

Supplementary Material for Expandable Factor Analysis

BY SANVESH SRIVASTAVA

Department of Statistics and Actuarial Science, University of Iowa, 241 Schaeffer Hall, 20 East Washington Street, Iowa City, Iowa 52242, U.S.A.

sanvesh-srivastava@uiowa.edu

BARBARA E. ENGELHARDT

Department of Computer Science, Center for Statistics and Machine Learning, Princeton University, 35 Olden Street, Princeton, New Jersey 08540, U.S.A.

bee@princeton.edu

AND DAVID B. DUNSON

Department of Statistical Science, Duke University, Box 90251, Durham, North Carolina 27708, U.S.A.

dunson@duke.edu

1. EXPECTATION-MAXIMIZATION ALGORITHM FOR EXPANDABLE FACTOR ANALYSIS

1.1. Estimation of Λ and Σ

Define the following quantities using mean-centered data:

$$S_{yy} = \frac{1}{n} \sum_{i=1}^n y_i y_i^T, \quad S_{zz} = \frac{1}{n} \sum_{i=1}^n z_i z_i^T, \quad S_{yz} = \frac{1}{n} \sum_{i=1}^n y_i z_i^T, \quad \Omega = \Lambda \Lambda^T + \Sigma,$$

$$\Delta = I_k - \Lambda^T \Omega^{-1} \Lambda, \quad G = \Omega^{-1} \Lambda, \quad F = \Delta + G^T S_{yy} G, \quad L = S_{yy} G,$$

where I_k is the $k \times k$ identity matrix. We place Jeffreys' prior on the error variances, $\pi(\sigma_d) \propto \sigma_d^{-1}$ ($d = 1, \dots, p$). Let $\Lambda^{(t)}$ and $\Sigma^{(t)}$ be the estimates of Λ and Σ at iteration t , then the conditional expectations of S_{zz} , S_{yz} , and complete data log likelihood at iteration $(t + 1)$ are

$$E(S_{zz} | Y, \Lambda^{(t)}, \Sigma^{(t)}) = \Delta^{(t)} + G^{(t)T} S_{yy} G^{(t)} = F^{(t)}, \quad E(S_{yz} | Y, \Lambda^{(t)}, \Sigma^{(t)}) = L^{(t)},$$

$$Q(\Lambda, \Sigma | \Lambda^{(t)}, \Sigma^{(t)}) = E\{(npk)^{-1} \log p(Z, \Lambda, \Sigma | Y, \Lambda^{(t)}, \Sigma^{(t)}, \alpha_{1:k}, \eta_{1:k})\}$$

$$= - \sum_{d=1}^p \left[\frac{1}{2pk} \frac{(S_{yy})_{dd} + \{\Lambda E(S_{zz} | Y, \Lambda^{(t)}, \Sigma^{(t)}) \Lambda^T\}_{dd}}{\sigma_d^2} - \frac{1}{2pk} \frac{2\{E(S_{yz} | Y, \Lambda^{(t)}, \Sigma^{(t)}) \Lambda^T\}_{dd}}{\sigma_d^2} \right]$$

$$- \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{npk} \log \left(1 + \frac{|\lambda_{dj}|}{\eta_j} \right) - \frac{n+2}{2npk} \sum_{d=1}^p \log \sigma_d^2$$

$$\equiv - \sum_{d=1}^p \log p_{\text{mis}}(\lambda_d, \sigma_d^2 | S_{yy}, F^{(t)}, L^{(t)}) - \frac{n+2}{2npk} \sum_{d=1}^p \log \sigma_d^2, \quad (1)$$

where the superscript (t) denotes the dependence on $\Lambda^{(t)}$ and $\Sigma^{(t)}$. The objective (1) splits into p separate terms, and term d depends on λ_d and σ_d^2 ; therefore, (1) is maximized by repeating the following two until steps until convergence to a fixed point:

1. For $d = 1, \dots, p$,
 - a. fix σ_d^2 at $\sigma_d^{2(t)}$ in

$$\log p_{\text{mis}}(\lambda_d, \sigma_d^2 \mid S_{yy}, F^{(t)}, L^{(t)}) + \frac{n+2}{2npk} \log \sigma_d^2, \quad (2)$$

and minimize with respect to λ_d to estimate $\lambda_d^{(t+1)}$;

- b. fix λ_d at $\lambda_d^{(t+1)}$ in (2) and minimize (2) with respect to σ_d^2 to estimate $\sigma_d^{2(t+1)}$.
2. Increment t to $(t+1)$.

1.2. Block coordinate descent algorithm for estimation of Λ

We use local linear approximation of the objective (2) to derive a new block coordinate descent algorithm. We suppress the superscript (t) in w_d and X to ease notation. The algorithm initializes $\tilde{\Lambda}^0$ at $\Lambda^{\text{lla}^{(t)}}$ and updates $\tilde{\Lambda}_{dj}^{(i)}$ using (2) as

$$\tilde{\lambda}_{dj}^{(i+1)} = \underset{\lambda_{dj}}{\operatorname{argmin}} \frac{\lambda_{dj}^2 X_j^T X_j + 2\lambda_{dj} (\tilde{\Lambda}_{d,(-j)}^{(i)T} X_{(-j)}^T X_j - X_j^T w_d)}{2} + \frac{(\alpha_j + 1)\sigma_d^{2(t)} |\lambda_{dj}|}{(\eta_j + |\lambda_{dj}^{(t)}|)n}$$

successively for $j = 1, \dots, k$ in the $(i+1)$ th cycle. This objective function is convex and its optimum is

$$\tilde{\lambda}_{dj}^{(i+1)} = \frac{\operatorname{sign}(s_{dj}^{(i)})}{f_{jj}} \left(|s_{dj}^{(i)}| - \frac{(\alpha_j + 1)\sigma_d^{2(t)}}{(\eta_j + |\lambda_{dj}^{(t)}|)n} \right)_+, \quad (3)$$

where $s_{dj}^{(i)} = X_j^T w_d - \tilde{\Lambda}_{d,(-j)}^{(i)T} X_{(-j)}^T X_j$ and $f_{jj} = X_j^T X_j$. We also exploit the form of (3) and use it to update the k th column of $\tilde{\Lambda}^{(i)}$. This leads to k block updates for $\tilde{\Lambda}^{(i)}$ in a single cycle of the coordinate descent algorithm. These updates are repeated until the change in $\tilde{\Lambda}$ is negligible. We then set $\Lambda^{\text{lla}^{(t+1)}} = \tilde{\Lambda}^{(\infty)}$. We have implemented this algorithm in R (R Development Core Team, 2016) using the glmnet package (Friedman et al., 2010).

1.3. Root- n consistent estimates of Λ and Σ

Let S_{yy} be the empirical covariance matrix of mean-centered data and $\hat{\zeta}_d$ and $\hat{\psi}_d$ ($d = 1, \dots, p$) be its eigenvalues and eigenvectors, then

$$S_{yy} = Y^T Y / n = \sum_{d=1}^p \hat{\zeta}_d \hat{\psi}_d \hat{\psi}_d^T \quad (4)$$

is the eigen decomposition of S_{yy} . Use (4) to define

$$\lambda_{dj}^0 = \hat{\zeta}_j^{1/2} \hat{\psi}_{dj} \quad (d = 1, \dots, p; j = 1, \dots, k).$$

An application of Theorem 2 in Kneip & Sarda (2011) shows that $\lambda_{dj}^0/p^{1/2}$ is a root- n consistent estimator of $\lambda_{dj}/p^{1/2}$ when $n \leq p$. Equations 4.3 and 4.4 in Kneip & Sarda (2011) and Assumptions A.1–A.4 in the main paper imply that there exist universal positive constants D_0, D_1 , and

C_0 such that

$$\frac{\lambda_{dj}^2}{p} \leq \frac{D_0 - D_1}{p}, \quad \frac{\lambda_{dj}^{0^2}}{p} \leq \frac{D_0 + C_0 (\log p/n)^{1/2}}{p}$$

with probability at least $A(n, p) = 1 - 8p^{2-C_0/2} \rightarrow 1$ as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$. This implies that

$$\left| \frac{\lambda_{dj}}{p^{1/2}} - \frac{\lambda_{dj}^0}{p^{1/2}} \right| \leq \left(\frac{D_0 - D_1}{p} \right)^{1/2} + \left\{ \frac{D_0 + C_0 (\log p/n)^{1/2}}{p} \right\}^{1/2} \quad (5)$$

with probability at least $A(n, p)$. Since $\log p/n \rightarrow 0$, $\log p/n \leq D_0^2/C_0^2$ for large n and p and (5) reduces to

$$\left| \frac{\lambda_{dj}^0}{p^{1/2}} - \frac{\lambda_{dj}}{p^{1/2}} \right| \leq \left(\frac{2D_0}{p} \right)^{1/2} \leq \left(\frac{2D_0}{n} \right)^{1/2}$$

with probability at least $A(n, p) \rightarrow 1$ as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$. This shows that $\lambda_{dj}^0/p^{1/2}$ is a root- n consistent estimator of $\lambda_{dj}/p^{1/2}$. Theorem 3 in Kneip & Sarda (2011) implies that $\sigma_d^{2^0} = (S_{yy} - \Lambda^0 \Lambda^{0T})_{dd}$ is a root- n consistent estimator of σ_d^2 for overfitted factor models.

We also prove a result that is used in the proof for asymptotic normality of nonzero loadings.

LEMMA 1. *If Assumptions A.0–A.4 in the main paper hold, then $E(\lambda_{dj}^{0^2}) < \infty$ ($d = 1, \dots, p$; $j = 1, \dots, k$).*

Proof. Using (4),

$$E(\lambda_{dj}^{0^2}) = E(\widehat{\zeta}_j^2 \widehat{\psi}_{dj}^2) \stackrel{(i)}{\leq} E(\widehat{\zeta}_j^2) = \int_0^\infty \text{pr}(\widehat{\zeta}_j^2 > t) dt \leq (D_2 + D_0)^2 + \int_{(D_2+D_0)^2}^\infty \text{pr}(\widehat{\zeta}_j^2 > t) dt, \quad (6)$$

where (i) follows because $\sum_{d=1}^p \widehat{\psi}_{dj}^2 = 1$. Equation 4.1 of Theorem 2 in Kneip & Sarda (2011) implies that for some $\zeta_j \geq 0$,

$$\begin{aligned} 8/p^{C_0/2-2} &\geq \text{pr} \left\{ |\widehat{\zeta}_j/p - \zeta_j/p| > D_2/p + C_0(\log p/n)^{1/2} \right\} \\ &\stackrel{(ii)}{\geq} \text{pr} \left\{ |\widehat{\zeta}_j/p - \zeta_j/p| > (C_0 D_2/D_0 + C_0)(\log p/n)^{1/2} \right\} \\ &\geq \text{pr} \left\{ \widehat{\zeta}_j/p > \zeta_j/p + (C_0 D_2/D_0 + C_0)(\log p/n)^{1/2} \right\} \\ &\geq \text{pr} \left\{ \widehat{\zeta}_j/p > (C_0 D_2/D_0 + C_0)(\log p/n)^{1/2} \right\} \\ &= \text{pr} \left\{ \widehat{\zeta}_j^2 > (C_0 D_2/D_0 + C_0)^2 p^2 \log p/n \right\}, \end{aligned}$$

where (ii) follows because $C_0(\log p/n)^{1/2} > D_0/p$ by Assumption A.4 in the main paper. Substituting $t = (C_0 D_2/D_0 + C_0)^2 p^2 \log p/n$ in (6) implies that

$$\text{pr}(\widehat{\zeta}_j^2 > t) \leq 8(C_0 D_2/D_0 + C_0)^{C_0/2-2} (\log p/n)^{C_0/2-2} t^{1-C_0/4}, \quad t \geq (D_0 + D_2)^2.$$

Therefore, $\int_{(D_2+D_0)^2}^{\infty} \text{pr}(\widehat{\zeta}_j^2 > t) dt < \infty$ for $C_0 \in (8, \infty)$, which in turn shows that $E(\lambda_{dj}^{0^2})$ is bounded because $\log p/n \rightarrow 0$.

1.4. Computational complexity

The computational complexity of the estimation algorithm equals the cost of performing p penalized regression problems of dimension $k = O(\log p)$. Our estimation algorithm requires $O(np^2 + p \log^2 p)$ time upfront to calculate S_{yy} and its eigen decomposition. Estimation of G, Δ, F , and L in (1) involves k -dimensional matrix multiplications and inversions of $O(\log^3 p)$ time complexity. Using these matrices, one iteration of the block coordinate descent algorithm has $O(\log p)$ time complexity for dimension d ($d = 1, \dots, p$). The total time complexity of each iteration is $O(p \log p + \log^3 p)$; therefore, the time complexity of T iterations of the expectation-maximization algorithm is $O(Tp \log p)$.

2. PROPERTIES OF THE MULTISCALE GENERALIZED DOUBLE PARETO PRIOR

2.1. Proof of Lemma 1

If \mathcal{C} is the support of multiscale generalized double Pareto prior on Λ , then

$$\text{pr}(\mathcal{C}) = \text{pr} \left(\Lambda \mid \max_{1 \leq d \leq p} \sum_{k=1}^{\infty} \lambda_{dk}^2 < \infty \right) \geq 1 - \lim_{t \uparrow \infty} \sum_{d=1}^p \text{pr} \left(\Lambda \mid \sum_{k=1}^{\infty} \lambda_{dk}^2 \geq t \right) \geq 1 - p \lim_{t \uparrow \infty} \frac{\sum_{k=1}^{\infty} V(\lambda_{1k})}{t}.$$

Since λ_{1k} follows generalized double Pareto distribution with parameters (α_k, η_k) , $V(\lambda_{1k}) = 2\eta_k^2(\alpha_k - 1)^{-1}(\alpha_k - 2)^{-1}$ for $\alpha_k > 2$ and

$$\sum_{k=1}^{\infty} V(\lambda_{1k}) \leq 2 \sum_{k=1}^{\infty} \frac{\eta_k^2}{\alpha_k^2} \left(1 - \frac{2}{\alpha_k}\right)^{-2} \leq \{2 + O(1)\} \sum_{k=1}^{\infty} \frac{\eta_k^2}{\alpha_k^2}. \quad (7)$$

This summation is finite if $\alpha_k > 2$ and $\eta_k/\alpha_k = O(k^{-m})$ for $m > 0.5$; therefore, $\text{pr}(\mathcal{C}) = 1$.

2.2. Proof of Lemma 2

Let $k(p, \delta, \epsilon)$ be such that $\text{pr}\{\Omega^k \mid d_{\infty}(\Omega, \Omega^k) \geq \epsilon\} \leq \epsilon$ for any $\epsilon > 0$. Then,

$$\text{pr}\{d_{\infty}(\Omega, \Omega^k) \geq \epsilon\} \stackrel{(i)}{\leq} \sum_{i=1}^p \sum_{j=1}^p \text{pr}(|\Omega_{ij} - \Omega_{ij}^k| \leq \epsilon) \stackrel{(ii)}{\leq} \frac{p^2}{\epsilon} \sum_{l=k+1}^{\infty} E(|\lambda_{1l}|^2),$$

where (i) follows from the union bound and (ii) follows from Markov's inequality and the independence of λ_{ik} s. The assumptions in Lemma 2 of the main paper and (7) imply that

$$\frac{p^2}{\epsilon} \sum_{l=k+1}^{\infty} E(|\lambda_{1l}|^2) = \text{constant} \frac{p^2}{\epsilon} \delta^{-2k} \leq \epsilon \implies k = O(\log^{-1} \delta \log \frac{p}{\epsilon}).$$

3. THEORETICAL PROPERTIES OF Λ^{lla} AND Σ^{lla}

3.1. Proof of Theorem 1

Let $\theta = (\Lambda, \Sigma)$. Then, the objective function in (1) is

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{ML}}(\theta) - \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{npk} \log \left(1 + \frac{|\lambda_{dj}|}{\eta_j} \right) - \frac{n+2}{2npk} \sum_{d=1}^p \log \sigma_d^2, \quad (8)$$

where $\mathcal{L}_{\text{ML}}(\theta)$ is the log likelihood of θ scaled by npk . This leads to the Q -function

$$Q(\theta | \theta^{(t)}) = - \sum_{d=1}^p \log p_{\text{mis}}(\lambda_d, \sigma_d^2 | S_{yy}, F^{(t)}, L^{(t)}) - \frac{n+2}{2npk} \sum_{d=1}^p \log \sigma_d^2. \quad (9)$$

The local linear approximation of (9) is

$$\begin{aligned} Q_{\text{LLA}}(\theta | \theta^{(t)}) &= - \sum_{d=1}^p \frac{(S_{yy})_{dd} + (\Lambda F^{(t)} \Lambda^T)_{dd} - 2(L^{(t)} \Lambda^T)_{dd}}{2pk\sigma_d^2} - \frac{n+2}{2npk} \sum_{d=1}^p \log \sigma_d^2 \\ &\quad - \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{npk} \left\{ \log \left(1 + \frac{|\lambda_{dj}^{(t)}|}{\eta_j} \right) + \frac{\text{sign}(\lambda_{dj}^{(t)})}{\eta_j + |\lambda_{dj}^{(t)}|} (\lambda_{dj} - \lambda_{dj}^{(t)}) \right\} \\ &= Q_{\text{ML}}(\theta | \theta^{(t)}) - \frac{n+2}{2npk} \sum_{d=1}^p \log \sigma_d^2 \\ &\quad - \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{npk} \left\{ \log \left(1 + \frac{|\lambda_{dj}^{(t)}|}{\eta_j} \right) + \frac{\text{sign}(\lambda_{dj}^{(t)})}{\eta_j + |\lambda_{dj}^{(t)}|} (\lambda_{dj} - \lambda_{dj}^{(t)}) \right\}, \quad (10) \end{aligned}$$

where $Q_{\text{ML}}(\theta | \theta^{(t)})$ is the Q -function that corresponds to $\mathcal{L}_{\text{ML}}(\theta)$. Theorem 1 of Dempster et al. (1977) shows that $Q_{\text{ML}}(\theta^{(t)} | \theta^{(t)}) = \mathcal{L}_{\text{ML}}(\theta^{(t)})$, and using this in (8) and (10) shows that $Q(\theta^{(t)} | \theta^{(t)}) = \mathcal{L}(\theta^{(t)})$ and $Q_{\text{LLA}}(\theta^{(t)} | \theta^{(t)}) = \mathcal{L}(\theta^{(t)})$. Subtracting (10) from (8) yields

$$\mathcal{L}(\theta) - Q_{\text{LLA}}(\theta | \theta^{(t)}) = \mathcal{L}_{\text{ML}}(\theta) - Q_{\text{ML}}(\theta | \theta^{(t)}) + \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{npk} l_{dj}(\lambda_{dj} | \lambda_{dj}^{(t)}), \quad (11)$$

where

$$l_{dj}(\lambda_{dj} | \lambda_{dj}^{(t)}) = \log \left(1 + \frac{|\lambda_{dj}^{(t)}|}{\eta_j} \right) + \frac{\text{sign}(\lambda_{dj}^{(t)})}{\eta_j + |\lambda_{dj}^{(t)}|} (\lambda_{dj} - \lambda_{dj}^{(t)}) - \log \left(1 + \frac{|\lambda_{dj}|}{\eta_j} \right). \quad (12)$$

The log function is concave and is majorized by its tangent, so $l_{dj}(\lambda_{dj} | \lambda_{dj}^{(t)}) \geq 0$ for any $|\lambda_{dj}| \geq 0$; therefore, $\mathcal{L}(\theta) - Q_{\text{LLA}}(\theta | \theta^{(t)}) \geq 0$ because $\mathcal{L}_{\text{ML}}(\theta) - Q_{\text{ML}}(\theta | \theta^{(t)}) \geq 0$ using Lemma 1 and Theorem 1 in Dempster et al. (1977). If $\theta^{(t+1)}$ maximizes $Q_{\text{LLA}}(\theta | \theta^{(t)})$, then

$$\mathcal{L}(\theta^{(t+1)}) \geq Q_{\text{LLA}}(\theta^{(t+1)} | \theta^{(t)}) \geq Q_{\text{LLA}}(\theta^{(t)} | \theta^{(t)}) = \mathcal{L}(\theta^{(t)}), \quad (13)$$

where the last equality follows from (10). The objective (1) is bounded in probability on the parameter space, so the sequence $\{\mathcal{L}(\theta^{(t)})\}_{t=1}^{\infty}$ converges to some $\mathcal{L}(\theta^{(\infty)})$. Using Proposition 1 in Zou & Li (2008), $\theta^{(t)}$ converges to the stationary point $\theta^{(\infty)}$.

3.2. Proof of asymptotic normality of nonzero loadings and consistency of estimated Λ

The proof has two steps. First, we show asymptotic normality of nonzero loadings. Second, we use results of the first step to show consistency of the estimated loadings.

Step 1. Let $\lambda_{dj}^0/p^{1/2}$ and $\sigma_d^{2^0}$ are the root- n consistent sequence of estimators of $\lambda_{dj}^*/p^{1/2}$ and $\sigma_d^{2^*}$ ($d = 1, \dots, p$; $j = 1, \dots, k$) as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$, then imputing Z

based on the eigen decomposition of $Y^T Y/n$ in (4) implies that

$$\hat{\Lambda} = \underset{\lambda_d}{\operatorname{argmin}} \sum_{d=1, \dots, p} \frac{\|y_d/p^{1/2} - Z^0 \lambda_d/p^{1/2}\|^2}{2\sigma_d^{2^0}/p} + \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{\eta_j + |\lambda_{dj}^0|/p^{1/2}} |\lambda_{dj}/p^{1/2}|, \quad (14)$$

where $\hat{\Lambda}$ is the estimate of Λ obtained using the estimation algorithm of expandable factor analysis,

$$\begin{aligned} \sigma_d^{2^0} &= \sum_{d=k+1}^p \hat{\zeta}_d \hat{\psi}_{dj}^2, \quad \lambda^0 = \hat{\zeta}_j^{1/2} \hat{\psi}_{dj}, \quad (d = 1, \dots, p; j = 1, \dots, k), \\ Z^0 &= Y \left(\hat{\zeta}_1^{-1/2} \hat{\psi}_1, \dots, \hat{\zeta}_k^{-1/2} \hat{\psi}_k \right). \end{aligned} \quad (15)$$

Again using (4),

$$\frac{Z^{0T} Z^0}{n} = \left(\hat{\zeta}_1^{-1/2} \hat{\psi}_1, \dots, \hat{\zeta}_k^{-1/2} \hat{\psi}_k \right)^T Y^T Y/n \left(\hat{\zeta}_1^{-1/2} \hat{\psi}_1, \dots, \hat{\zeta}_k^{-1/2} \hat{\psi}_k \right) = I_k. \quad (16)$$

If U is a $p \times k$ matrix independent of n and p and u_d^T represents row d of U , then define

$$V_n(U) = \sum_{d=1}^p \frac{\left\| \frac{y_d}{p^{1/2}} - Z^0 \left(\frac{\lambda_d^*}{p^{1/2}} + \frac{u_d}{(np)^{1/2}} \right) \right\|^2}{2\sigma_d^{2^0}/p} + \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{\eta_j + \frac{|\lambda_{dj}^0|}{p^{1/2}}} \left| \frac{\lambda_{dj}^*}{p^{1/2}} + \frac{u_{dj}}{(np)^{1/2}} \right|, \quad (17)$$

where vectors are added component-wise. Substitute $u_{dj} = 0$ ($d = 1, \dots, p; j = 1, \dots, k$) in (17) to obtain

$$V_n(0) = \sum_{d=1}^p \frac{\left\| \frac{y_d}{p^{1/2}} - Z^0 \frac{\lambda_d^*}{p^{1/2}} \right\|^2}{2\sigma_d^{2^0}/p} + \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{\eta_j + \frac{|\lambda_{dj}^0|}{p^{1/2}}} \left| \frac{\lambda_{dj}^*}{p^{1/2}} \right|. \quad (18)$$

Using (15) and (16),

$$\begin{aligned} V_n(U) - V_n(0) &= \sum_{d=1}^p \frac{u_d^T u_d}{2\sigma_d^{2^0}} - \sum_{d=1}^p \frac{n^{1/2} u_d^T}{\sigma_d^{2^0}} \left(\frac{Z^{0T} y_d}{n} - \lambda_d^* \right) + \\ &\quad \sum_{d=1}^p \sum_{j=1}^k \frac{\alpha_j + 1}{\eta_j + \frac{|\lambda_{dj}^0|}{p^{1/2}}} \left(\left| \frac{\lambda_{dj}^*}{p^{1/2}} + \frac{u_{dj}}{(np)^{1/2}} \right| - \left| \frac{\lambda_{dj}^*}{p^{1/2}} \right| \right) \\ &\equiv \sum_{d=1}^p T_{1d} - \sum_{d=1}^p T_{2d} + \sum_{d=1}^p \sum_{j=1}^k T_{3dj}. \end{aligned} \quad (19)$$

The limiting forms of all the terms in (19) are derived next. First, we obtain the limiting form of T_{1d} in (19). Because $\sigma_d^{2^0}$ is a root- n consistent estimator of $\sigma_d^{2^*}$, $T_{1d} \rightarrow (u_d^T u_d)/(2\sigma_d^{2^*})$ in probability as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$ using Slutsky's theorem. Second, we obtain the limiting form of T_{2d} in (19). Lemma 1 shows that variance of λ_{dj}^0 ($d = 1, \dots, p; j = 1, \dots, k$) is bounded, so using (4), Slutsky's theorem, and the central limit theorem,

$$T_{2d} = n^{1/2} (\lambda_{d1}^0 - \lambda_{d1}^*, \dots, \lambda_{dk}^0 - \lambda_{dk}^*) \frac{u_d}{\sigma_d^{2^0}} \rightarrow \frac{u_d^T r_d}{\sigma_d^{2^*}} \quad (d = 1, \dots, p) \quad (20)$$

as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$, where the convergence is in distribution and $r_d \sim N_k(0_{k \times 1}, C_d)$ for some symmetric positive definite matrix C_d . Let \mathcal{M}_d^* be the set of j s such that λ_{dj}^* is nonzero, then $\mathcal{M}_d^* = \{j : (d, j) \in \mathcal{M}^*\}$ and $\mathcal{M}_d^{*c} = \{j : (d, j) \notin \mathcal{M}^*\}$. If $A_{\mathcal{B}\mathcal{B}}$ denotes a sub-matrix that contains the rows and the columns of matrix A with indices in \mathcal{B} , then the block partitioned form of the covariance matrix of r_d in (20) based on \mathcal{M}_d^* is

$$C_d^* = \begin{bmatrix} C_{d_{\mathcal{M}_d^* \mathcal{M}_d^*}} & C_{d_{\mathcal{M}_d^* \mathcal{M}_d^{*c}}} \\ C_{d_{\mathcal{M}_d^{*c} \mathcal{M}_d^*}} & C_{d_{\mathcal{M}_d^{*c} \mathcal{M}_d^{*c}}} \end{bmatrix}, \quad (r_{d_{\mathcal{M}_d^*}}, r_{d_{\mathcal{M}_d^{*c}}})^T \sim N_k(0_{k \times 1}, C_d^*), \quad (d = 1, \dots, p), \quad (21)$$

where $r_{d_{\mathcal{M}_d^*}}$ and $r_{d_{\mathcal{M}_d^{*c}}}$ include elements of r_d with indices in \mathcal{M}_d^* and \mathcal{M}_d^{*c} , respectively. Finally, the limiting form of T_{3dj} is found using arguments in Zou & Li (2008). If $\lambda_{dj}^* \neq 0$, then $\eta_j + p^{-1/2}|\lambda_{dj}^0| = \eta_j + p^{-1/2}|\lambda_{dj}^*| + O_P\{(np)^{-1/2}\}$, $(np)^{1/2}(|p^{-1/2}\lambda_{dj}^* + (np)^{-1/2}u_{dj}| - |p^{-1/2}\lambda_{dj}^*|) = \text{sign}(\lambda_{dj}^*)u_{dj}$, and

$$T_{3dj} = \frac{\{n^{-1/2}(\alpha_j + 1)\} [(np)^{1/2}\{|p^{-1/2}\lambda_{dj}^* + (np)^{-1/2}u_{dj}| - |p^{-1/2}\lambda_{dj}^*|\}]}{\{p^{1/2}\eta_j + |\lambda_{dj}^*| + O_P(n^{-1/2})\}} \rightarrow 0$$

in probability by Slutsky's theorem and the continuous mapping theorem as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$. Similarly, if $\lambda_{dj}^* = 0$, then $\eta_j + p^{-1/2}|\lambda_{dj}^0| = \eta_j + O_P\{(np)^{-1/2}\}$, $(np)^{1/2}(|p^{-1/2}\lambda_{dj}^* + (np)^{-1/2}u_{dj}| - |p^{-1/2}\lambda_{dj}^*|) = |u_{dj}|$, and

$$T_{3dj} = \frac{(\alpha_j + 1)[(np)^{1/2}\{|p^{-1/2}\lambda_{dj}^* + (np)^{-1/2}u_{dj}| - |p^{-1/2}\lambda_{dj}^*|\}]}{\{(np)^{1/2}\eta_j + O_P(1)\}} \rightarrow \begin{cases} 0, & u_{dj} = 0, \\ \infty, & u_{dj} \neq 0, \end{cases} \quad (22)$$

in probability by Slutsky's theorem and the continuous mapping theorem as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$.

Let $\hat{U}_n = \underset{U}{\operatorname{argmin}}\{V_n(U) - V_n(0)\}$, then $\hat{\lambda}_{dj}/p^{1/2} = \lambda_{dj}^*/p^{1/2} + \hat{u}_{dj_n}/(np)^{1/2}$ or $n^{1/2}(\hat{\lambda}_{dj} - \lambda_{dj}^*) = \hat{u}_{dj_n}$. The limiting forms of T_{1d} , T_{2d} , and T_{3dj} ($d = 1, \dots, p$; $j = 1, \dots, k$), and Slutsky's theorem imply that $V_n(U) - V_n(0) \rightarrow V^*(U)$ in distribution for every U as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$, where

$$V^*(U) = \begin{cases} \sum_{(d,j) \in \mathcal{M}^*} \frac{u_{dj}^2}{2\sigma_d^{*2}} - \sum_{(d,j) \in \mathcal{M}^*} \frac{u_{dj}r_{dj}}{\sigma_d^{*2}}, & u_{dj} = 0 \text{ for all } (d, j) \notin \mathcal{M}^*, \\ \infty, & \text{otherwise.} \end{cases} \quad (23)$$

Since $V_n(U) - V_n(0)$ is convex, the unique minimizer of $V^*(U)$ is

$$U^* \text{ such that } u_{dj}^* = \begin{cases} 0, & (d, j) \notin \mathcal{M}^*, \\ r_{dj}, & (d, j) \in \mathcal{M}^*. \end{cases} \quad (24)$$

Following the epi-convergence results of Geyer (1994) and Knight & Fu (2000), $\hat{u}_{dj_n} \rightarrow u_{dj}^*$ in distribution ($d = 1, \dots, p$; $j = 1, \dots, k$) as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$. Let $\hat{\lambda} = (\hat{\lambda}_1^T, \dots, \hat{\lambda}_p^T)^T$, $\lambda^* = (\lambda_1^{*T}, \dots, \lambda_p^{*T})^T$, and $|A|$ be the cardinality of set A , then

$$n^{1/2}(\hat{\lambda}_{\mathcal{M}^*} - \lambda_{\mathcal{M}^*}^*) \rightarrow (r_{1_{\mathcal{M}_1^*}}, \dots, r_{p_{\mathcal{M}_p^*}})^T \equiv r_{\mathcal{M}^*}, \quad \hat{\lambda}_{\mathcal{M}^{*c}} \rightarrow 0_{|\mathcal{M}^{*c}| \times 1}$$

in distribution as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$ using (21), (22), (24), and $n^{1/2}(\hat{\lambda}_{dj} - \lambda_{dj}^*) = \hat{u}_{djn} \rightarrow u_{dj}^*$ in distribution ($d = 1, \dots, p; j = 1, \dots, k$). Further,

$$r_{\mathcal{M}^*} \sim N_{|\mathcal{M}^*|} (0_{|\mathcal{M}^*| \times 1}, C_{\mathcal{M}^* \mathcal{M}^*}), \quad C_{\mathcal{M}^* \mathcal{M}^*} = \text{bdiag}(C_{1_{\mathcal{M}_1^* \mathcal{M}_1^*}}, \dots, C_{p_{\mathcal{M}_p^* \mathcal{M}_p^*}}),$$

where $\text{bdiag}(C_{1_{\mathcal{M}_1^* \mathcal{M}_1^*}}, \dots, C_{p_{\mathcal{M}_p^* \mathcal{M}_p^*}})$ is a block diagonal matrix with $C_{1_{\mathcal{M}_1^* \mathcal{M}_1^*}}, \dots, C_{p_{\mathcal{M}_p^* \mathcal{M}_p^*}}$ forming the diagonal blocks. This proves the asymptotic normality of nonzero loadings.

Step 2. We now prove the consistency of $\hat{\lambda}_{dj}$ ($d = 1, \dots, p; j = 1, \dots, k$). For every $(d, j) \in \mathcal{M}^*$, asymptotic normality of $\hat{\lambda}_{dj}$ implies that $\lambda_{dj} \rightarrow \lambda_{dj}^*$ in probability, so $\text{pr}\{(d, j) \in \hat{\mathcal{M}}\} \rightarrow 1$, where $\hat{\mathcal{M}}$ is the estimated set of the locations of nonzero loadings based on $\hat{\Lambda}$. The proof is completed by showing that for all $(\tilde{d}, \tilde{j}) \notin \mathcal{M}^*$, $\text{pr}\{(\tilde{d}, \tilde{j}) \in \hat{\mathcal{M}}\} \rightarrow 0$. Let $(\tilde{d}, \tilde{j}) \in \hat{\mathcal{M}}$, then Karush-Kuhn-Tucker optimality condition implies that

$$n^{-1/2} z_j^{0T} (y_{\tilde{d}} - Z^0 \hat{\lambda}_{\tilde{d}}) = \text{sign}(\hat{\lambda}_{\tilde{d}\tilde{j}}) \frac{\sigma_{\tilde{d}}^{2^0} (\alpha_{\tilde{j}} + 1)}{\eta_{\tilde{j}} (np)^{1/2} + O_P(1)}. \quad (25)$$

The right hand side of (25) is unbounded in probability as $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$ because $(\tilde{d}, \tilde{j}) \notin \mathcal{M}^*$. The left hand side of (25) is

$$n^{1/2} \left(\frac{z_j^{0T} y_{\tilde{d}}}{n} - \frac{z_j^{0T} Z^0}{n} \lambda_{\tilde{d}}^* \right) + \frac{z_j^{0T} Z^0}{n} \left\{ n^{1/2} (\lambda_{\tilde{d}}^* - \hat{\lambda}_{\tilde{d}}) \right\}. \quad (26)$$

Following arguments similar to those used to derive (20), the first term in (26) is asymptotically normal. The second term in (26) is also asymptotically normal from asymptotic normality of the estimates of nonzero loadings shown previously. By Slutsky's theorem, the left hand side of (25) is asymptotically normal; therefore,

$$\text{pr}\{(\tilde{d}, \tilde{j}) \in \hat{\mathcal{M}}\} \leq \text{pr} \left\{ n^{-1/2} z_j^{0T} (y_{\tilde{d}} - Z^0 \hat{\lambda}_{\tilde{d}}) = \text{sign}(\hat{\lambda}_{\tilde{d}\tilde{j}}) \frac{\sigma_{\tilde{d}}^{2^0} (\alpha_{\tilde{j}} + 1)}{\eta_{\tilde{j}} (np)^{1/2} + O_P(1)} \right\} \rightarrow 0 \quad (27)$$

in probability because asymptotic normality of $n^{-1/2} z_j^{0T} (y_{\tilde{d}} - Z^0 \hat{\lambda}_{\tilde{d}})$ implies that it is bounded in probability. This proves the consistency of $\hat{\lambda}_{dj}$ ($d = 1, \dots, p; j = 1, \dots, k$).

3.3. Proof of asymptotic normality and consistency of estimated Σ

We now prove asymptotic normality and consistency of $\hat{\sigma}_d^2$ ($d = 1, \dots, p$). We first show that $\hat{\sigma}_d^2$ is consistent. For the root- n consistent sequence of estimators $\lambda_{dj}^0/p^{1/2}$, ($d = 1, \dots, p; j = 1, \dots, k$), Assumption A.5 in the main paper and the continuous mapping theorem imply that if $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$, then $L^0 = \{\Omega^* + o_P(1)\} \{\Omega^{*-1} \Lambda^* + o_P(1)\}$, where convergence is element-wise, and

$$\begin{aligned} \hat{\sigma}_d^2 &= \{1 + o(1)\} \left\{ \lambda_d^{*T} \lambda_d^* + o_P(1) - 2\lambda_d^{*T} \lambda_d^* + o_P(1) + (\Omega^*)_{dd} + o_P(1) \right\} \\ &= -\lambda_d^{*T} \lambda_d^* + (\Omega^*)_{dd} + o_P(1) = \sigma_d^{2^*} + o_P(1), \end{aligned} \quad (28)$$

which proves the consistency of $\hat{\sigma}_d^2$.

The asymptotic normality of $\hat{\sigma}_d^2$ follows from Equation (5.19) and Exercise 5.20 in van der Vaart (2000) because the objective for estimating $\hat{\sigma}_d^2$ has two continuous derivatives with respect to σ_d^2 for any Y and Λ .

3.4. Lemma required to prove Theorem 3

We use the eigen decomposition of $Y^T Y/n$ to impute Σ and Z in Equation (3) of the main paper. Using the notation of Algorithm 1 in the main paper, impute Σ by Σ^0 and Z by Z^0 and let $y = \text{vec}(Y)$, $\lambda = \text{vec}(\Lambda^T)$, $\epsilon = \text{vec}(E^T)$, and $X = I_p \otimes Z^0 \in \mathfrak{R}^{pn \times pk}$. Then, the hierarchical model for the joint distribution of y and λ after scaling Equation (3) in the main paper by $p^{1/2}$ is

$$p^{-1/2}y \mid \lambda \sim N_{np}(X p^{-1/2}\lambda, p^{-1}\Sigma^0 \otimes I_n),$$

$$\lambda \mid \delta, \rho \sim \text{multiscale generalized double Pareto}\{\alpha_1(\delta), \dots, \alpha_k(\delta), p^{1/2}\eta_1(\rho), \dots, p^{1/2}\eta_k(\rho)\}. \quad (29)$$

The density of the prior for loadings that are estimated to be nonzero in \mathcal{M} is $\prod_{(d,j) \in \mathcal{M}} p_{\text{gdP}}(\lambda_{dj})$, where $p_{\text{gdP}}(\cdot)$ is the density of the generalized double Pareto prior in Section 2.2 of the main paper. The log likelihood of $\lambda_{\mathcal{M}}$ given \mathcal{M} is

$$\begin{aligned} \log f_G(y \mid \lambda_{\mathcal{M}}) &= \frac{np \log p}{2} - \frac{np}{2} \frac{\sum_{d=1}^p \log(\Sigma^0)_{dd}}{p} - \frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^n \sum_{d=1}^p y_{id}^2 / (\Sigma^0)_{dd} - \\ &\quad \frac{n}{2} \sum_{(d,j) \in \mathcal{M}} \lambda_{dj}^2 / (\Sigma^0)_{dd} + n \sum_{(d,j) \in \mathcal{M}} \lambda_{dj} \lambda_{dj}^0 / (\Sigma^0)_{dd} \end{aligned} \quad (30)$$

and the log joint density of y and $\lambda_{\mathcal{M}}$ given \mathcal{M} is

$$\log f(y, \lambda_{\mathcal{M}} \mid \delta, \rho) = \log f_G(y \mid \lambda_{\mathcal{M}}) + \sum_{(d,j) \in \mathcal{M}} \log p_{\text{gdP}}(\lambda_{dj}). \quad (31)$$

The following lemma describes the order of $\log f_G(y \mid \lambda_{\mathcal{M}})$ and $\log f(y, \lambda_{\mathcal{M}} \mid \delta, \rho)$ when $\lambda_{\mathcal{M}}$ is replaced by a consistent estimator of $\lambda_{\mathcal{M}}^*$ and $n \rightarrow \infty$, $n \leq p \rightarrow \infty$, and $\log p/n \rightarrow 0$.

LEMMA 2. *If $\tilde{\lambda}_{\mathcal{M}}$ and $\hat{\lambda}_{\mathcal{M}}$ are root- n consistent estimators of $\lambda_{\mathcal{M}}^*$ and Assumptions A.0–A.7 in the main paper hold, then*

$$2 \log f_G(y \mid \tilde{\lambda}_{\mathcal{M}}) / (np \log p) = 2 \log f(Y, \hat{\lambda}_{\mathcal{M}} \mid \delta, \rho) / (np \log p) = 1 + o_P(1).$$

Proof. We first show that

$$2 \log f_G(y \mid \tilde{\lambda}_{\mathcal{M}}) / (np \log p) = 1 + o_P(1).$$

Using (30), $2 \log f_G(y \mid \tilde{\lambda}_{\mathcal{M}}) / (np \log p)$

$$\begin{aligned} &= 1 - \frac{1}{\log p} \frac{\sum_{d=1}^p \log(\Sigma^0)_{dd}}{p} - \frac{\log 2\pi}{\log p} - \frac{1}{\log p} \frac{\sum_{i=1}^n \sum_{d=1}^p y_{id}^2 / (\Sigma^0)_{dd}}{np} - \\ &\quad \frac{\sum_{(d,j) \in \mathcal{M}} \tilde{\lambda}_{dj}^2 / (\Sigma^0)_{dd}}{p \log p} + 2 \frac{\sum_{(d,j) \in \mathcal{M}} \tilde{\lambda}_{dj} \lambda_{dj}^0 / (\Sigma^0)_{dd}}{p \log p} \\ &= 1 - \frac{1}{\log p} \frac{\sum_{d=1}^p \log(\Sigma^*)_{dd}}{p} + o_P(1) - o(1) - \frac{1}{\log p} \frac{\sum_{i=1}^n \sum_{d=1}^p y_{id}^2 / (\Sigma^*)_{dd}}{np} O_P(1) - \\ &\quad \frac{\sum_{(d,j) \in \mathcal{M}} \lambda_{dj}^{*2} / (\Sigma^*)_{dd}}{p \log p} O_P(1) + 2 \frac{\sum_{(d,j) \in \mathcal{M}} \lambda_{dj}^{*2} / (\Sigma^*)_{dd}}{p \log p} O_P(1), \end{aligned}$$

where the last equality follows because $(\Sigma^0)_{dd}$ and $\tilde{\lambda}_{dj}$ are consistent estimators of $(\Sigma^*)_{dd}$ and λ_{dj}^* ($d = 1, \dots, p$; $j = 1, \dots, k$). Since $E(y_{id}^2) \leq D_0$ and $D_1 \leq (\Sigma^*)_{dd} \leq D_2$ ($i = 1, \dots, n$;

$d = 1, \dots, p$) using Assumption A.1 in the main paper,

$$\sum_{i=1}^n \sum_{d=1}^p (\Sigma^*)_{dd}^{-1} y_{id}^2 / (np) = O_P(1)$$

by an application of Markov's inequality and

$$\begin{aligned} \log D_1 &\leq \sum_{d=1}^p \log(\Sigma^*)_{dd}/p \leq \log D_2, \\ 0 &\leq \sum_{(d,j) \in \mathcal{M}} (\Sigma^*)_{dd}^{-1} \lambda_{dj}^{2*} / (p \log p) \leq \text{tr}(\Omega^*) / (D_1 p \log p) \leq D_0 / (D_1 \log p). \end{aligned}$$

Therefore,

$$\frac{2 \log f_G(y \mid \tilde{\lambda}_{\mathcal{M}}, \Sigma^0)}{np \log p} = 1 + o_P(1) + \frac{O_P(1)}{\log p} + \frac{O_P(1)}{\log p} = 1 + o_P(1).$$

Proceeding similarly,

$$\frac{2 \log f_G(y \mid \hat{\lambda}_{\mathcal{M}})}{np \log p} = 1 + o_P(1)$$

using the consistency of $\hat{\lambda}_{\mathcal{M}}$.

We complete the proof by showing that

$$\frac{\sum_{(d,j) \in \mathcal{M}} \log p_{\text{gdP}}(\hat{\lambda}_{dj})}{np \log p} = o_P(1).$$

Using the analytic form of p_{gdP} in (31),

$$\sum_{(d,j) \in \mathcal{M}} \log p_{\text{gdP}}(\hat{\lambda}_{dj}) = \sum_{(d,j) \in \mathcal{M}} \log \frac{\alpha_j}{p^{1/2} \eta_j} - \sum_{(d,j) \in \mathcal{M}} (\alpha_j + 1) \log \left(1 + \frac{|\hat{\lambda}_{dj}|}{p^{1/2} \eta_j} \right). \quad (32)$$

The first term on the right hand side of (32) after scaling by $np \log p$ is

$$\begin{aligned} \frac{1}{np \log p} \sum_{(d,j) \in \mathcal{M}} \log \frac{\alpha_j}{p^{1/2} \eta_j} &= \frac{1}{p \log p} \sum_{(d,j) \in \mathcal{M}} \left[\frac{\log \alpha_j}{n} - \frac{\log \{(np)^{1/2} \eta_j\}}{n} + \frac{\log n}{2n} \right] \\ &= o(1) \frac{|\mathcal{M}|}{p \log p} = o(1) O(1) = o(1). \end{aligned}$$

The last equality follows from Assumption A.5 in the main paper and using conditions that $|\mathcal{M}| \leq pk$ and $k = O(\log p)$. The second term on the right hand side of (32) after scaling by $np \log p$ is

$$\begin{aligned} \frac{1}{np \log p} \sum_{(d,j) \in \mathcal{M}} (\alpha_j + 1) \log \left(1 + \frac{|\hat{\lambda}_{dj}|}{p^{1/2} \eta_j} \right) &= \frac{1}{p \log p} \sum_{(d,j) \in \mathcal{M}} \frac{\alpha_j + 1}{n^{1/2}} \frac{\log \{(np)^{1/2} \eta_j + n^{1/2} \hat{\lambda}_{dj}\}}{n^{1/2}} \\ &\quad - \frac{1}{p \log p} \sum_{(d,j) \in \mathcal{M}} \frac{\alpha_j + 1}{n^{1/2}} \frac{\log \{(np)^{1/2} \eta_j\}}{n^{1/2}} \\ &= o_P(1) \frac{|\mathcal{M}|}{p \log p} - o(1) \frac{|\mathcal{M}|}{p \log p} = o_P(1). \end{aligned}$$

The last equality follows from Assumption A.5 in the main paper, from consistency of $\hat{\lambda}_{dj}$, and using conditions that $|\mathcal{M}| \leq pk$ and $k = O(\log p)$. The proof is completed by using (31) to obtain that

$$\frac{2 \log f(y, \hat{\lambda}_{\mathcal{M}} | \delta, \rho)}{np \log p} = \frac{2 \log f_G(y | \hat{\lambda}_{\mathcal{M}})}{np \log p} + \frac{2 \sum_{(d,j) \in \mathcal{M}} \log p_{\text{gdP}}(\hat{\lambda}_{dj})}{np \log p} = 1 + o_P(1).$$

3.5. Proof of Theorem 3

The proof consists of three steps: derive the asymptotic form of $\log \pi_{\mathcal{M}}$; show that $-2 \log \pi_{\mathcal{M}} / \text{EBIC}_{\gamma}(\mathcal{M}) = 1 + o_P(1)$; and show that the sufficient condition for model selection consistency of $\text{EBIC}_{\gamma}(\mathcal{M})$ holds under the assumptions of Theorem 3 in the main paper.

We use the following notation for ease of presentation. If \mathcal{B} is a set of indices and X is a matrix, then $X_{\mathcal{B}}$ is a sub-matrix that contains columns of X with indices in \mathcal{B} and $X_{\mathcal{B}, \mathcal{B}}$ is a sub-matrix that contains rows and columns of X with indices in \mathcal{B} .

Step 1. Using (29), the density of the prior for loadings that are estimated to be nonzero in \mathcal{M} is $\prod_{(d,j) \in \mathcal{M}} p_{\text{gdP}}(\lambda_{dj})$; see Section 3.4 also. Use the Gaussian scale mixture representation for the density of generalized double Pareto prior to write $|\lambda_{dj}|$ in form of differentiable functions when $\lambda_{dj} \neq 0$; see the equation for E-step in Section 4.4.1 of Armagan et al. (2013) for details related to the Gaussian scale mixture representation for the generalized double Pareto density. Define the diagonal matrix D as

$$D = \frac{d^2 \log \left\{ \prod_{(d,j) \in \mathcal{M}} p_{\text{gdP}}(\lambda_{dj}) \right\}}{d\lambda_{\mathcal{M}} d\lambda_{\mathcal{M}}^T},$$

and let

$$D_{(d,j),(d,j)} = \frac{\alpha_j(\delta) + 1}{\{p^{1/2}\eta_j(\rho) + |\lambda_{dj}|\}^2}$$

be the diagonal element of D corresponding to λ_{dj} such that $(d, j) \in \mathcal{M}$. If $f(y, \lambda_{\mathcal{M}} | \delta, \rho)$ is the joint density of y and $\lambda_{\mathcal{M}}$ defined using (29), then define another diagonal matrix $H_{\mathcal{M}}$ as

$$H_{\mathcal{M}} = -\frac{d^2 \log f(y, \lambda_{\mathcal{M}} | \delta, \rho)}{d\lambda_{\mathcal{M}} d\lambda_{\mathcal{M}}^T} = n(\Sigma^{0^{-1}} \otimes I_n)_{\mathcal{M}, \mathcal{M}} - D. \quad (33)$$

If $\hat{H}_{\mathcal{M}}$ represents $H_{\mathcal{M}}$ in (33) evaluated at $\hat{\lambda}_{\mathcal{M}}$, then the diagonal element of $\hat{H}_{\mathcal{M}}$ that corresponds to the index $(d, j) \in \mathcal{M}$ is

$$\frac{n}{\sigma_d^{2^0}} - \frac{\alpha_j(\delta) + 1}{\{p^{1/2}\eta_j(\rho) + |\hat{\lambda}_{dj}|\}^2} = \begin{cases} \frac{n}{\sigma_d^{2^*}} \{1 + o_P(n^{-1/2})\}, & (d, j) \in \mathcal{M}^*, \\ \frac{n}{\sigma_d^{2^*}} \{1 + o_P(n^{1/2})\}, & (d, j) \notin \mathcal{M}^*. \end{cases} \quad (34)$$

The equality in (34) follows because $\hat{\lambda}_{dj} = \lambda_{dj}^* + o_P(n^{-1/2})$, $\sigma_d^{2^0} = \sigma_d^{2^*} + o_P(n^{-1/2})$, and $\alpha_j(\delta) = o(n^{1/2})$ from Theorem 2 in the main paper and Assumptions A.0–A.6 in the main paper. The posterior probability of \mathcal{M} , denoted as $\pi_{\mathcal{M}}$, equals

$$\text{pr}(\mathcal{M} | Y, \delta, \rho) \propto m(Y | \mathcal{M}) \text{pr}(\mathcal{M} | \delta, \rho), \quad (35)$$

where $m(Y | \mathcal{M})$ is the marginal likelihood of the factor model in (29) with the locations of nonzero loadings contained in the set \mathcal{M} , $m(y | \mathcal{M}) = \int f(y, \lambda_{\mathcal{M}} | \delta, \rho) d\lambda_{\mathcal{M}}$, and $\text{pr}(\mathcal{M} | \delta, \rho)$

is prior defined in Equation (9) in the main paper. Using Laplace approximation and (34),

$$2 \log m(Y | \mathcal{M}) = 2 \log f(Y, \hat{\lambda}_{\mathcal{M}} | \delta, \rho) - |\mathcal{M}| \log n [1 + \{c + o_P(\log n)\} / \log n], \quad (36)$$

where $c = \log(2\pi) + \sum_{(d,j) \in \mathcal{M}} \sigma_d^{2*} / |\mathcal{M}| = O(1)$ using Assumption A.1 in the main paper. Further, using (35),

$$-2 \log \pi_{\mathcal{M}} = -2 \log m(Y | \mathcal{M}) - 2 \log \text{pr}(\mathcal{M} | \delta, \rho).$$

Sterling's approximation and Theorem 2 imply that $\log \text{pr}(\mathcal{M} | \delta, \rho) = -|\mathcal{M}| \log(pk) \{1 + o_P(1)\}$; therefore, the previous equation after using (36) reduces to

$$-2 \log \pi_{\mathcal{M}} = -2 \log f(Y, \hat{\lambda}_{\mathcal{M}} | \delta, \rho) + |\mathcal{M}| \{\log n + 2 \log(pk)\} \{1 + o_P(1)\}. \quad (37)$$

Step 2. The definition of $\text{EBIC}_{\gamma}(\mathcal{M})$ in Chen & Chen (2008) for regression models implies that

$$\begin{aligned} \text{EBIC}_{\gamma}(\mathcal{M}) &= -2 \log f_G(y | \tilde{\lambda}_{\mathcal{M}}) + |\mathcal{M}| \{\log(np) + 2\gamma \log(pk)\}, \\ &= -2 \log f_G(y | \tilde{\lambda}_{\mathcal{M}}) + |\mathcal{M}| \{\log n + (2\gamma + 1) \log p\} \{1 + o_P(1)\}, \end{aligned} \quad (38)$$

where $\tilde{\lambda}_{dj}$ is a root- n consistent estimate of λ_{dj}^* ($d = 1, \dots, p; j = 1, \dots, k$) in (29), $f_G(y | \lambda_{\mathcal{M}})$ is the Gaussian likelihood defined using (29), and $0 < \gamma < 1$ is a tuning parameter such that $\gamma > 1 - 1/(2\kappa)$. Lemma 3.4 implies that there exists a universal constant b^* such that

$$-2 \log f_G(y | \tilde{\lambda}_{\mathcal{M}}) / (np \log p) = -2 \log f(Y, \hat{\lambda}_{\mathcal{M}} | \delta, \rho) / (np \log p) = b^* + o_P(1). \quad (39)$$

Let $r = -2 \log \pi_{\mathcal{M}} / \text{EBIC}_{\gamma}(\mathcal{M})$. Then, Theorem 2 in the main paper and (39) imply that

$$\begin{aligned} r &= \frac{-2 \log f(Y, \hat{\lambda}_{\mathcal{M}} | \delta, \rho) / (np \log p) + |\mathcal{M}| \{\log n + 2 \log(pk)\} / (np \log p) \{1 + o_P(1)\}}{-2 \log f_G(y | \tilde{\lambda}_{\mathcal{M}}) / (np \log p) + |\mathcal{M}| \{\log n + (2\gamma + 1) \log p\} / (np \log p) \{1 + o_P(1)\}} \\ &= \frac{b^* + o_P(1) + \{|\mathcal{M}^*| + o_P(1)\} o_P(1)}{b^* + o_P(1) + \{|\mathcal{M}^*| + o_P(1)\} o_P(1)} = 1 + o_P(1). \end{aligned} \quad (40)$$

Step 3. Let l be an upper bound on k^* in (29) such that $X \in \mathbb{R}^{pn \times pl}$. If $(np)^{-1} X^T X$ has positive eigen values for any l such that $k \leq l \leq 2k$, $\mathcal{M} \neq \mathcal{M}^*$, and $|\mathcal{M}| \in \{1, \dots, pk\}$, then uniformly for any such \mathcal{M} there is a universal positive constant C_0 and a positive constant $C_{\mathcal{M}}$ depending on \mathcal{M} such that

$$\text{EBIC}_{\gamma}(\mathcal{M}) - \text{EBIC}_{\gamma}(\mathcal{M}^*) \geq \begin{cases} C_0 \log n \{1 + o_P(1)\}, & \mathcal{M}^* \subset \mathcal{M}, \\ C_{\mathcal{M}} \log n, & \text{otherwise;} \end{cases} \quad (41)$$

see the definition of asymptotic identifiability condition on pages 762–763 in Chen & Chen (2008) and the proof of Theorem 1 in Chen & Chen (2008). Using (40) and (41), $2 \log(\pi_{\mathcal{M}^*} / \pi_{\mathcal{M}}) = \{\text{EBIC}_{\gamma}(\mathcal{M}) - \text{EBIC}_{\gamma}(\mathcal{M}^*)\} \{1 + o_P(1)\} \rightarrow \infty$ as $n \rightarrow \infty$ for any \mathcal{M} such that $\mathcal{M} \neq \mathcal{M}^*$ and $|\mathcal{M}| \in \{1, \dots, pk\}$. The proof is completed by showing $(np)^{-1} X^T X$ has positive eigen values for any l such that $k \leq l \leq 2k$. Assumption A.7 implies that $Y^T Y / n$ has at least $2k$ positive eigen values, so $(np)^{-1} X^T X = I_p \otimes (np)^{-1} Z^{0T} Z^0 = I_p \otimes p^{-1} I_l$, which has pl positive eigenvalues equal to $p^{-1} > 0$ for any $k \leq l \leq 2k$.

4. MICROARRAY DATA ANALYSIS

The AGEMAP data (Zahn et al., 2007) were obtained from <http://statweb.stanford.edu/~owen/data/AGEMAP/>.

The δ - ρ grid in expandable factor analysis had 20 different δ and 20 different ρ values: $\delta_i = 10^{a_i}$, where $a_i = \log_{10} 2 + (i - 1)(\log_{10} 10 - \log_{10} 2)/20$ ($i = 1, \dots, 20$), and $\rho_i = 10^{b_i}$, where $b_i = \log_{10} 10^{-3} + (i - 1)(\log 10^6 - \log 10^{-3})/20$ ($i = 1, \dots, 20$). Our estimation algorithm estimated Λ at grid points (δ_r, ρ_s) ($r = 1, \dots, 20$; $s = 1, \dots, 20$). The results of our estimation algorithm were stable in that the estimated rank of Λ was the same at most points on the δ - ρ grid across 10 folds of cross-validation (Table 1).

Table 1: Estimated rank of loadings matrix in AGEMAP data analysis across δ - ρ grid. The results are averaged over 10 folds of cross-validation and the maximum Monte Carlo error is 0.52 across the 10 folds.

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}	δ_{14}	δ_{15}	δ_{16}	δ_{17}	δ_{18}	δ_{19}	δ_{20}
ρ_{20}	10	10	10	10	10	10	10	9	9	8	8	7	7	7	6	6	6	6	6	5
ρ_{19}	10	10	10	10	10	10	9	9	8	8	7	7	6	6	6	6	6	5	5	5
ρ_{18}	10	10	10	10	9	9	8	8	7	7	6	6	6	6	5	5	5	5	4	4
ρ_{17}	10	10	10	9	9	8	7	7	6	6	6	5	5	5	4	4	4	4	4	4
ρ_{16}	10	10	9	8	8	7	7	6	6	6	5	5	5	4	4	4	4	4	4	4
ρ_{15}	10	9	8	7	7	6	6	6	6	5	5	4	4	4	4	4	4	4	4	3
ρ_{14}	8	7	7	6	6	6	6	5	4	4	4	4	4	4	4	4	3	3	3	3
ρ_{13}	7	6	6	6	6	5	4	4	4	4	4	4	4	3	3	3	3	3	3	3
ρ_{12}	6	6	5	4	4	4	4	4	4	4	3	3	3	3	3	3	2	2	2	2
ρ_{11}	5	4	4	4	4	4	3	3	3	3	3	2	2	2	2	2	2	2	2	2
ρ_{10}	4	4	4	3	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	2
ρ_9	3	3	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
ρ_8	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1
ρ_7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ρ_6	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ρ_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ρ_4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ρ_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ρ_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ρ_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

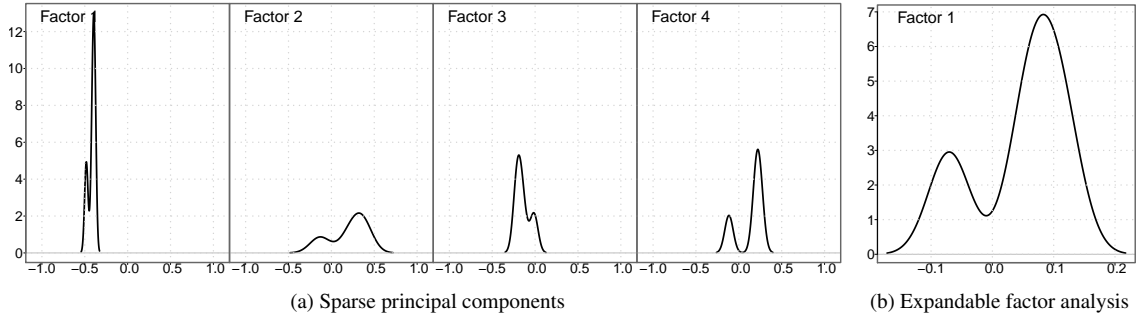


Fig. 1: Density plots for the estimated factors in a test data for cerebrum tissue samples.

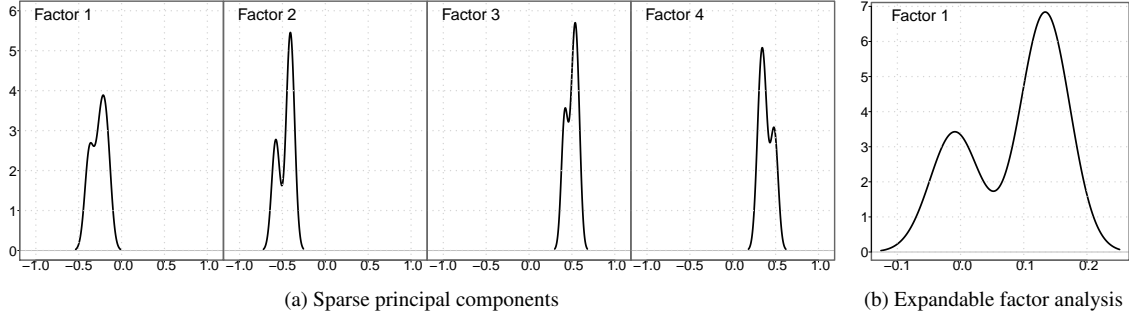


Fig. 2: Density plots for the estimated factors in a test data for cerebellum tissue samples.

5. CODE FOR COMPETING METHODS

We used R package *fanc* to obtain the results for Hirose and Yamamoto’s method (Hirose et al., 2015). Let $Y \in \mathfrak{R}^{n \times p}$ be the simulated data matrix. The following R code was used to estimate the loadings matrix, Λ_{HY} , and its rank, r_{HY} , using Hirose and Yamamoto’s method:

1. `nfactor = 20; tol = 1e-5`
2. `ctrl = list(length.rho = 20, length.gamma = 20, maxit.em = 1000, maxit.cd = 1000, tol.cd = tol, tol.em = tol)`
3. `fancfit = fanc(Y, factors = nfactor, control = ctrl, normalize = FALSE)`
4. `idx = which(fancfit$BIC == min(fancfit$BIC), arr.ind = TRUE)[1,]`
5. `fancout = out(fancfit, rho = fancfit$rho[idx[1]], gamma = fancfit$gamma[idx[2]])`
6. `fancLoad = as.matrix(fancout$loadings)`
7. `fancIdx = which(colSums(abs(fancLoad)) > 0)`
8. `rHY = length(fancIdx)`
9. `Λ_{HY} = as.matrix(fancLoad[, fancIdx])`

We used R package *PMA* to obtain the results for Witten et al.’s method (Witten et al., 2013). Let $Y \in \mathfrak{R}^{n \times p}$ be the simulated data matrix. The following R code was used to estimate the loadings matrix, Λ_W , and its rank, r_W , using Witten et al.’s method:

1. `nfactor = 20`
2. `spccv = SPC.cv(Y)`
3. `spcfit = SPC(Y, K = nfactor, sumabsv = spccv$bestsumabs)`
4. `rW = which(diff(spcfit$prop.var.explained) < 0.05)[1]`
5. `spcLoad = spcfit$v * matrix(sqrt(spcfit$d), nrow = ndim, ncol = length(spcfit$d), byrow = TRUE)`
6. `Λ_W = spcLoad[, 1:rW]`

The R function *FACTOR_ROTATE* implemented the first version of Ročková and George’s method in Table 1 of Ročková & George (2016). It also had an option for varimax rotation of the loadings matrix in the second version of Ročková and George’s method. The R code was provided to us by Veronika Ročková. There were two tuning parameters in Ročková and George’s method: λ_0 and λ_1 . We used $\lambda_1 = 0.001$ and $\lambda_0 = 30$ for both versions of Ročková and George’s method. These choices were based on the empirical results reported in Ročková & George (2016). Let $Y \in \mathfrak{R}^{n \times p}$ be the simulated data matrix. The following R code was used

to estimate the loadings matrix, Λ_{RG} , and its rank, r_{RG} , using the first version of Ročková and George's method:

1. $n = nrow(\text{train}); G = ncol(\text{train}); p = 10; K = 20; \alpha = 1/G; \text{epsilon} = 0.05$
2. $\text{lambda1} = 0.001; \text{startB} = \text{matrix}(\text{rnorm}(G*K), G, K)$
3. $\text{start} = \text{list}(B = \text{startB}, \text{sigma} = \text{rep}(1, p), \text{theta} = \text{rep}(0.5, K))$
4. $\text{lambda0} = 5; \text{result}_5 = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{start}, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 0)$
5. $\text{lambda0} = 10; \text{result}_{10} = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{result}_5, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 0)$
6. $\text{lambda0} = 20; \text{result}_{20} = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{result}_{10}, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 0)$
7. $\text{lambda0} = 30; \text{result}_{30} = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{result}_{20}, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 0)$
8. $r_{RG} = \text{sum}(\text{colSums}(\text{abs}(\text{result}_{30}\$B)) > 0)$
9. $\Lambda_{RG} = \text{result}_{30}\$B[, \text{rev}(\text{order}(\text{sqrt}(\text{colSums}((\text{result}_{30}\$B)^2)))]$

The following R code was used to estimate the loadings matrix, Λ_{RG+} , and its rank, r_{RG+} , using the second version of Ročková and George's method:

1. $n = nrow(\text{train}); G = ncol(\text{train}); p = 10; K = 20; \alpha = 1/G; \text{epsilon} = 0.05$
2. $\text{lambda1} = 0.001; \text{startB} = \text{matrix}(\text{rnorm}(G*K), G, K)$
3. $\text{start} = \text{list}(B = \text{startB}, \text{sigma} = \text{rep}(1, p), \text{theta} = \text{rep}(0.5, K))$
4. $\text{lambda0} = 5; \text{result}_5 = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{start}, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 1)$
5. $\text{lambda0} = 10; \text{result}_{10} = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{result}_5, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 1)$
6. $\text{lambda0} = 20; \text{result}_{20} = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{result}_{10}, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 1)$
7. $\text{lambda0} = 30; \text{result}_{30} = \text{FACTOR_ROTATE}(Y, \text{lambda0}, \text{lambda1}, \text{result}_{20}, K, \text{epsilon}, \alpha, \text{TRUE}, \text{TRUE}, 100, 1)$
8. $r_{RG+} = \text{sum}(\text{colSums}(\text{abs}(\text{result}_{30}\$B)) > 0)$
9. $\Lambda_{RG+} = \text{result}_{30}\$B[, \text{rev}(\text{order}(\text{sqrt}(\text{colSums}((\text{result}_{30}\$B)^2)))]$

The complete R code used for data analysis, including the code for *FACTOR_ROTATE* function, are available online.

REFERENCES

- ARMAGAN, A., DUNSON, D. B. & LEE, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23**, 119–143.
- CHEN, J. & CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.
- DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**, 1–38.
- FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.
- GEYER, C. J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics*, 1993–2010.
- HIROSE, K., YAMAMOTO, M. & NAGATA, H. (2015). *fanc: Penalized Likelihood Factor Analysis via Nonconvex Penalty*. R package version 1.25.
- KNEIP, A. & SARDA, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics* **39**, 2410–2447.

- KNIGHT, K. & FU, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics* , 1356–1378.
- R DEVELOPMENT CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- ROČKOVÁ, V. & GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association* **111**, 1608–1622.
- VAN DER VAART, A. W. (2000). *Asymptotic Statistics*, vol. 3. Cambridge University Press.
- WITTEN, D., TIBSHIRANI, R. J., GROSS, S. & NARASIMHAN, B. (2013). *PMA: Penalized Multivariate Analysis*. R package version 1.0.9.
- ZAHN, J. M., POOSALA, S., OWEN, A. B., INGRAM, D. K., LUSTIG, A., CARTER, A., WEERARATNA, A. T., TAUB, D. D., GOROSPE, M., MAZAN-MAMCZARZ, K. et al. (2007). AGEMAP: a gene expression database for aging in mice. *PLoS Genet* **3**, e201.
- ZOU, H. & LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533.