

Supplementary Material for “A regression modeling approach to structured shrinkage estimation”

Sihai Dave Zhao*

Department of Statistics, University of Illinois at Urbana-Champaign
and
William Biscarri
Capital One Financial Corporation

January 10, 2021

Abstract

These materials contain illustrations of the regression approach to shrinkage estimation, an estimator for the asymptotic variance of the proposed unconstrained estimator, and proofs of all results.

Keywords: Compound decision, Shrinkage, Empirical Bayes, James-Stein

*The authors gratefully acknowledge Dr. Jingshu Wang for her advice on the single-cell transcriptomics analysis.

1 Examples of regression modeling in shrinkage estimation

This section illustrates how the proposed regression approach can be applied to common simultaneous estimation problems.

Example 1. Let $Y_i \sim N(\theta_i, \sigma^2)$ with σ^2 known. In model (3), let \mathbf{C}_{Ci} and \mathbf{C}_{ki} be null and $C_{Yi} = 1$, so that $\delta_i(\mathbf{Y}, \mathbf{C}, \mathcal{N}) = \beta Y_i$. The vector \mathbf{X}_i equals Y_i and the empirical risk function (5) becomes

$$\hat{R}_n(\beta) = -\sigma^2 + 2\sigma^2\beta + \frac{1}{n} \sum_{i=1}^n (Y_i - Y_i\beta)^2.$$

This is minimized by $\hat{\beta} = 1 - n\sigma^2 / \sum_i Y_i^2$. The resulting estimate $\hat{\beta}Y_i$ for θ is nearly identical to the estimator of James and Stein (1961).

Example 2. Let $Y_i \sim N(\theta_i, \sigma_i^2)$ with σ_i^2 known. The σ_i^2 vary with i and may be informative for θ_i , so they should be included as a covariate in model (3). Let $\mathbf{C}_{Ci} = (1, \sigma_i^2)^\top$, $\mathbf{C}_{Yi} = 1$, and \mathbf{C}_{ki} be null, so that $\delta_i(\mathbf{Y}, \mathbf{C}, \mathcal{N}) = \beta_0 + \beta_1\sigma_i^2 + \beta_2Y_i$. Though not at first evident, it can be shown (Biscarri, 2019) that the resulting estimate for θ after empirical risk minimization is nearly identical to the subspace shrinkage estimator of Oman (1982).

Example 3. Let Y_i follow the state-space model in equation (14) of Greenshtein et al. (2019), where for $i = 1, \dots, n$,

$$Y_i = \theta_i + \epsilon_i \sim N(0, 1), \quad \theta_i = \Phi\theta_{i-1} + U_i$$

for some Φ , where U_i are independent random variables and ϵ_i and U_i are independent. This suggests that both Y_i and Y_{i-1} should be used to estimate θ_i . Define the index

neighborhood $\mathcal{N}_i = \{i-1\}$ and let $\mathbf{C}_{Ci} = 1$, $\mathbf{C}_{Yi} = 1$, and $\mathbf{C}_{ki} = 1$ for $k = 1$. Then model (3) becomes $\delta(\mathbf{Y}, \mathbf{C}, \mathcal{N}) = \beta_0 + \beta_1 Y_i + \beta_2 Y_{i-1}$.

Though not discussed in detail here, the regression approach also encompasses more complex regression models. For example, let $Y_i \sim N(\theta_i, \sigma_i^2)$ with σ_i^2 known. Motivated by a Bayesian hierarchical model, several authors (Kou and Yang, 2017; Xie et al., 2012, 2016) have considered semiparametric estimators of the form $(1 - b_i)Y_i + b_i\tilde{\theta}_i$, where $\tilde{\theta}_i$ can depend on covariates but not Y_i and the $b_i \in [0, 1]$ obey $b_i \leq b_j$ whenever $\sigma_i^2 \leq \sigma_j^2$. This satisfies the intuition that θ_i should be close to Y_i for observations with small variance σ_i^2 . Similarly, model (3) could be extended to allow components of the regression coefficient β to depend on i , and the constrained estimator (7) could be used to impose the same ordering constraints on the coefficients.

2 Estimating \mathbf{V}_n (12)

An estimate of the matrix \mathbf{V}_n (12) from Theorem 5 is presented here. As in model (3), partition the covariate vector \mathbf{C}_i into \mathbf{C}_{Ci} , \mathbf{C}_{Yi} , and \mathbf{C}_{ki} , $k = 1, \dots, q$, and denote the lengths of the component vectors by p_C , p_Y , and p_k respectively. Define the $p \times p$ block matrix

$$\hat{\mathbf{V}}_n = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} & \mathbf{I} \end{pmatrix} \quad (1)$$

where \mathbf{A} , \mathbf{B} , and \mathbf{C} have rows indexed by $j = 1, \dots, p_C$ and entries equal to

$$\begin{aligned} A_{jj'} &= \frac{1}{n} \sum_{i=1}^n C_{Cij} C_{Cij'} \sigma_i^2, j' = 1, \dots, p_C, \\ B_{jj'} &= \frac{1}{n} \sum_{i=1}^n C_{Cij} C_{Yij'} (\sigma_i^2 Y_i + \kappa_{3i}), j' = 1, \dots, p_Y, \\ C_{j,kj'} &= \frac{1}{n} \sum_{i=1}^n C_{Cij} C_{kij'} \sigma_i^2 Y_{i_k}, k = 1, \dots, q, j' = 1, \dots, p_k, \end{aligned}$$

\mathbf{D} , \mathbf{E} and \mathbf{F} have rows indexed by $j = 1, \dots, p_Y$ and entries equal to

$$\begin{aligned} D_{jj'} &= B_{j'j}, j' = 1, \dots, p_C, \\ E_{jj'} &= \frac{1}{n} \sum_{i=1}^n C_{Yij} C_{Yij'} (\sigma_i^2 Y_i^2 + 2Y_i \kappa_{3i} + \kappa_{4i} - 2\sigma_i^4), j' = 1, \dots, p_Y, \\ F_{j,kj'} &= \frac{1}{n} \sum_{i=1}^n C_{Yij} C_{kij'} Y_{i_k} (\sigma_i^2 Y_i + \kappa_{3i}), k = 1, \dots, q, j' = 1, \dots, p_k, \end{aligned}$$

and \mathbf{G} , \mathbf{H} , and \mathbf{I} have rows indexed by kj , $k = 1, \dots, q$, $j = 1, \dots, p_k$ and entries equal to

$$\begin{aligned} G_{kj,j'} &= C_{j',kj}, j' = 1, \dots, p_C, \\ H_{kj,j'} &= F_{j',kj}, j' = 1, \dots, p_Y, \\ I_{kj,k'j'} &= \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} \sigma_i^2 Y_{i_k} Y_{i_{k'}} + \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \sigma_i^2 \sigma_l^2 \delta_{il_{k'}} \delta_{i_k l}, k' = 1, \dots, q, j' = 1, \dots, p_k, \end{aligned}$$

where $\delta_{ab} = 1$ if $a = b$ and 0 otherwise.

3 Proof of Proposition 1

Let $\epsilon_i = Y_i - \theta_i$. Then the true risk $R_n(\boldsymbol{\beta})$ obeys

$$\begin{aligned} R_n(\boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^n E \{ \epsilon_i^2 - 2\epsilon_i(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}) + (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2 \} \\ &= -\frac{1}{n} \sum_{i=1}^n \sigma_i^2 - \frac{2}{n} \left(\sum_{i=1}^n E\epsilon_i \mathbf{X}_i \right)^\top \boldsymbol{\beta} + \frac{1}{n} \sum_{i=1}^n E\{(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2\}. \end{aligned}$$

By the definition of \mathbf{X}_i in , each coordinate of \mathbf{X}_i is of the form C_{ij} , $Y_i C_{ij}$, or $Y_{i_k} C_{ij}$, for $j = 1, \dots, p$ and $k = 1, \dots, q$. Since $E\epsilon_i C_{ij} = 0$, $E\epsilon_i Y_{i_k} C_{ij} = 0$ because $i_k \neq i$ by definition, and $E\epsilon_i Y_i C_{ij} = \sigma_i^2 C_{ij}$, $E\epsilon_i \mathbf{X}_i$ can be expressed as $\sigma_i^2 \partial \mathbf{X}_i / \partial Y_i$. Finally, $\sum_i \sigma_i^2 \partial \mathbf{X}_i / \partial Y_i = E\mathbf{Z}$ because $\hat{\sigma}_i^2$ is an unbiased estimate of σ_i^2 .

4 Proof of Theorem 1

Let $\epsilon_i = Y_i - \theta_i$. Since

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2 = \frac{1}{n} \sum_{i=1}^n \{ \epsilon_i^2 + 2\epsilon_i(\theta_i - \mathbf{X}_i^\top \boldsymbol{\beta}) + (\theta_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2 \},$$

it follows that

$$\begin{aligned} \sup_{\boldsymbol{\beta} \in \mathcal{M}_n} |\hat{R}_n(\boldsymbol{\beta}) - \ell_n(\boldsymbol{\beta})| &= \sup_{\boldsymbol{\beta} \in \mathcal{M}_n} \left| -\frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2 + \frac{2}{n} \mathbf{Z}^\top \boldsymbol{\beta} + \frac{1}{n} \sum_{i=1}^n (\epsilon_i^2 + 2\epsilon_i \theta_i - 2\epsilon_i \mathbf{X}_i^\top \boldsymbol{\beta}) \right| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2 - \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \right| + \left| \frac{2}{n} \sum_{i=1}^n \epsilon_i \theta_i \right| + 2 \sup_{\boldsymbol{\beta} \in \mathcal{M}_n} \left| \left(\frac{1}{n} \mathbf{Z} - \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbf{X}_i \right)^\top \boldsymbol{\beta} \right|. \end{aligned} \quad (2)$$

Each term in (2) can be shown to converge to zero in expectation.

The first term obeys

$$\left(E \left| \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2 - \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \right| \right)^2 \leq E \left[\left\{ \frac{1}{n} \sum_{i=1}^n (\hat{\sigma}_i^2 - \epsilon_i^2) \right\}^2 \right] = \frac{1}{n^2} \sum_{i=1}^n \text{var}(\hat{\sigma}_i^2 - \epsilon_i^2).$$

This is at most

$$\begin{aligned}
& \frac{1}{n^2} \sum_{i=1}^n \{ \text{var}(\hat{\sigma}_i^2) + \text{var}(\epsilon_i^2) + 2|\text{cov}(\hat{\sigma}_i^2, \epsilon_i^2)| \} \\
& \leq \frac{1}{n^2} \sum_{i=1}^n \{ \text{var}(\hat{\sigma}_i^2) + \text{var}(\epsilon_i^2) + 2\text{var}(\hat{\sigma}_i^2)^{1/2} \text{var}(\epsilon_i^2)^{1/2} \} \\
& \leq \frac{1}{n^2} \sum_{i=1}^n \text{var}(\hat{\sigma}_i^2) + \frac{1}{n^2} \sum_{i=1}^n \text{var}(\epsilon_i^2) + 2 \left\{ \frac{1}{n^2} \sum_{i=1}^n \text{var}(\hat{\sigma}_i^2) \right\}^{1/2} \left\{ \frac{1}{n^2} \sum_{i=1}^n \text{var}(\epsilon_i^2) \right\}^{1/2},
\end{aligned}$$

which converges to zero by Assumption 1a because $0 \leq \text{var}(\epsilon_i^2) = \kappa_{4i} - \sigma_i^4 \leq \kappa_{4i}$, where κ_{4i} is the fourth central moment of Y_i . For the second term,

$$\left(E \left| \frac{2}{n} \sum_{i=1}^n \epsilon_i \theta_i \right| \right)^2 \leq E \left\{ \left(\frac{2}{n} \sum_{i=1}^n \epsilon_i \theta_i \right)^2 \right\} = \frac{4}{n^2} \sum_{i=1}^n \text{var}(\epsilon_i \theta_i) = \frac{4}{n^2} \sum_{i=1}^n \sigma_i^2 \theta_i^2,$$

which also converges to zero by Assumption 1a.

To bound the third term in (2), first partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_C^\top, \boldsymbol{\beta}_Y^\top, \boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_q^\top)^\top$ as in model (3). Denote the lengths of $\boldsymbol{\beta}_C$, $\boldsymbol{\beta}_Y$, and $\boldsymbol{\beta}_k$ by p_C , p_Y , and p_k respectively. In other words, $p_C + p_Y + p_1 + \dots + p_q$ equals the total number of covariates p . Finally, given the definition of \mathbf{X}_i and $\mathbf{Z} = \sum_i \hat{\sigma}_i^2 \partial \mathbf{X}_i / \partial Y_i$ in (5),

$$\mathbf{Z} = \left(0, \dots, 0, \sum_{i=1}^n \hat{\sigma}_i^2 C_{i1}, \dots, \sum_{i=1}^n \hat{\sigma}_i^2 C_{ip}, 0, \dots, 0, \dots, 0, \dots, 0 \right)^\top.$$

Then ignoring the constant factor 2, the third term in (2) is upper-bounded by

$$\begin{aligned}
& \sup_{\boldsymbol{\beta} \in \mathcal{M}_n} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbf{C}_{Ci}^\top \boldsymbol{\beta}^C \right| + \left| \frac{1}{n} \sum_{i=1}^n (\hat{\sigma}_i^2 - \epsilon_i Y_i) \mathbf{C}_{Yi}^\top \boldsymbol{\beta}_Y \right| + \sum_{k=1}^q \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i Y_{ik} \mathbf{C}_{ki}^\top \boldsymbol{\beta}_k \right| \right\} \\
& \leq \sum_{j=1}^{p_C} \left| \frac{M_n}{n} \sum_{i=1}^n \epsilon_i C_{Cij} \right| + \sum_{j=1}^{p_Y} \left| \frac{M_n}{n} \sum_{i=1}^n (\hat{\sigma}_i^2 - \epsilon_i Y_i) C_{Yij} \right| + \sum_{k=1}^q \sum_{j=1}^{p_k} \left| \frac{M_n}{n} \sum_{i=1}^n \epsilon_i Y_{ik} C_{kij} \right|, \quad (3)
\end{aligned}$$

where $M_n = \sup_{\beta \in \mathcal{M}_n} \|\beta\|_\infty = o(n^{1/2})$ by assumption.

The first term of (3) can be upper-bounded because

$$\left(E \left| \frac{M_n}{n} \sum_{i=1}^n \epsilon_i C_{Cij} \right| \right)^2 \leq \frac{M_n^2}{n^2} \sum_{i=1}^n \text{var}(\epsilon_i C_{Cij}) = \frac{M_n^2}{n^2} \sum_{i=1}^n \sigma_i^2 C_{Cij}^2,$$

which converges to zero by Assumption 1b and the fact that $M_n^2 = o(n)$. The second term of (3) can be bounded because $E(\hat{\sigma}_i^2 - \epsilon_i Y_i) = 0$ from the proof of Proposition 1, so

$$\begin{aligned} & \left(E \left| \frac{M_n}{n} \sum_{i=1}^n (\hat{\sigma}_i^2 - \epsilon_i Y_i) C_{Yij} \right| \right)^2 \leq \frac{M_n^2}{n^2} \sum_{i=1}^n \text{var}(\hat{\sigma}_i^2 - \epsilon_i Y_i) C_{Yij}^2 \\ & \leq \frac{M_n^2}{n^2} \sum_{i=1}^n \{ \text{var}(\hat{\sigma}_i^2 - \epsilon_i^2) + \text{var}(\epsilon_i \theta_i) + 2|\text{cov}(\hat{\sigma}_i^2 - \epsilon_i^2, \epsilon_i \theta_i)| \} C_{Yij}^2. \end{aligned}$$

Showing that the first two terms above converge to zero under Assumption 1b is similar to controlling the first two terms of (2). The third term above is no more than

$$2 \left\{ \frac{M_n^2}{n^2} \sum_{i=1}^n \text{var}(\hat{\sigma}_i^2 - \epsilon_i^2) C_{Yij}^2 \right\}^{1/2} \left\{ \frac{M_n^2}{n^2} \sum_{i=1}^n \text{var}(\epsilon_i \theta_i) C_{Yij}^2 \right\}^{1/2},$$

which therefore also converges to zero.

The third term of (3) can be upper-bounded because

$$\left(E \left| \frac{M_n}{n} \sum_{i=1}^n \epsilon_i Y_{i_k} C_{kij} \right| \right)^2 \leq \frac{M_n^2}{n^2} \sum_{i=1}^n E(\epsilon_i^2 Y_{i_k}^2 C_{kij}^2) + \frac{M_n^2}{n^2} \sum_{i \neq l} |E(\epsilon_i Y_{i_k} C_{kij} \epsilon_l Y_{l_k} C_{klj})|.$$

Since $i \neq l$ and ϵ_i and Y_{i_k} are independent for all i , $E(\epsilon_i Y_{i_k} C_{kij} \epsilon_l Y_{l_k} C_{klj}) = 0$ unless $i = l_k$ and $l = i_k$. Therefore

$$\left(E \left| \frac{M_n}{n} \sum_{i=1}^n \epsilon_i Y_{i_k} C_{kij} \right| \right)^2 \leq \frac{M_n^2}{n^2} \sum_{i=1}^n \sigma_i^2 (\sigma_{i_k}^2 + \theta_{i_k}^2) C_{kij}^2 + \frac{M_n^2}{n^2} \sum_{i \neq l, i=l_k, l=i_k} |E(\epsilon_i Y_{i_k} C_{kij} \epsilon_l Y_{l_k} C_{klj})|.$$

The first term above converges to zero by Assumption 1c and $M_n^2 = o(n)$. By Cauchy-Schwarz, the second term is at most

$$\left\{ \frac{M_n^2}{n^2} \sum_{i \neq l, i=l_k, l=i_k} E(\epsilon_i^2 Y_{i_k}^2 C_{kij}^2) \right\}^{1/2} \left\{ \frac{M_n^2}{n^2} \sum_{i \neq l, i=l_k, l=i_k} E(\epsilon_l^2 Y_{l_k}^2 C_{klj}^2) \right\}^{1/2}.$$

But for each i there is only one l such that $i_k = l$, so the previous line is at most

$$\frac{M_n^2}{n^2} \sum_{i=1}^n E(\epsilon_i^2 Y_{i_k}^2 C_{kij}^2),$$

which converges to zero by previous arguments.

5 Proof of Theorem 2

Because both $\hat{\beta}_n^M$ and $\hat{\beta}_n^{M\star}$ lie in \mathcal{M}_n ,

$$\begin{aligned} 0 \leq \ell_n(\hat{\beta}_n^M) - \ell_n(\hat{\beta}_n^{M\star}) &= \ell_n(\hat{\beta}_n^M) - \hat{R}_n(\hat{\beta}_n^M) + \hat{R}_n(\hat{\beta}_n^M) - \hat{R}_n(\hat{\beta}_n^{M\star}) + \hat{R}_n(\hat{\beta}_n^{M\star}) - \ell_n(\hat{\beta}_n^{M\star}) \\ &\leq 2 \sup_{\beta \in \mathcal{M}_n} |\ell_n(\beta) - \hat{R}_n(\beta)| + \hat{R}_n(\hat{\beta}_n^M) - \hat{R}_n(\hat{\beta}_n^{M\star}). \end{aligned}$$

By construction, $\hat{R}_n(\hat{\beta}_n^M) \leq \hat{R}_n(\hat{\beta}_n^{M\star})$, so

$$0 \leq E\ell_n(\hat{\beta}_n^M) - \ell_n(\hat{\beta}_n^{M\star}) \leq 2E \sup_{\beta \in \mathcal{M}_n} |\ell_n(\beta) - \hat{R}_n(\beta)|,$$

which converges to zero by Theorem 1.

6 Proof of Theorem 3

Define $\epsilon_i = Y_i - \theta_i$ and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$. Then $\hat{\beta}_n - \hat{\beta}_n^\star = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \epsilon - \mathbf{Z})$. The bound on (3) in the proof of Theorem 1 shows that each component of

$$\frac{1}{n}(\mathbf{X}^\top \epsilon - \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n (\epsilon_i \mathbf{X}_i - \mathbf{Z})$$

has zero mean and variance converging to zero under Assumption 1b. By Chebyshev's inequality, each component therefore converges to zero in probability. It remains to show that each entry of $(\mathbf{X}^\top \mathbf{X}/n)^{-1}$ converges to a constant in probability, as Slutsky's theorem will then imply that $\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^* = o_P(1)$.

The entries of $\mathbf{X}^\top \mathbf{X}/n$ consist of six types of terms:

1. $n^{-1} \sum_{i=1}^n C_{Cij} C_{Cij'}$
2. $n^{-1} \sum_{i=1}^n C_{Cij} C_{Yij'} Y_i$
3. $n^{-1} \sum_{i=1}^n C_{Cij} C_{kij'} Y_{i_k}$
4. $n^{-1} \sum_{i=1}^n C_{Yij} C_{Yij'} Y_i^2$
5. $n^{-1} \sum_{i=1}^n C_{Yij} C_{kij'} Y_i Y_{i_k}$
6. $n^{-1} \sum_{i=1}^n C_{kij} C_{k'ij'} Y_{i_k} Y_{i_{k'}}.$

Since $E(\mathbf{X}^\top \mathbf{X}/n)$ converges by assumption, the expected value of each type of term converges to a constant. The variance of the first type is zero and the variances of the second and third types are given by

$$\begin{aligned} \text{var} \left(\frac{1}{n} \sum_{i=1}^n C_{Cij} C_{Yij'} Y_i \right) &= \frac{1}{n^2} \sum_{i=1}^n C_{Cij}^2 C_{Yij'}^2 \sigma_i^2, \\ \text{var} \left(\frac{1}{n} \sum_{i=1}^n C_{Cij} C_{kij'} Y_{i_k} \right) &= \frac{1}{n^2} \sum_{i=1}^n C_{Cij}^2 C_{kij'}^2 \sigma_{i_k}^2. \end{aligned}$$

The variance of the fourth type equals

$$\begin{aligned}
& \frac{1}{n^2} \sum_{i=1}^n C_{Yij}^2 C_{Yij'}^2 \text{var}(\theta_i^2 + 2\theta_i \epsilon_i + \epsilon_i^2) \\
&= \frac{1}{n^2} \sum_{i=1}^n C_{Yij}^2 C_{Yij'}^2 \{ \text{var}(2\theta_i \epsilon_i) + \text{var}(\epsilon_i^2) + 2\text{cov}(2\theta_i \epsilon_i, \epsilon_i^2) \} \\
&\leq \frac{1}{n^2} \sum_{i=1}^n C_{Yij}^2 C_{Yij'}^2 (4\theta_i^2 \sigma_i^2 + \kappa_{4i} - \sigma_i^4) + \\
&\quad \left\{ \frac{1}{n^2} \sum_{i=1}^n C_{Yij}^2 C_{Yij'}^2 \text{var}(2\theta_i \epsilon_i) \right\}^{1/2} \left\{ \frac{1}{n^2} \sum_{i=1}^n C_{Yij}^2 C_{Yij'}^2 \text{var}(\epsilon_i^2) \right\}^{1/2},
\end{aligned}$$

where κ_{4i} is the fourth central moment of Y_i and the last line is due to the Cauchy-Schwarz inequality. Therefore the variances of the first four types of terms converge to zero under Assumptions 2a–b.

The variance of the fifth type equals

$$\frac{1}{n^2} \sum_{i=1}^n \text{var}(C_{Yij} C_{kij'} Y_i Y_{i_k}) + \frac{1}{n^2} \sum_{i \neq l} \text{cov}(C_{Yij} C_{kij'} Y_i Y_{i_k}, C_{Ylj} C_{klj'} Y_l Y_{l_k}).$$

The second term above is at most

$$\frac{1}{n^2} \sum_{(i,l) \in \mathcal{S}} |\text{cov}(C_{Yij} C_{kij'} Y_i Y_{i_k}, C_{Ylj} C_{klj'} Y_l Y_{l_k})|,$$

where the set \mathcal{S} consists of all pairs (i, l) such that $i \neq l$ and $i = l_k$, $i_k = l$, or $i_k = l_k$. By Cauchy-Schwarz this is upper-bounded by

$$\left\{ \frac{1}{n^2} \sum_{(i,l) \in \mathcal{S}} \text{var}(C_{Yij} C_{kij'} Y_i Y_{i_k}) \right\}^{1/2} \left\{ \frac{1}{n^2} \sum_{(i,l) \in \mathcal{S}} \text{var}(C_{Ylj} C_{klj'} Y_l Y_{l_k}) \right\}^{1/2}.$$

For fixed k , for each i there is at most one l such that $i_k = l_k$. By (11), there are at most D_n indices i such that $i = l_k$ or $i_k = l$. Since $D_n = o(n)$ by assumption, the previous line

is at most

$$\frac{o(n)}{n^2} \sum_{i=1}^n \text{var}(C_{Yij} C_{kij'} Y_i Y_{i_k}) = \frac{o(n)}{n^2} \sum_{i=1}^n C_{ij}^2 C_{ij'}^2 (\theta_i^2 \sigma_{i_k}^2 + \theta_{i_k}^2 \sigma_i^2 + \sigma_i^2 \sigma_{i_k}^2),$$

so

$$\text{var} \left(\frac{1}{n} \sum_{i=1}^n C_{Yij} C_{kij'} Y_i Y_{i_k} \right) \leq \frac{o(n)}{n^2} \sum_{i=1}^n \text{var}(C_{Yij} C_{kij'} Y_i Y_{i_k}) \rightarrow 0$$

under Assumption 2c.

Finally, the variance of the sixth type of term when $k = k'$ takes the same form as the variance of the fourth type of term and so converges to zero. When $k \neq k'$, the variance is

$$\frac{1}{n^2} \sum_{i=1}^n C_{kij}^2 C_{k'ij'}^2 \text{var}(Y_{i_k} Y_{i_{k'}}) = \frac{1}{n^2} \sum_{i=1}^n C_{kij}^2 C_{k'ij'}^2 (\theta_{i_k}^2 \sigma_{i_{k'}}^2 + \theta_{i_{k'}}^2 \sigma_{i_k}^2 + \sigma_{i_k}^2 \sigma_{i_{k'}}^2),$$

which also converges to zero by Assumption 2d.

Chebyshev's inequality implies that $\mathbf{X}^\top \mathbf{X}/n - E(\mathbf{X}^\top \mathbf{X}/n) = o_p(1)$. Since $E(\mathbf{X}^\top \mathbf{X}/n)$ converges to a positive-definite matrix by assumption, by the continuous mapping theorem $(\mathbf{X}^\top \mathbf{X}/n)^{-1}$ converges in probability to the inverse of the limit of $E(\mathbf{X}^\top \mathbf{X}/n)$.

7 Proof of Theorem 4

This proof is similar to the proof of Theorem 3.4 in Rigollet and Hütter (2017). The result is clearly true if $\hat{\beta}_n = \hat{\beta}_n^*$. When $\hat{\beta}_n \neq \hat{\beta}_n^*$, since $\hat{R}_n(\hat{\beta}_n) \leq \hat{R}_n(\beta)$ for any β and

$$\hat{R}_n(\beta) = -\frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2 + \frac{2}{n} \mathbf{Z}^\top \beta + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 + \frac{2}{n} \sum_{i=1}^n \epsilon_i (\theta_i - \mathbf{X}_i^\top \beta) + \ell_n(\beta),$$

it follows that

$$0 \leq \ell_n(\hat{\beta}_n) - \ell_n(\hat{\beta}_n^*) \leq \frac{2}{n} \mathbf{Z}^\top (\hat{\beta}_n^* - \hat{\beta}_n) + \frac{2}{n} \sum_{i=1}^n \epsilon_i \mathbf{X}_i^\top (\hat{\beta}_n - \hat{\beta}_n^*),$$

where $\epsilon_i = Y_i - \theta_i$. Define $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$. Since $\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*) \neq 0$,

$$0 \leq \ell_n(\hat{\boldsymbol{\beta}}_n) - \ell_n(\hat{\boldsymbol{\beta}}_n^*) \leq \frac{2}{n} \mathbf{Z}^\top (\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n) + \frac{2}{n} \boldsymbol{\epsilon}^\top \frac{\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)}{\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)\|_2} \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)\|_2.$$

Young's inequality implies that

$$\frac{2}{n} \boldsymbol{\epsilon}^\top \frac{\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)}{\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)\|_2} \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)\|_2 \leq \frac{2}{n} \left\{ \boldsymbol{\epsilon}^\top \frac{\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)}{\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)\|_2} \right\}^2 + \frac{1}{2n} \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)\|_2^2.$$

Furthermore,

$$\begin{aligned} \ell_n(\hat{\boldsymbol{\beta}}_n) &= \frac{1}{n} \|\boldsymbol{\theta} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^* + \mathbf{X}\hat{\boldsymbol{\beta}}_n^* - \mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2 \\ &= \ell_n(\hat{\boldsymbol{\beta}}_n^*) + \frac{2}{n} (\boldsymbol{\theta} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^*)^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n) + \frac{1}{n} \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n)\|_2^2 \\ &= \ell_n(\hat{\boldsymbol{\beta}}_n^*) + \frac{2}{n} \boldsymbol{\theta}^\top \{\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\} \mathbf{X}(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n) + \frac{1}{n} \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n)\|_2^2 \\ &= \ell_n(\hat{\boldsymbol{\beta}}_n^*) + \frac{1}{n} \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n)\|_2^2. \end{aligned}$$

Therefore

$$0 \leq \ell_n(\hat{\boldsymbol{\beta}}_n) - \ell_n(\hat{\boldsymbol{\beta}}_n^*) \leq \frac{4}{n} \mathbf{Z}^\top (\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n) + \frac{4}{n} \left\{ \boldsymbol{\epsilon}^\top \frac{\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)}{\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*)\|_2} \right\}^2.$$

Since

$$\mathbf{Z} = \left(0, \dots, 0, \sum_{i=1}^n \hat{\sigma}_i^2 \mathbf{C}_{Yi}^\top, 0, \dots, 0, \dots, 0, \dots, 0 \right)^\top$$

and by assumption $n^{-1}E\mathbf{Z}$ converges to a constant, Assumption 1b combined with Chebyshev's inequality show that $n^{-1}\{\mathbf{Z} - E(\mathbf{Z})\} = o_P(1)$. Since $\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^* = o_P(1)$ by Theorem 3, the first term above is $o_P(1)$.

To show that the second term above is $o_P(1)$, let $\boldsymbol{\Phi}$ be a $n \times p$ matrix whose columns constitute an orthonormal basis of the column space of \mathbf{X} , as in the proof of Theorem 2.2

of Rigollet and Hütter (2017). Then there exists a $\mathbf{v} \in \mathbb{R}^p$ such that $\Phi \mathbf{v} = \mathbf{X}(\hat{\beta}_n - \hat{\beta}_n^*)$. Therefore

$$\frac{1}{n} \left\{ \epsilon^\top \frac{\mathbf{X}(\hat{\beta}_n - \hat{\beta}_n^*)}{\|\mathbf{X}(\hat{\beta}_n - \hat{\beta}_n^*)\|_2} \right\}^2 = \frac{1}{n} \left(\epsilon^\top \frac{\Phi \mathbf{v}}{\|\Phi \mathbf{v}\|_2} \right)^2 = \frac{1}{n} \left(\epsilon^\top \Phi \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right)^2 \leq \left(\frac{1}{n^{1/2}} \sup_{\mathbf{u} \in \mathcal{B}_2} |\epsilon^\top \Phi \mathbf{u}| \right)^2,$$

where \mathcal{B}_2 denotes the closed unit ball in \mathbb{R}^p . By the arguments in the proof of Theorem 1.19 of Rigollet and Hütter (2017),

$$P \left(\sup_{\mathbf{u} \in \mathcal{B}_2} |\epsilon^\top \Phi \mathbf{u}| \geq n^{1/2} t^{1/2} \right) \leq P \left(2 \sup_{\mathbf{u} \in \mathcal{N}} |\epsilon^\top \Phi \mathbf{u}| \geq n^{1/2} t^{1/2} \right)$$

for any $t > 0$, where \mathcal{N} is an $1/2$ -net of \mathcal{B}_2 . For each $\mathbf{u} \in \mathcal{N}$, by Markov's inequality

$$P(|\epsilon^\top \Phi \mathbf{u}| \geq n^{1/2} t^{1/2} / 2) \leq \frac{4E\{(\epsilon^\top \Phi \mathbf{u})^2\}}{nt}.$$

Let $\mathbf{c} = (c_1, \dots, c_n)$, where c_j is the j th coordinate of $\Phi \mathbf{u}$. Then $\|\mathbf{c}\|_2 = \|\Phi \mathbf{u}\|_2 = 1$, so

$$E\{(\epsilon^\top \Phi \mathbf{u})^2\} = E(\mathbf{c}^\top \epsilon \epsilon^\top \mathbf{c}) \leq \max_i \sigma_i^2.$$

Since the $1/2$ -net \mathcal{N} has cardinality at most 6^p by Lemma 1.18 of Rigollet and Hütter (2017),

$$P \left(\sup_{\mathbf{u} \in \mathcal{B}_2} |\epsilon^\top \Phi \mathbf{u}| \geq n^{1/2} t^{1/2} \right) \leq 6^p \frac{4 \max_i \sigma_i^2}{nt},$$

which goes to zero by Assumption 2. Therefore

$$P \left[\frac{1}{n} \left\{ \epsilon^\top \frac{\mathbf{X}(\hat{\beta}_n - \hat{\beta}_n^*)}{\|\mathbf{X}(\hat{\beta}_n - \hat{\beta}_n^*)\|_2} \right\}^2 > t \right] \rightarrow 0$$

for every $t > 0$, which implies that $0 \leq \ell_n(\hat{\beta}_n) - \ell_n(\hat{\beta}_n^*) \leq o_P(1)$.

8 Proof of Theorem 5

Define

$$\mathbf{V}_i = \frac{1}{n^{1/2}} \left(\epsilon_i \mathbf{X}_i - \hat{\sigma}_i^2 \frac{\partial \mathbf{X}_i}{\partial Y_i} \right).$$

Since $\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^* = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \boldsymbol{\epsilon} - \mathbf{Z})$,

$$n^{1/2} \hat{\boldsymbol{\Sigma}}_n^{-1/2} (\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*) = \hat{\boldsymbol{\Sigma}}_n^{-1/2} (\mathbf{X}^\top \mathbf{X} / n)^{-1} \sum_{i=1}^n \mathbf{V}_i.$$

In the proof of Theorem 3, it was shown that $\mathbf{X}^\top \mathbf{X} / n$ converges to a positive-definite matrix in probability. It remains to show that $\sum_i \mathbf{V}_i$ is asymptotically normal.

The standard Lindeberg-Feller central limit theorem cannot be applied because the \mathbf{V}_i are not independent. Specifically, \mathbf{V}_i and \mathbf{V}_l are dependent when $i \in \mathcal{N}(l)$ or $l \in \mathcal{N}(i)$, due to the inclusion of Y_{i_k} in \mathbf{X}_i . On the other hand, the assumption that D_n (11) is $O(1)$ guarantees that this dependence is sufficiently local such that a central limit theorem still holds. Let Γ be a graph with n vertices where an edge exists between vertex i and vertex l if and only if \mathbf{V}_i and \mathbf{V}_l are dependent. Then Γ is a dependency graph for the \mathbf{V}_i with maximum degree D_n , and the local dependence central limit theorem in Corollary 1 of Raič (2004) implies that $\sum_i \mathbf{V}_n^{-1/2} \mathbf{V}_i \rightarrow N(0, \mathbf{I}_p)$ as long as

$$\lim_{n \rightarrow \infty} (1 + D_n) \sum_{i=1}^n E \left\{ \|\mathbf{V}_n^{-1/2} \mathbf{V}_i\|_2^2 I \left(\|\mathbf{V}_n^{-1/2} \mathbf{V}_i\|_2^2 > \frac{\eta}{\mathbf{V}_n} \right) \right\} = 0$$

for every $\eta > 0$, where

$$\mathbf{V}_n = (1 + D_n)^2 \sum_{i=1}^n E \|\mathbf{V}_n^{-1/2} \mathbf{V}_i\|_2^2.$$

To establish this last condition, first bound

$$\begin{aligned}\|\mathbf{V}_n^{-1/2}\mathbf{V}_i\|_2^2 &\leq \lambda_1(\mathbf{V}_n)^{-1}\|\mathbf{V}_i\|_2^2 \\ &= \lambda_1(\mathbf{V}_n)^{-1}\frac{1}{n}\sum_{j=1}^p\left\{C_{Cij}^2\epsilon_i^2 + C_{Yij}^2(\epsilon_i Y_i - \hat{\sigma}_i^2)^2 + \sum_{k=1}^q C_{kij}^2\epsilon_i^2 Y_{i_k}^2\right\},\end{aligned}$$

where $\lambda_1(\mathbf{V}_n)$ is the minimum eigenvalue of \mathbf{V}_n . Therefore

$$\mathbf{V}_n \leq \frac{(1 + D_n)^2}{n\lambda_1(\mathbf{V}_n)}\sum_{i=1}^n\sum_{j=1}^p\left\{C_{Cij}^2\sigma_i^2 + C_{Yij}^2\text{var}(\epsilon_i Y_i - \hat{\sigma}_i^2) + \sum_{k=1}^q C_{kij}^2\sigma_i^2(\theta_{i_k}^2 + \sigma_{i_k}^2)\right\}$$

As in the proof of Theorem 1, two applications of Cauchy-Schwarz give

$$\begin{aligned}\sum_{i=1}^n C_{Yij}^2\text{var}(\epsilon_i Y_i - \hat{\sigma}_i^2) &\leq \sum_{i=1}^n C_{Yij}^2\text{var}(\epsilon_i \theta_i) + \sum_{i=1}^n C_{Yij}^2\text{var}(\epsilon_i^2 - \hat{\sigma}_i^2) + \\ &\quad 2\left\{\sum_{i=1}^n C_{Yij}^2\text{var}(\epsilon_i \theta_i)\right\}^{1/2}\left\{\sum_{i=1}^n C_{Yij}^2\text{var}(\epsilon_i^2 - \hat{\sigma}_i^2)\right\}^{1/2}.\end{aligned}$$

This, combined with Assumption 1 and the assumption that $\lim_n \lambda_1(\mathbf{V}_n) > 0$, implies

$$\mathbf{V}_n \leq \frac{(1 + O(1))^2}{n\lambda_1(\mathbf{V}_n)}O(n) = O(1).$$

This in turn implies that for any $\tau > 0$,

$$\begin{aligned}&P\left\{I\left(\|\mathbf{V}_n^{-1/2}\mathbf{V}_i\|_2^2 > \frac{\eta}{\mathbf{V}_n}\right) > \tau\right\} \leq \frac{1}{\tau}P\left(\|\mathbf{V}_n^{-1/2}\mathbf{V}_i\|_2^2 > \frac{\eta}{\mathbf{V}_n}\right) \\ &\leq \frac{1}{\tau}P\left\{\lambda_1(\mathbf{V}_n)^{-1}\|\mathbf{V}_i\|_2^2 > \frac{\eta}{\mathbf{V}_n}\right\} \leq \frac{\mathbf{V}_n}{\eta\tau\lambda_1(\mathbf{V}_n)}E\|\mathbf{V}_i\|^2 \\ &= \frac{\mathbf{V}_n}{\eta\tau\lambda_1(\mathbf{V}_n)}\frac{1}{n}\sum_{j=1}^p\left\{C_{Cij}^2\sigma_i^2 + C_{Yij}^2\text{var}(\epsilon_i Y_i - \hat{\sigma}_i^2) + \sum_{k=1}^q C_{kij}^2\sigma_i^2\theta_{i_k}^2\right\} \rightarrow 0.\end{aligned}$$

Therefore

$$n\|\mathbf{V}_n^{-1/2}\mathbf{V}_i\|_2^2 I\left(\|\mathbf{V}_n^{-1/2}\mathbf{V}_i\|_2^2 > \frac{\eta}{\mathbf{V}_n}\right)$$

converges to zero in probability. By the dominated convergence theorem, this is also true in expectation, so that

$$(1 + D_n) \sum_{i=1}^n E \left\{ \|\mathbf{V}_n^{-1/2} \mathbf{V}_i\|_2^2 I \left(\|\mathbf{V}_n^{-1/2} \mathbf{V}_i\|_2^2 > \frac{\eta}{\mathbf{V}_n} \right) \right\} = \{1 + O(1)\} \frac{1}{n} \sum_{i=1}^n o(1) = o(1).$$

9 Proof of Proposition 2

Define

$$\mathbf{V}_i = \frac{1}{n^{1/2}} \left(\epsilon_i \mathbf{X}_i - \hat{\sigma}_i^2 \frac{\partial \mathbf{X}_i}{\partial Y_i} \right).$$

so that

$$\mathbf{V}_n = \sum_{i=1}^n \text{var}(\mathbf{V}_i) + \sum_{i \neq l} \text{cov}(\mathbf{V}_i, \mathbf{V}_l).$$

If σ_i^2 , κ_{3i} , and κ_{4i} are known, the variance term can be written as the $p \times p$ block matrix

$$\sum_{i=1}^n \text{var}(\mathbf{V}_i) = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} & \mathbf{I} \end{pmatrix},$$

where the jj' th entries of the $p \times p$ blocks \mathbf{A} , $\mathbf{B} = \mathbf{D}^\top$, and \mathbf{E} are

$$\begin{aligned} A_{jj'} &= \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} E(\epsilon_i^2) = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} \sigma_i^2 \\ B_{jj'} &= D_{j'j} = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} E\{\epsilon_i(\epsilon_i Y_i - \sigma_i^2)\} = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} (\sigma_i^2 \theta_i + \kappa_{3i}) \\ E_{jj'} &= \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} E\{(\epsilon_i Y_i - \sigma_i^2)^2\} = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} (\sigma_i^2 \theta_i^2 + 2\theta_i \kappa_{3i} + \kappa_{4i} - \sigma_i^4) \end{aligned}$$

for $j, j' = 1, \dots, p$. The matrices $\mathbf{C} = \mathbf{G}^\top$ and $\mathbf{F} = \mathbf{H}^\top$ are $p \times pq$ -dimensional, with rows indexed by j and columns indexed by pairs (j', k) . Their entries take the form

$$C_{j,(j',k)} = G_{(j',k),j} = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} E(\epsilon_i^2 Y_{i_k}) = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} \sigma_i^2 \theta_{i_k}$$

$$F_{j,(j',k)} = H_{(j',k),j} = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} E\{(\epsilon_i Y_i - \sigma_i^2) \epsilon_i Y_{i_k}\} = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} \theta_{i_k} (\sigma_i^2 \theta_i + \kappa_{3i})$$

for $j = 1, \dots, p$ and columns indexed by pairs (j', k) for $j' = 1, \dots, p$ and $k = 1, \dots, q$. Finally, the matrix \mathbf{I} is $pq \times pq$ -dimensional with rows and columns indexed by pairs (j, k) and (j', k') with entries

$$I_{(j,k),(j',k')} = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} E(\epsilon_i^2 Y_{i_k}^2) = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} \sigma_i^2 (\theta_{i_k}^2 + \sigma_{i_k}^2) \text{ if } k = k',$$

$$\frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} E(\epsilon_i^2 Y_{i_k} Y_{i_{k'}}) = \frac{1}{n} \sum_{i=1}^n C_{ij} C_{ij'} \sigma_i^2 \theta_{i_k} \theta_{i_{k'}} \text{ if } k \neq k'$$

The covariance term can be written as the $p \times p$ block matrix

$$\sum_{i \neq l} \text{cov}(\mathbf{V}_i, \mathbf{V}_l) = \begin{pmatrix} \mathbf{A}' & \mathbf{B}' & \mathbf{C}' \\ \mathbf{D}' & \mathbf{E}' & \mathbf{F}' \\ \mathbf{G}' & \mathbf{H}' & \mathbf{I}' \end{pmatrix},$$

where the entries of the blocks are

$$\begin{aligned}
A'_{jj'} &= \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \text{cov}(\epsilon_i, \epsilon_l) = 0, \\
B'_{jj'} &= D'_{j'j} = \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \text{cov}(\epsilon_i, \epsilon_l Y_l - \sigma_l^2) = 0, \\
C'_{j,(j',k)} &= G'_{(j',k),j} = \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \text{cov}(\epsilon_i, \epsilon_l Y_{l_k}) = \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} E(\epsilon_i Y_{l_k}) E(\epsilon_l) = 0, \\
E'_{jj'} &= \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \text{cov}(\epsilon_i Y_i - \sigma_i^2, \epsilon_l Y_l - \sigma_l^2) = 0, \\
F'_{j,(j',k)} &= H'_{(j',k),j} = \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \text{cov}(\epsilon_i Y_i - \sigma_i^2, \epsilon_l Y_{l_k}) = 0, \\
I'_{(j,k),(j',k')} &= \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \text{cov}(\epsilon_i Y_{i_k}, \epsilon_l Y_{l_{k'}}) = \frac{1}{n} \sum_{i \neq l} C_{ij} C_{lj'} \sigma_i^2 \sigma_l^2 \delta_{il_{k'}} \delta_{i_k l},
\end{aligned}$$

where $\delta_{ab} = 1$ if $a = b$ and 0 otherwise.

In each of the above terms, replacing θ_i with Y_i and $\theta_i^2 + \sigma_i^2$ with Y_i^2 gives an unbiased estimate.

References

- W. D. Biscarri. *Statistical methods for binomial and Gaussian sequences*. PhD thesis, University of Illinois at Urbana-Champaign, 2019.
- E. Greenshtein, A. Mantzura, Y. Ritov, et al. Nonparametric empirical bayes improvement of shrinkage estimators with applications to time series. *Bernoulli*, 25(4B):3459–3478, 2019.
- W. James and C. M. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth*

- Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 367–379. Berkeley and Los Angeles, University of California Press, 1961.
- S. Kou and J. J. Yang. Optimal shrinkage estimation in heteroscedastic hierarchical linear models. In *Big and Complex Data Analysis*, pages 249–284. Springer, 2017.
- S. D. Oman. Contracting towards subspaces when estimating the mean of a multivariate normal distribution. *Journal of Multivariate Analysis*, 12(2):270–290, 1982.
- M. Raič. A multivariate CLT for decomposable random vectors with finite second moments. *Journal of Theoretical Probability*, 17(3):573–603, 2004.
- P. Rigollet and J.-C. Hütter. High-dimensional statistics. Lecture notes, Massachusetts Institute of Technology, 2017. <http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf>.
- X. Xie, S. Kou, and L. D. Brown. SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.
- X. Xie, S. C. Kou, and L. Brown. Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *The Annals of Statistics*, 44(2):564, 2016.