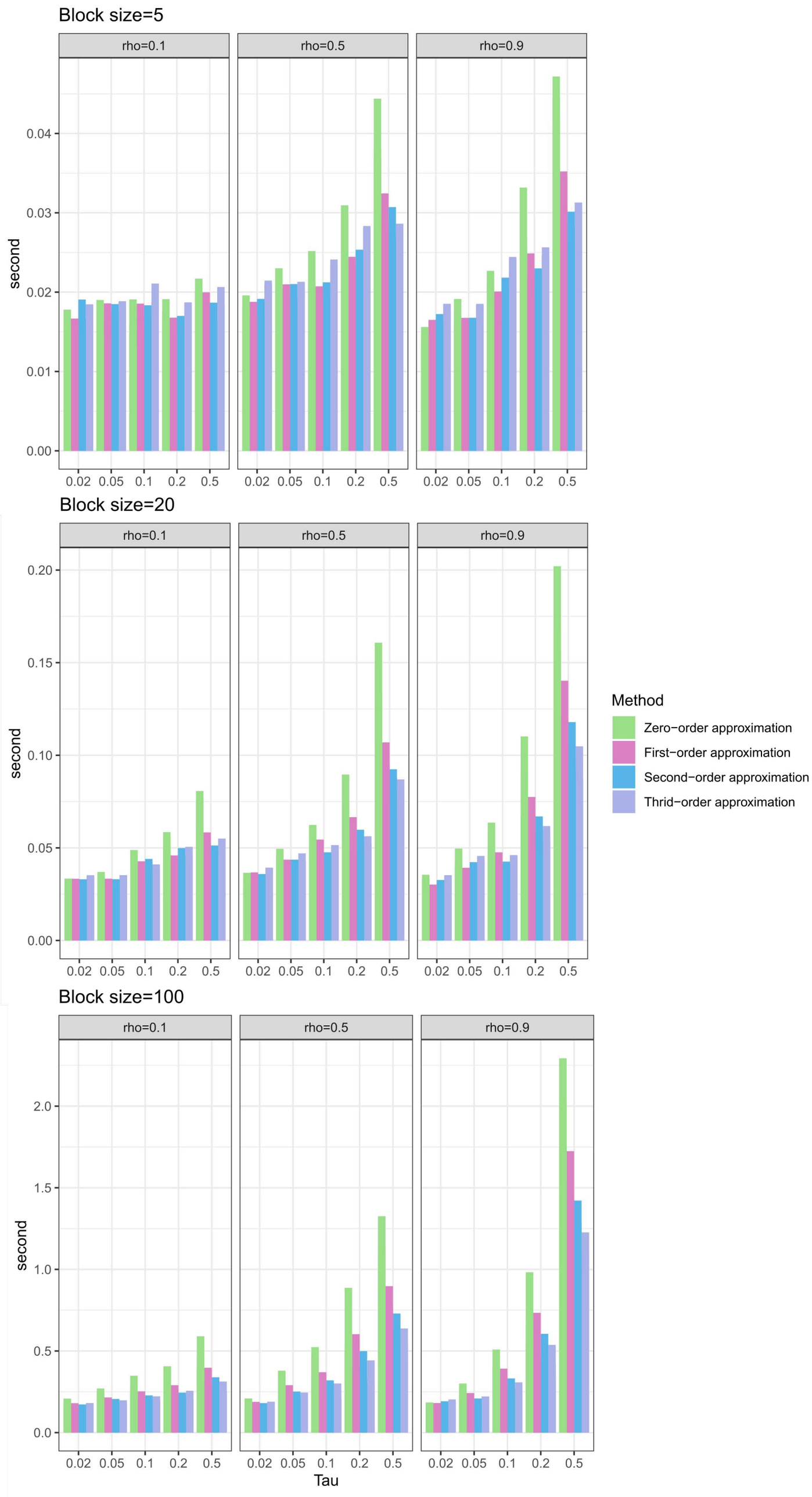

Block size=5
rho=0.1
rho=0.5
rho=0.9
second
0.00
0.01
0.02
0.03
0.04
0.02 0.05 0.1 0.2 0.5
Block size=20
rho=0.1
rho=0.5
rho=0.9
second
0.00
0.05
0.10
0.15
0.20
0.02 0.05 0.1 0.2 0.5
Method
Zero−order approximation
First−order approximation
Second−order approximation
Thrid−order approximation
Block size=100
rho=0.1
rho=0.5
rho=0.9
second
0.0
0.5
1.0
1.5
2.0
0.02 0.05 0.1 0.2 0.5
Tau

Figure S5. Evaluation of the computational performance of three methods (RcppEigen::LDLT, Matrix::solve using the Cholesky decomposition, and RcppEigen::CG with diagonal preconditioned) for solving the sparse linear system in COXMEG-sparse. The relatedness matrix used in the simulation is a block-diagonal correlation matrix with the block size varying between 5-500 and the correlation ρ being 0.5. For each setting, the convergence rate was measured by the time for estimating the HRs of one predictor given a variance component τ ranging from 0.02 to 0.5.

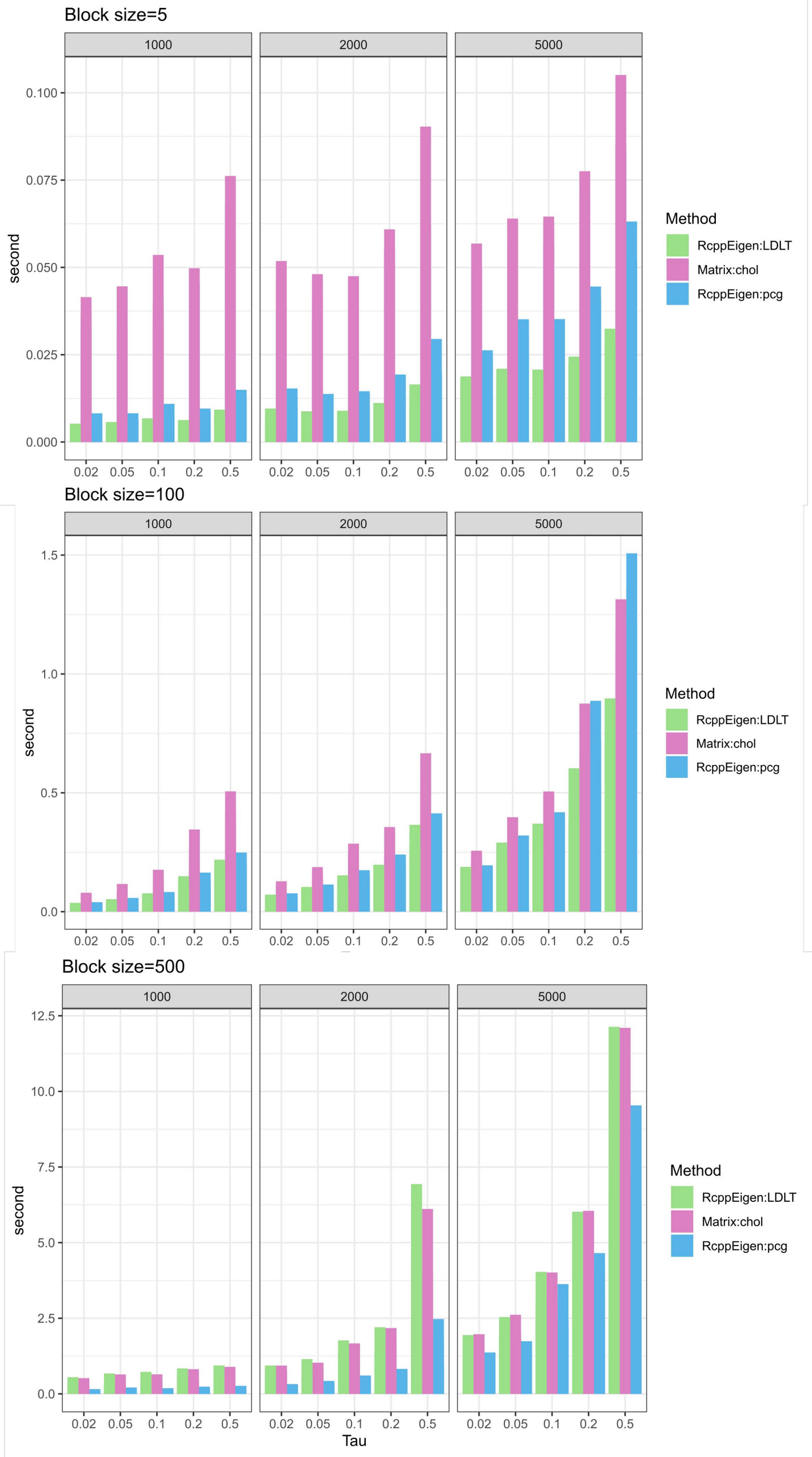

Block size=5
1000
2000
5000
second
0.100
0.075
0.050
0.025
0.000
0.02
0.05
0.1
0.2
0.5
Method
RcppEigen:LDLT
Matrix:chol
RcppEigen:pcg
Block size=100
1.5
1.0
0.5
0.0
Block size=500
12.5
10.0
7.5
5.0
2.5
Tau

Figure S6. Comparison of estimated variance components by coxme, COXMEG-sparse with the exact log-determinant, and COXMEG-sparse with a diagonal approximated log-determinant. (A) COXMEG-sparse with a diagonal approximated log-determinant vs. coxme; (B) COXMEG-sparse with the exact log-determinant vs. coxme; (C) COXMEG-sparse with a diagonal approximated log-determinant vs. the exact log-determinant.

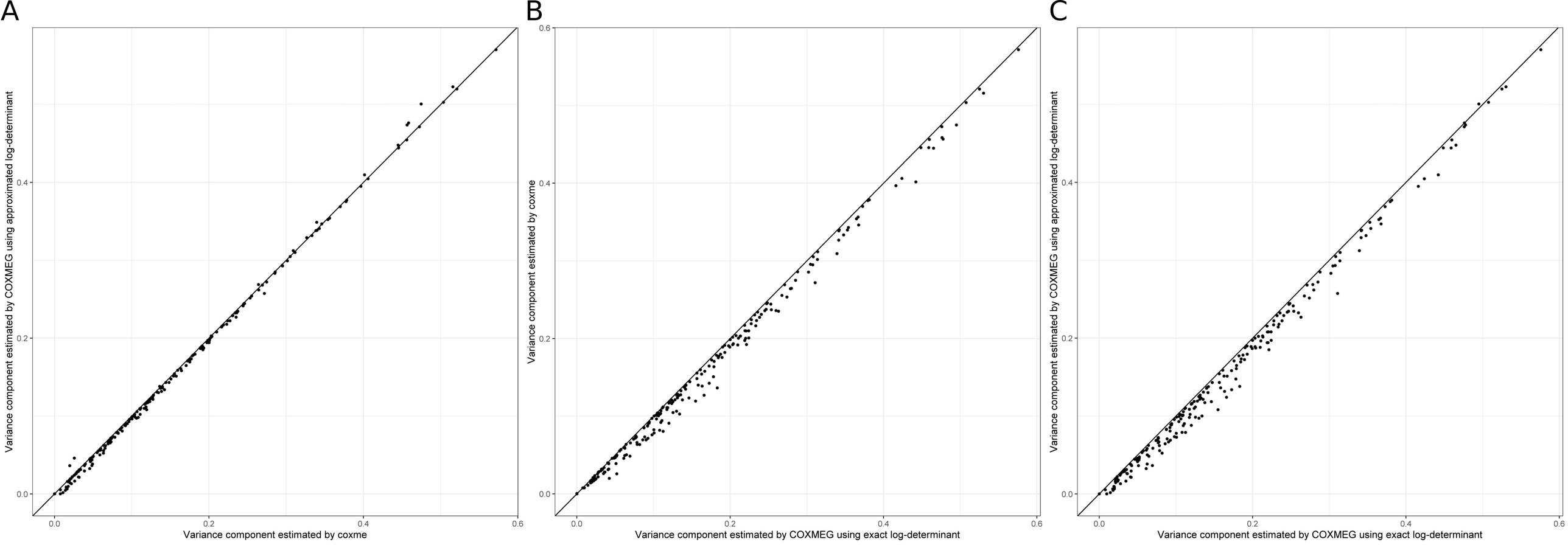

A
Variance component estimated by COXMEG using approximated log-determinant
Variance component estimated by coxme
0.0
0.2
0.4
0.6
B
Variance component estimated by coxme
Variance component estimated by COXMEG using exact log-determinant
C
Variance component estimated by COXMEG using approximated log-determinant
Variance component estimated by COXMEG using exact log-determinant

Figure S7. Comparison of estimated variance components by COXMEG-score with the exact log-determinant, and COXMEG-score with the SLQ approximation under sample sizes of 5000 and 10,000. The relatedness matrix used in the simulation is a block-diagonal correlation matrix with the block size varying between 5-500 and the correlation ρ between 0.1 and 0.9.

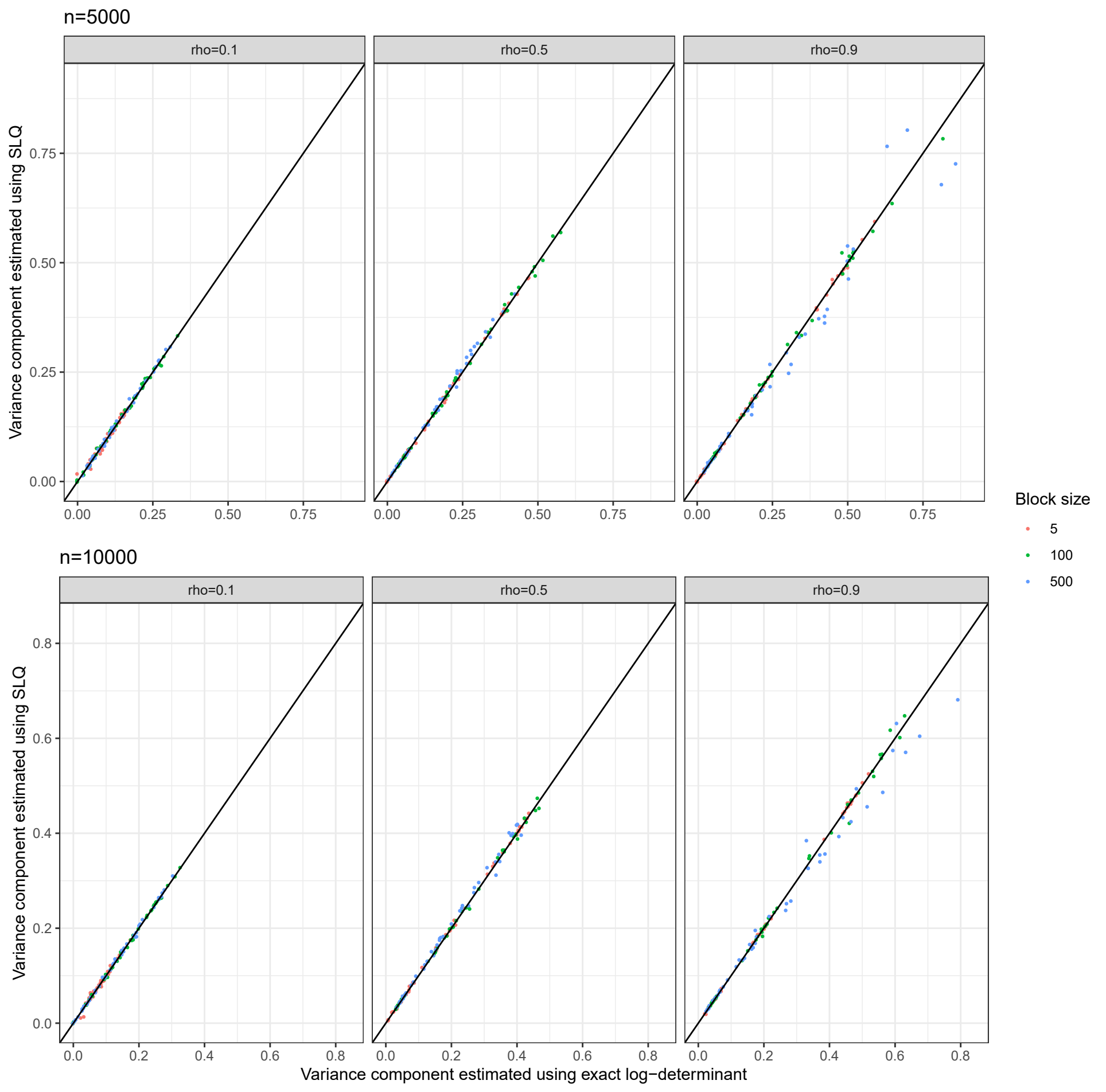
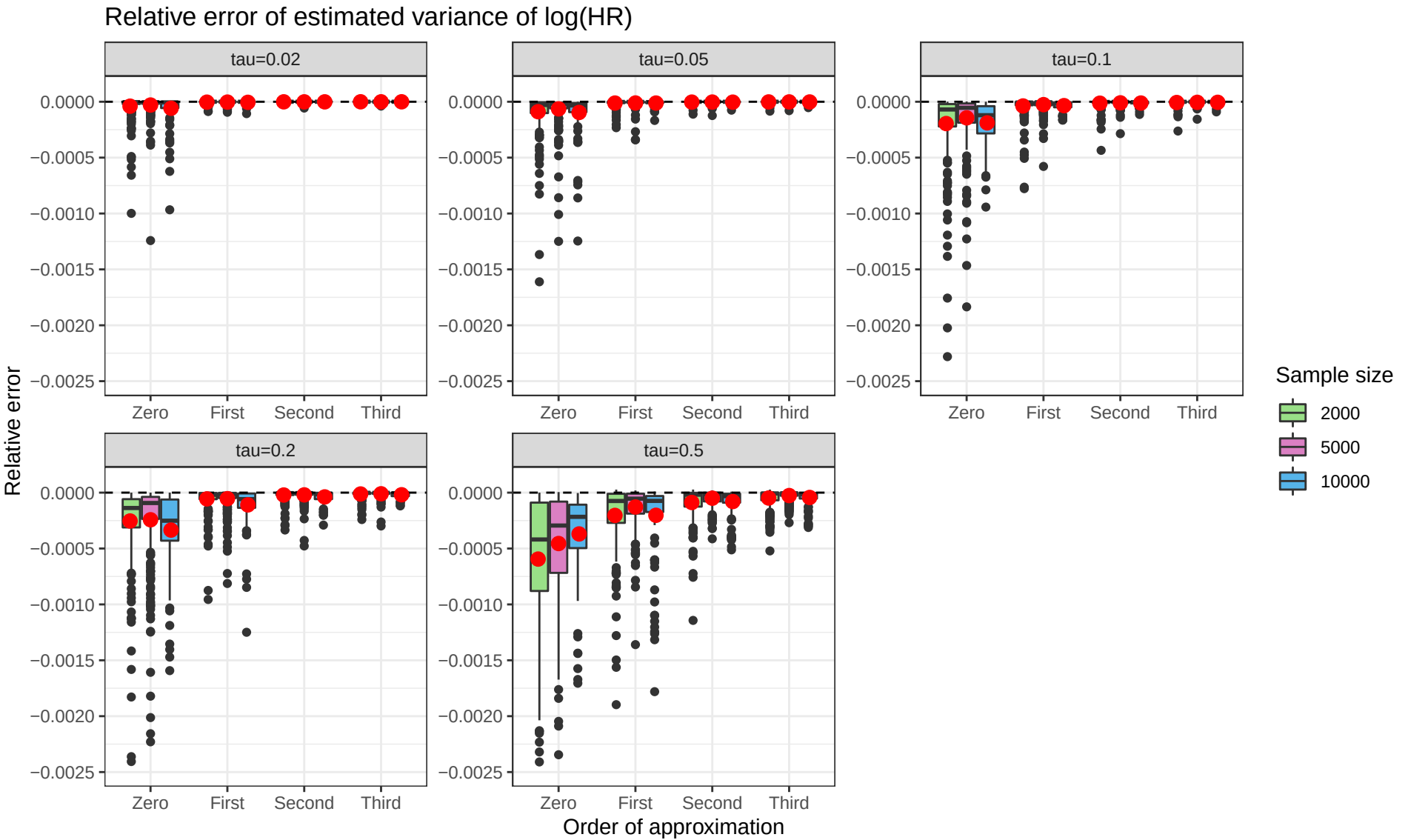

n=5000
rho=0.1
rho=0.5
rho=0.9
Variance component estimated using SLQ
0.00
0.25
0.50
0.75
n=10000
rho=0.1
rho=0.5
rho=0.9
Variance component estimated using SLQ
0.0
0.2
0.4
0.6
0.8
Variance component estimated using exact log−determinant
Block size
5
100
500

Figure S8. Relative error between the variance of log(HR) estimated using an approximated V_{22}^{-1} in COXMEG-sparse and using an exact Hessian matrix. Four approximations of V_{22}^{-1} (Zero-order to Third-order) were evaluated under settings of different sample sizes, and variance components. The red dots are the mean values.


Relative error of estimated variance of log(HR)
tau=0.02
tau=0.05
tau=0.1
tau=0.2
tau=0.5
0.0000
−0.0005
−0.0010
−0.0015
−0.0020
−0.0025
Zero
First
Second
Third
Relative error
Order of approximation
Sample size
2000
5000
10000