Biometrika (2017), **104**, 3, *pp.* 649–663
Printed in Great Britain

doi: 10.1093/biomet/asx030
Advance Access publication 16 June 2017

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

Summary

Bayesian sparse factor models have proven useful for characterizing dependence in multivariate data, but scaling computation to large numbers of samples and dimensions is problematic. We propose expandable factor analysis for scalable inference in factor models when the number of factors is unknown. The method relies on a continuous shrinkage prior for efficient maximum a posteriori estimation of a low-rank and sparse loadings matrix. The structure of the prior leads to an estimation algorithm that accommodates uncertainty in the number of factors. We propose an information criterion to select the hyperparameters of the prior. Expandable factor analysis has better false discovery rates and true positive rates than its competitors across diverse simulation settings. We apply the proposed approach to a gene expression study of ageing in mice, demonstrating superior results relative to four competing methods.

Some key words: Expectation-maximization algorithm; Factor analysis; Shrinkage prior; Sparsity; Variable selection.

1. Introduction

Factor analysis is a popular approach to modelling covariance matrices. Letting k^*, p and Ω denote the true number of factors, number of dimensions and $p \times p$ covariance matrix, respectively, factor models set $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$, where $\Lambda \in \mathbb{R}^{p \times k^*}$ is the loadings matrix and Σ is a diagonal matrix of positive residual variances. To allow computation to scale to large p, Λ is commonly assumed to be of low rank and sparse. These assumptions imply that $k^* \ll p$ and the number of nonzero loadings is small. A practical problem is that k^* and the locations of zeros in Λ are unknown. A number of Bayesian approaches exist to model this uncertainty in k^* and sparsity (Carvalho et al., 2008; Knowles & Ghahramani, 2011), but conventional approaches that rely on posterior sampling are intractable for large sample sizes n and dimensions p. Continuous shrinkage priors have been proposed that lead to computationally efficient sampling algorithms

(Bhattacharya & Dunson, 2011), but the focus is on estimating Ω, with Λ treated as a non-identifiable nuisance parameter. Our goal is to develop a computationally tractable approach for inference on Λ that models the uncertainty in k^* and the locations of zeros in Λ. To do this, we propose a novel shrinkage prior and a corresponding class of efficient inference algorithms for factor analysis.

Penalized likelihood methods provide computationally efficient approaches for point estimation of Λ and Σ. If k^* is known, then many such methods exist (Kneip & Sarda, 2011; Bai & Li, 2012). Sparse principal components analysis estimates a sparse Λ assuming $\Sigma = \sigma^2 I_p$, where I_p is the $p \times p$ identity matrix (Jolliffe et al., 2003; Zou et al., 2006; Shen & Huang, 2008; Witten et al., 2009). The assumptions of spherical residual covariance and known k^* are restrictive in practice. There are several approaches to estimating k^*. In econometrics, it is popular to rely on test statistics based on the eigenvalues of the empirical covariance matrix (Onatski, 2009; Ahn & Horenstein, 2013). It is also common to fit the model for different choices of k^*, and choose the best value based on an information criterion (Bai & Ng, 2002). Recent approaches instead use the trace norm or the sum of column norms of Λ as a penalty in the objective function to estimate k^* (Caner & Han, 2014). Alternatively, Ročková & George (2016) use a spike-and-slab prior to induce sparsity in Λ with an Indian buffet process allowing uncertainty in k^*; a parameter-expanded expectation-maximization algorithm is then used for estimation.

We propose a Bayesian approach for estimation of a low-rank and sparse Λ, allowing k^* to be unknown. Our approach relies on a novel multi-scale generalized double Pareto prior, inspired by the generalized double Pareto prior for variable selection (Armagan et al., 2013) and by the multiplicative gamma process prior for loadings matrices (Bhattacharya & Dunson, 2011). The latter approach focuses on estimation of Ω, but does not explicitly estimate Λ or k^*. The proposed prior leads to an efficient and scalable computational algorithm for obtaining a sparse estimate of Λ with appealing practical and theoretical properties. We refer to our method as expandable factor analysis because it allows the number of factors to increase as more dimensions are added and as p increases.

Expandable factor analysis combines the representational strengths of Bayesian approaches with the computational benefits of penalized likelihood methods. The multi-scale generalized double Pareto prior is concentrated near low-rank matrices; in particular, a high probability is placed around matrices with rank $O(\log p)$. Local linear approximation of the penalty imposed by the prior equals a sum of weighted ℓ_1 penalties on the elements of Λ. This facilitates maximum a posteriori estimation of a sparse Λ using an extension of the coordinate descent algorithm for weighted ℓ_1-regularized regression (Zou & Li, 2008). The hyperparameters of our prior are selected using a version of the Bayesian information criterion for factor analysis. Under the theoretical set-up for high-dimensional factor analysis in Kneip & Sarda (2011), we show that the estimates of loadings are consistent and that the estimates of nonzero loadings are asymptotically normal.

2. Expandable factor analysis

2·1. *Factor analysis model*

Consider the usual factor model. Let $Y \in \mathbb{R}^{n \times p}$, $Z \in \mathbb{R}^{n \times k^*}$ and $E \in \mathbb{R}^{n \times p}$ be the mean-centred data matrix, latent factor matrix and residual error matrix, respectively, where Z and E are unknown. We use index i for samples, index d for dimensions, and index j for factors. If $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is the residual error variance matrix, then the factor model for y_{id} is

$$y_{id} = \sum_{j=1}^{k^*} z_{ij}\lambda_{dj} + e_{id}, \quad z_{ij} \sim N(0,1), \quad e_{id} \mid \sigma_d^2 \sim N(0,\sigma_d^2), \tag{1}$$

where z_{ij} and e_{id} are independent ($i = 1,\ldots,n$; $j = 1,\ldots,k^*$; $d = 1,\ldots,p$). Equivalently,

$$y_i = \Lambda z_i + e_i, \quad y_i = (y_{i1},\ldots,y_{ip})^{\mathrm{T}}, \quad z_i = (z_{i1},\ldots,z_{ik^*})^{\mathrm{T}}, \quad e_i = (e_{i1},\ldots,e_{ip})^{\mathrm{T}} \tag{2}$$

for sample i and $\mathrm{cov}(y_i) = \Lambda\Lambda^{\mathrm{T}} + \Sigma$ ($i = 1,\ldots,n$). Similarly, model (1) reduces to regression in the space of latent factors

$$y_d = Z\lambda_d + e_d, \quad y_d = (y_{1d},\ldots,y_{nd})^{\mathrm{T}}, \quad \lambda_d = (\lambda_{d1},\ldots,\lambda_{dk^*})^{\mathrm{T}}, \quad e_d = (e_{1d},\ldots,e_{nd})^{\mathrm{T}} \tag{3}$$

for dimension d ($d = 1,\ldots,p$). Unlike usual regression, the design matrix Z in (3) is unknown.

Penalized estimation of Λ is typically based on (2) or (3). The loss is estimated as the regression-type squared error after imputing Z using the eigendecomposition of the empirical covariance matrix $Y^{\mathrm{T}}Y/n$ or an expectation-maximization algorithm. The choice of penalty on Λ presents a variety of options. If the goal is to select factors that affect any of the p variables, then the sum of column norms of Λ can be used as a penalty; a recent example is the group bridge penalty, $\sum_{j=1}^{k}(\sum_{d=1}^{p}\lambda_{dj}^2/p)^{\alpha}$, where $0 < \alpha < 1/2$ and k is an upper bound on k^*. The selected factors correspond to the nonzero columns of the estimated Λ (Caner & Han, 2014). To further obtain elementwise sparsity, a nonconcave variable selection penalty can be applied to the elements in Λ. The estimate of Λ depends on the choice of criterion for selecting the tuning parameters (Hirose & Yamamoto, 2015).

Our expandable factor analysis differs from this typical approach in several important ways. We start from a Bayesian perspective, and place a prior on Λ that is structured to allow uncertainty in k^* while shrinking towards loadings matrices with many zeros and $k^* \ll p$. If k is an upper bound on k^*, then the prior is designed to automatically allow a slow rate of growth in k as the number of dimensions p increases by concentrating in neighbourhoods of matrices with rank bounded above by $k = O(\log p)$. To our knowledge, this is a unique feature of our approach, justifying its name. Expandability is an appealing characteristic, as more factors should be needed to accurately model the dependence structure as the dimension of the data increases.

2·2. *Multi-scale generalized double Pareto prior*

We would like to design a prior on Λ such that maximum a posteriori estimates of Λ have the following four characteristics:

(a) the estimate of a loading with large magnitude should be nearly unbiased;
(b) a thresholding rule, such as soft-thresholding, is used to estimate the loadings so that loadings estimates with small magnitudes are automatically set to zero;
(c) the estimator of any loading is continuous in the data to limit instability; and
(d) the ℓ_2-norm of the ith column of the estimated Λ does not increase as i increases.

The first three properties are related to nonconcave variable selection (Fan & Li, 2001). Properties (b) and (d) together ensure existence of a column index after which all estimated loadings are identically zero. Automatic relevance determination and multiplicative gamma process priors satisfy (d) but fail to satisfy (b). No existing prior for loadings matrices satisfies properties (a)–(d) simultaneously (Carvalho et al., 2008; Bhattacharya & Dunson, 2011; Knowles & Ghahramani, 2011).

In order to satisfy these four properties and obtain a computationally efficient inference procedure, it is convenient to start with a prior for a loadings matrix $\Lambda \in \mathbb{R}^{p\times\infty}$ having infinitely many columns; in practice, all of the elements will be estimated to be zero after a finite column index that corresponds to the estimated number of factors. Bhattacharya & Dunson (2011) showed that the set of loadings matrices $\Lambda \in \mathbb{R}^{p\times\infty}$ that leads to well-defined covariance matrices is

$$\mathcal{C} = \left\{ \Lambda : \max_{1\leqslant d\leqslant p} \sum_{j=1}^{\infty} \lambda_{dj}^2 < \infty \right\}.$$

We propose a multi-scale generalized double Pareto prior for Λ having support on $\mathcal{C}$. This prior is constructed to concentrate near low-rank matrices, placing high probability around matrices with rank at most $k = O(\log p)$.

The multi-scale generalized double Pareto prior on Λ specifies independent generalized double Pareto priors on λ_{dj} $(d = 1, \ldots, p; j = 1, \ldots, \infty)$ so that the density of Λ is

$$p_{\text{mgdP}}(\Lambda) = \prod_{d=1}^{p}\prod_{j=1}^{\infty} p_{\text{gdP}}(\lambda_{dj} \mid \alpha_j, \eta_j), \quad p_{\text{gdP}}(\lambda_{dj} \mid \alpha_j, \eta_j) = \frac{\alpha_j}{2\eta_j}\left(1 + \frac{|\lambda_{dj}|}{\eta_j}\right)^{-(\alpha_j+1)}, \qquad (4)$$

where $p_{\text{gdP}}(\cdot \mid \alpha_j, \eta_j)$ is the generalized double Pareto density with parameters α_j and η_j (Armagan et al., 2013). This prior on λ_{dj} ensures that properties (a)–(c) are satisfied. Property (d) is satisfied by choosing parameter sequences α_j and η_j $(j = 1, \ldots, \infty)$ such that two conditions hold: the prior measure P_l on $\mathcal{C}$ has density p_{mgdP} in (4), and P_l has $\mathcal{C}$ as its support. These conditions hold for the form of α_j and η_j $(j = 1, \ldots, \infty)$ specified by the following lemma.

LEMMA 1. *If* $\alpha_j > 2$, $\eta_j/\alpha_j = O(j^{-m})$ $(j = 1, \ldots, \infty)$ *and* $m > 1/2$, *then* $P_l(\mathcal{C}) = 1$.

The proof is given in the Supplementary Material, along with the other proofs.

As in Bhattacharya & Dunson (2011), we truncate to a finite number of columns for tractable computation. This truncation is accomplished by mapping $\Lambda \in \mathcal{C}$ to $\Lambda^k \in \mathcal{C}$, with Λ^k retaining the first k columns of Λ. The choice of k is such that $\Omega^k = \Lambda^k {\Lambda^k}^{\mathrm{T}} + \Sigma$ is arbitrarily close to $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$, where distance between Ω^k and Ω is measured using the ℓ_∞-norm of their elementwise difference. In addition, for computational convenience, we assume that the hyperparameters α_j and η_j $(j = 1, \ldots, \infty)$ are analytic functions of the parameters δ and ρ, respectively, with these functions satisfying the conditions of Lemma 1.

The following lemma defines the forms of α_j and η_j $(j = 1, \ldots, \infty)$ in terms of δ and ρ.

LEMMA 2. *If* $\delta > 2$, $\rho > 0$, $\alpha_j(\delta) = \delta^j$ *and* $\eta_j(\rho) = \rho$ *for* $j = 1, \ldots, \infty$, *then* $P_l(\mathcal{C}) = 1$, *where* P_l *has density* p_{mgdP} *in* (4) *with hyperparameters* $\alpha_j(\delta)$ *and* $\eta_j(\rho)$ $(j = 1, \ldots, \infty)$. *Furthermore, given* $\epsilon > 0$, *there exists a positive integer* $k(p, \delta, \epsilon) = O\{\log^{-1}\delta \log(p/\epsilon^2)\}$ *for every* Ω *such that for all* $r \geqslant k$, $\alpha_j(\delta) = \delta^j$, $\eta_j(\rho) = \rho$ $(j = 1, \ldots, r)$ *and* $\Omega^r = \Lambda^r {\Lambda^r}^{\mathrm{T}} + \Sigma$, *we have that* $\text{pr}\{\Omega^r \mid d_\infty(\Omega, \Omega^r) < \epsilon\} > 1 - \epsilon$ *where* $d_\infty(A, B) = \max_{1\leqslant i,j\leqslant p}|a_{ij} - b_{ij}|$.

The penalty imposed on the loadings by the prior grows exponentially with δ as the column index increases. This property of the prior ensures that all the loadings are estimated to be zero after a finite column index, which corresponds to the estimated number of factors.

3. Estimation algorithm

3·1. *Expectation-maximization algorithm*

We rely on an adaptation of the expectation-maximization algorithm to estimate Λ and Σ. Choose a positive integer k of order $\log p$ as the upper bound on k^*; the estimate of the number of factors will be less than or equal to k. The results are not sensitive to the choice of k due to the properties of the multi-scale generalized double Pareto prior, provided k is sufficiently large. If k is too small, then the estimated number of factors will be equal to the upper bound, suggesting that this bound should be increased. Given k, define $\alpha_j(\delta)$ and $\eta_j(\rho)$ $(j = 1, \ldots, k)$ as in Lemma 2, with $\delta > 2$ and $\rho > 0$ being prespecified constants.

We present the objective function as a starting point for developing the coordinate descent algorithm and provide derivations in the Supplementary Material. Let $F^{(t)} = n^{-1}E(Z^{\mathrm{T}}Z \mid Y, \Lambda^{(t)}, \Sigma^{(t)})$ and $L^{(t)} = n^{-1}E(\sum_{i=1}^{n} y_i z_i^{\mathrm{T}} \mid Y, \Lambda^{(t)}, \Sigma^{(t)})$, where the superscript (t) denotes an estimate at iteration t and $E(\cdot \mid Y, \Lambda^{(t)}, \Sigma^{(t)})$ denotes the conditional expectation given Y, $\Lambda^{(t)}$ and $\Sigma^{(t)}$ based on (1). The objective function for parameter updates in iteration $t+1$ is

$$\underset{\substack{\lambda_d, \sigma_d^2 \\ d=1,\ldots,p}}{\arg\min} \sum_{d=1}^{p} \Bigg(\frac{n+2}{2npk} \log \sigma_d^2 + \frac{\|w_d^{(t)} - X^{(t)}\lambda_d\|^2 - w_d^{(t)\mathrm{T}} w_d^{(t)} + (Y^{\mathrm{T}}Y/n)_{dd}}{2pk\sigma_d^2} + \left[\sum_{j=1}^{k} \frac{\alpha_j(\delta)+1}{npk} \log\left\{1 + \frac{|\lambda_{dj}|}{\eta_j(\rho)}\right\}\right]\Bigg), \tag{5}$$

where $X^{(t)} = F^{(t)1/2}$ and $w_d^{(t)} = F^{(t)-1/2} l_d^{(t)}$ $(d = 1, \ldots, p)$.

3·2. *Estimating parameters using a convex objective function*

The objective (5) is written as a sum of p terms. The dth term corresponds to the objective function for the regularized estimation of the dth row of the loadings matrix, λ_d^{T}, with a specific form of log penalty on λ_d (Zou & Li, 2008). Local linear approximation at $\lambda_{dj}^{(t)}$ of the log penalty on λ_{dj} in (5) implies that each row of Λ is estimated separately at iteration $t+1$:

$$\lambda_d^{\mathrm{lla}(t+1)} = \underset{\lambda_d}{\arg\min} \frac{\|w_d^{(t)} - X^{(t)}\lambda_d\|^2}{2pk\sigma_d^{2(t)}} + \sum_{j=1}^{k} \frac{\alpha_j(\delta)+1}{npk\{\eta_j(\rho) + |\lambda_{dj}^{(t)}|\}} |\lambda_{dj}| \quad (d = 1, \ldots, p). \tag{6}$$

This problem corresponds to regularized estimation of regression coefficients λ_d with $w_d^{(t)}$ as the response, $X^{(t)}$ as the design matrix, $\sigma_d^{2(t)}$ as the error variance, and a weighted ℓ_1 penalty on λ_d.

The solution to (6) is found using block coordinate descent. Let column j of F and row d of Λ without the jth element be written as $f_{(-j),j}$ and $\lambda_{d,(-j)}^{\mathrm{T}}$. Then the update to estimate λ_d^{lla} is

$$\lambda_{dj}^{\mathrm{lla}(t+1)} = \frac{\mathrm{sign}(\tilde{\lambda}_{dj}^{(t)})}{f_{jj}^{(t)}} \big(|\tilde{\lambda}_{dj}^{(t)}| - c_{dj}^{(t)}\big)_+, \quad c_{dj}^{(t)} = \frac{\sigma_d^{2(t)}\{\alpha_j(\delta)+1\}}{n\{\eta_j(\rho) + |\lambda_{dj}^{(t)}|\}} \quad (j = 1, \ldots, k), \tag{7}$$

where $\tilde{\lambda}_{dj}^{(t)} = l_{dj}^{(t)} - \lambda_{d,(-j)}^{\mathrm{lla}(t)\,\mathrm{T}} f_{(-j),j}^{(t)}$ and $(x)_+ = \max(x, 0)$. Fix λ_d at $\lambda_d^{\mathrm{lla}(t+1)}$ in (5) to update σ_d^2 in iteration $t+1$ as

$$\sigma_d^{2(t+1)} = \frac{n}{n+2}\big\{(Y^{\mathrm{T}}Y/n)_{dd} + \lambda^{\mathrm{lla}(t+1)\,\mathrm{T}} F^{(t)} \lambda^{\mathrm{lla}(t+1)} - 2 l_d^{(t)\mathrm{T}} \lambda^{\mathrm{lla}(t+1)}\big\}. \tag{8}$$

If any root-n-consistent estimate of λ_{dj} is used instead of $|\lambda_{dj}^{(t)}|$ in (6), then it acts as a warm starting point for the estimation algorithm. This leads to a consistent estimate of λ_{dj} in one step of coordinate descent (Zou & Li, 2008). An implementation of this approach for known values of δ and ρ is summarized in steps (i)–(iv) of Algorithm 1 using the R (R Development Core Team, 2017) package glmnet (Friedman et al., 2010).

Algorithm 1. Estimation algorithm for expandable factor analysis.

```
Notation:
1. diag($A$) is the diagonal matrix containing diagonal elements of a symmetric matrix $A$.
2. Chol($A$) is the upper triangular Cholesky factorization of a symmetric positive-definite
   matrix $A$.
3. bdiag($A_1, \ldots, A_p$) is a block-diagonal matrix with $A_1, \ldots, A_p$ forming the diagonal
   blocks.
4. $\mathrm{vec}(A) = (a_1^{\mathrm{T}}, \ldots, a_d^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{cd \times 1}$, where $A \in \mathbb{R}^{c \times d}$.

Input:
1. Data $Y \in \mathbb{R}^{n \times p}$ and upper bound $k = O(\log p)$ on the rank of the loadings matrix.
2. The $\delta$-$\rho$ grid with $RS$ grid indices ($\delta_1 < \cdots < \delta_R$; $\rho_1 < \cdots < \rho_S$).

Do:
1. Centre data about their mean $\hat{y}_{ij} = y_{ij} - (n^{-1}) \sum_{m=1}^{n} y_{mj}$ ($i = 1, \ldots, n; j = 1, \ldots, p$).
2. Let $S_{\hat{y}\hat{y}} = \hat{Y}^{\mathrm{T}}\hat{Y}/n$. Then estimate eigenvalues and eigenvectors of $S_{\hat{y}\hat{y}}$: $\hat{\zeta}_1, \ldots, \hat{\zeta}_p$ and
   $\hat{\psi}_1, \ldots, \hat{\psi}_p$.
3. Define $\Lambda^0$ to be the matrix $\{(\hat{\zeta}_1)^{1/2}\hat{\psi}_1, \ldots, (\hat{\zeta}_k)^{1/2}\hat{\psi}_k\}$.
4. Begin estimation of $\Lambda$, $\Sigma$ and $\pi$ across the $\delta$-$\rho$ grid:
     For $r = 1, \ldots, R$
       For $s = S, \ldots, 1$
         (i) Define $\alpha_j = \delta_r^j$, $\eta_j = \rho_s p^{1/2}$ if $n \leqslant p$, and $\eta_j = \rho_s$ if $n > p$ ($j = 1, \ldots, k$).
         (ii) Initialize the following statistics required in (7):

   $\Sigma^0 = \mathrm{diag}(S_{\hat{y}\hat{y}} - \Lambda^0\Lambda^{0\mathrm{T}})$,  $\Omega^0 = \Lambda^0\Lambda^{0\mathrm{T}} + \Sigma^0$,  $G^0 = (\Omega^0)^{-1}\Lambda^0$,  $L^0 = S_{\hat{y}\hat{y}}G^0$,
   $\Delta^0 = I_k - \Lambda^{0\mathrm{T}}G^0$,  $F^0 = \Delta^0 + G^{0\mathrm{T}}S_{\hat{y}\hat{y}}G^0$,  $R^0 = \mathrm{Chol}(F^0)\cdot$

         (iii) Define $X \in \mathbb{R}^{pk \times pk}$, $w \in \mathbb{R}^{pk \times 1}$, $y \in \mathbb{R}^{pk \times 1}$ and $v \in \mathbb{R}^{pk \times 1}$ required to
               solve (6):

   $X = \mathrm{bdiag}(\underbrace{R^0, \ldots, R^0}_{p \text{ times}})$,  $w = \{\underbrace{(\Sigma^0)_{11}^{-1}, \ldots, (\Sigma^0)_{11}^{-1}}_{k \text{ times}}, \ldots, \underbrace{(\Sigma^0)_{pp}^{-1}, \ldots, (\Sigma^0)_{pp}^{-1}}_{k \text{ times}}\}$,
   $y = \mathrm{vec}\big\{(R^0)^{-1}L^{0\mathrm{T}}\big\}$,  $v = \frac{1}{npk}\left(\frac{\alpha_1+1}{\eta_1+|\lambda_{11}^0|}, \ldots, \frac{\alpha_k+1}{\eta_k+|\lambda_{1k}^0|}, \ldots, \frac{\alpha_1+1}{\eta_1+|\lambda_{p1}^0|}, \ldots, \frac{\alpha_k+1}{\eta_k+|\lambda_{pk}^0|}\right)\cdot$
```

```
            (iv) Estimate $\Lambda^{\text{lla}}$ in (7) and $\Sigma^{\text{lla}}$ in (8) using the R package glmnet in three steps:
              • result $\leftarrow$ glmnet$\big($x = $X$, y = $y$, weights = $w$, intercept = FALSE,
                        standardize = FALSE, penalty.factor = $v/\sum_{j=1}^{pk} v_j\big)$.
              • vec$\big(\Lambda^{\text{lla}^{\text{T}}}\big)$ $\leftarrow$ coef$\big($result, s = $\sum_{j=1}^{pk} v_j$, exact = TRUE$\big)$ [-1, ].
              • $(\Sigma^{\text{lla}})_{dd} \leftarrow \{n/(n+2)\}\big\{(S_{\hat{y}\hat{y}})_{dd} + \lambda_d^{\text{lla}^{\text{T}}} F^0 \lambda_d^{\text{lla}} - 2 l_d^{0^{\text{T}}} \lambda_d^{\text{lla}}\big\}$ $(d = 1, \ldots, p)$.
            (v) Set $\Lambda^{(r,s)} = \Lambda^{\text{lla}}$, $\Sigma^{(r,s)} = \Sigma^{\text{lla}}$, $\Lambda^0 = \Lambda^{\text{lla}}$, and estimate the posterior weight
                $\pi^{(r,s)}$ in (10).
        End for.
        Set $\Lambda^0 = \Lambda^{(r,S)}$.
      End for.
  5. Obtain grid index $(\hat{r}, \hat{s})$ for the estimate of $(\delta, \rho)$, where $\pi^{(\hat{r},\hat{s})} = \max_{(r,s)} \pi^{(r,s)}$.

  Return:
    $\Lambda^{(\hat{r},\hat{s})}$, $\Sigma^{(\hat{r},\hat{s})}$ and $\mathcal{M}^{(\hat{r},\hat{s})} = \{(d,j) : \lambda_{dj}^{(\hat{r},\hat{s})} \neq 0;\ d = 1, \ldots, p;\ j = 1, \ldots, k\}$.
```

The estimate of Λ obtained using (7) satisfies properties (a)–(d) described earlier. The adaptive threshold $c_{dj}^{(t)}$ in (7) ensures that property (a) is satisfied. The soft-thresholding rule to estimate λ_{dj} ensures that property (b) is satisfied. The local linear approximation (6) has continuous first derivatives in the parameter space excluding zero, so property (c) is also satisfied (Zou & Li, 2008). The Λ estimate satisfies property (d) due to the structured penalty imposed by the prior.

We comment briefly on the choice of prior and uncertainty quantification. We build on the generalized double Pareto prior instead of other shrinkage priors not only because the estimate of Λ satisfies properties (a)–(d), but also because local linear approximation of the resulting penalty has a weighted ℓ_1 form. We exploit this for efficient computations and use a warm starting point to estimate a sparse Λ in one step using Algorithm 1. Uncertainty estimates of the nonzero loadings are obtained from Laplace approximation, and the remaining loadings are estimated as zero without uncertainty quantification.

3·3. *Root-n-consistent estimate of* λ_{dj}

The root-n-consistent estimate of λ_{dj} exists under Assumptions A0–A4 given in the Appendix. If $\hat{\zeta}_d$ and $\hat{\psi}_d$ $(d = 1, \ldots, p)$ are the eigenvalues and eigenvectors of the empirical covariance matrix $Y^{\text{T}}Y/n$, then $\sum_{d=1}^{p} \hat{\zeta}_d \hat{\psi}_d \hat{\psi}_d^{\text{T}}$ is the eigendecomposition of $Y^{\text{T}}Y/n$. It is known that $(\hat{\zeta}_d)^{1/2}\hat{\psi}_{dj}$ is a root-n-consistent estimator of λ_{dj} if p is fixed and $n \to \infty$. If $n \to \infty$, $n \leqslant p \to \infty$ and $\log p/n \to 0$, then $p^{-1/2}(\hat{\zeta}_d)^{1/2}\hat{\psi}_{dj}$ is a root-n-consistent estimator of $p^{-1/2}\lambda_{dj}$; see the Supplementary Material for a proof. Scaling by $p^{1/2}$ is required because the largest eigenvalue of Ω tends to infinity as $p \to \infty$ (Kneip & Sarda, 2011). This scaling does not change our estimation algorithm for λ_{dj} in (7), except that $\eta_j(\rho)$ is changed to $\eta_j(\rho)p^{1/2}$ $(j = 1, \ldots, k)$.

3·4. *Bayesian information criterion to select* δ *and* ρ

The parameter estimates in (7) and (8) depend on the hyperparameters through δ and ρ, both of which are unknown. To estimate δ and ρ, we use a grid search. Let $\delta_1 < \cdots < \delta_R$ and $\rho_1 < \cdots < \rho_S$ form a δ-ρ grid. If (δ_r, ρ_s) is the value of $(\delta,\ \rho)$ at grid index (r, s), then $\alpha_j(\delta_r)$

and $\eta_j(\rho_s)$ $(j = 1, \ldots, k)$ are the hyperparameters of our prior defined using Lemma 2, and $\Lambda^{(r,s)}$ and $\Sigma^{(r,s)}$ are the parameter estimates based on this prior. Algorithm 1 first estimates $\Lambda^{(r,s)}$ and $\Sigma^{(r,s)}$ for every (r, s) by choosing warm starting points and then estimates (δ, ρ) using all the estimated Λ and Σ. These two steps in the estimation of (δ, ρ) are described next.

The structured penalty imposed by our prior implies that $\Lambda^{(1,S)}$ has the maximum number of nonzero loadings. Algorithm 1 exploits this structure by first estimating $\Lambda^{(1,S)}$ and then other loadings matrices along the δ-ρ grid by successively thresholding nonzero loadings in $\Lambda^{(1,S)}$ to 0. Let $\mathcal{M}^{(r,s)} = \{(d,j) : \lambda_{dj}^{(r,s)} \neq 0;\ d = 1, \ldots, p;\ j = 1, \ldots, k\}$ be the set that contains the locations of nonzero loadings in $\Lambda^{(r,s)}$. The estimation path of Algorithm 1 across the δ-ρ grid is such that $\mathcal{M}^{(r,1)} \subseteq \cdots \subseteq \mathcal{M}^{(r,S)}$ $(r = 1, \ldots, R)$ and $\mathcal{M}^{(R,S)} \subseteq \cdots \subseteq \mathcal{M}^{(1,S)}$.

After the estimation of $\Lambda^{(r,s)}$ and $\Sigma^{(r,s)}$ $(r = 1, \ldots, R;\ s = 1, \ldots, S)$, (δ, ρ) is set to $(\delta_{\hat{r}}, \rho_{\hat{s}})$ if $\mathcal{M}^{(\hat{r},\hat{s})}$ has the maximum posterior probability. Let $|A|$ be the cardinality of set A. Given (δ_r, ρ_s), there are $\binom{pk}{|\mathcal{M}^{(r,s)}|}$ loadings matrices that have $|\mathcal{M}^{(r,s)}|$ nonzero loadings but differ in the locations of the nonzero loadings. Assuming that each of these matrices is equally likely to represent the locations of nonzero loadings in the true loadings matrix, the prior for $\mathcal{M}^{(r,s)}$ is

$$\mathrm{pr}(\mathcal{M}^{(r,s)} \mid \delta_r, \rho_s) \propto \binom{pk}{|\mathcal{M}^{(r,s)}|}^{-1} \quad (r = 1, \ldots, R;\ s = 1, \ldots, S). \tag{9}$$

Let $\pi^{(r,s)}$ be the posterior probability of $\mathcal{M}^{(r,s)}$. Then an asymptotic approximation to $-2 \log \pi^{(r,s)}$ is

$$-2 \log f(Y, \Lambda^{(r,s)} \mid \delta_r, \rho_s) + |\mathcal{M}^{(r,s)}| \log n + 2\, |\mathcal{M}^{(r,s)}| \log(pk) \tag{10}$$

if terms of order smaller than $\log n + \log p$ are ignored, where $f(Y, \Lambda \mid \delta_r, \rho_s)$ is the joint density of Y and Λ based on (1). The first term in (10) measures the goodness-of-fit, and the last two terms penalize complexity of a factor model with n samples and pk loadings with the locations of nonzero loadings in $\mathcal{M}^{(r,s)}$. Theorem 3 in the next section shows that $-2 \log \pi^{(r,s)}$ and $\mathrm{EBIC}_\gamma(\mathcal{M}^{(r,s)})$ have the same asymptotic order under certain regularity assumptions, where EBIC_γ is the extended Bayesian information criteria of Chen & Chen (2008) and $0 \leqslant \gamma \leqslant 1$ is an unknown constant. The analytic forms of $-2 \log \pi^{(r,s)}$ and $\mathrm{EBIC}_\gamma(\mathcal{M}^{(r,s)})$ are the same when $\gamma = 0{\cdot}5$ and terms of order smaller than $\log n + \log p$ are ignored, so we use $\mathrm{EBIC}_{0{\cdot}5}$ for estimating $\mathcal{M}^{(\hat{r},\hat{s})}$ in our numerical experiments.

4. Theoretical properties

Let Λ_n^{lla} and Σ_n^{lla} be the fixed points of $\Lambda^{\mathrm{lla}(t)}$ and $\Sigma^{\mathrm{lla}(t)}$. The updates (7) and (8) define the map $g : \theta^{(t)} \mapsto \theta^{(t+1)}$, where $\theta = (\Lambda, \Sigma)$. The following theorem shows that our estimation algorithm retains the convergence properties of the expectation-maximization algorithm.

Theorem 1. *If $\mathcal{L}(\theta)$ represents the objective (5), then $\mathcal{L}(\theta)$ does not decrease at every iteration. Let Q be the local linear approximation of (5). Assume that $Q(\theta) = Q\{g(\theta)\}$ only for stationary points of Q; then the sequence $\{\theta^{(t)}\}_{t=1}^{\infty}$ converges to its stationary point θ_n^{lla}.*

Let Λ^* be the true loadings matrix and Σ^* the residual variance matrix. We define $\lambda_{dj}^* = 0$ $(d = 1, \ldots, p;\ j = k^* + 1, \ldots, k)$ and express Λ^* as having k columns. The locations of true nonzero loadings are in the set $\mathcal{M}^* = \{(d,j) : \lambda_{dj}^* \neq 0;\ d = 1, \ldots, p;\ j = 1, \ldots, k\}$. Let $\hat{\Lambda}$ and $\hat{\Sigma}$

be the estimates of Λ and Σ obtained using our estimation algorithm for a specific choice of $\alpha_j(\delta)$ and $\eta_j(\rho)$ $(j = 1,\ldots,k)$; then $\hat{\mathcal{M}} = \{(d,j) : \hat{\lambda}_{dj} \neq 0;\ d = 1,\ldots,p;\ j = 1,\ldots,k\}$ is an estimator of $\mathcal{M}^*$. If $\hat{\lambda} = \mathrm{vec}(\hat{\Lambda}^{\mathrm{T}})$ and $\lambda^* = \mathrm{vec}(\Lambda^{*\mathrm{T}})$, then $\hat{\lambda}_A$ and λ^*_A retain elements of $\hat{\lambda}$ and λ^* with indices in the set A. The following theorem specifies the asymptotic properties of $\hat{\Lambda}$, $\hat{\Sigma}$ and $\hat{\mathcal{M}}$.

Theorem 2. *Suppose that Assumptions A0–A6 in the Appendix hold and that $n \to \infty$, $n \leqslant p \to \infty$ and $\log p/n \to 0$. Then, for any $d = 1,\ldots,p$ and $j = 1,\ldots,k$:*

(i) *$\hat{\lambda}_{dj}$, $\hat{\sigma}^2_d$ and $\hat{\mathcal{M}}$ are consistent estimators of λ^*_{dj}, σ^{2*}_d and $\mathcal{M}^*$, respectively;*

(ii) *$n^{1/2}(\hat{\lambda}_{\mathcal{M}^*} - \lambda^*_{\mathcal{M}^*}) \to N_{|\mathcal{M}^*|}(0, C_*)$ and $n^{1/2}(\hat{\sigma}^2_d - \sigma^{2*}_d) \to N(0, c_*)$ in distribution, where C_* is a $|\mathcal{M}^*| \times |\mathcal{M}^*|$ symmetric positive-definite matrix and $c_* > 0$.*

Theorem 2 holds for any multi-scale generalized double Pareto prior with hyperparameters $\alpha_j(\delta)$ and $\eta_j(\rho)$ $(j = 1,\ldots,k)$ that satisfies Assumption A5. In practice, the estimate of Λ depends on the choice of δ and ρ. Restricting the search to the hyperparameters indexed along the δ-ρ grid, Algorithm 1 sets the values of the hyperparameters to $\alpha_j(\delta_{\hat{r}})$ and $\eta_j(\rho_{\hat{s}})$ $(j = 1,\ldots,k)$, where $\pi^{(r,s)}$ achieves its maximum at grid index $(\hat{r},\hat{s})$. The following theorem justifies this method of selecting hyperparameters and shows the asymptotic relationship between $-2\log\pi^{(r,s)}$ and $\mathrm{EBIC}_\gamma(\mathcal{M}^{(r,s)})$.

Theorem 3. *Suppose that the generalized double Pareto prior with hyperparameters defined using (δ_*, ρ_*) leads to estimation of $\mathcal{M}^*$. Let $\mathcal{M} \neq \mathcal{M}^*$ be another set that contains the locations of nonzero loadings in an estimated Λ for a given (δ, ρ). Define $\pi_{\mathcal{M}} = \mathrm{pr}(\mathcal{M} \mid Y)$ and $\pi_{\mathcal{M}^*} = \mathrm{pr}(\mathcal{M}^* \mid Y)$. If Assumptions A0–A7 in the Appendix hold, then for any $\mathcal{M}$ such that $|\mathcal{M}| \in \{1,\ldots,pk\}$:*

(i) *$-2\log\pi_{\mathcal{M}}/\mathrm{EBIC}_\gamma(\mathcal{M}) \to 1$ in probability as $n \to \infty$;*

(ii) *$\mathrm{pr}\{\max(\pi_{\mathcal{M}} : \mathcal{M} \neq \mathcal{M}^*) < \pi_{\mathcal{M}^*}\} \to 1$ as $n \to \infty$.*

Let $(\delta_{r^*}, \rho_{s^*})$ be a point on the δ-ρ grid that leads to estimation of $\mathcal{M}^*$. Then, Theorem 3 shows that Algorithm 1 selects $\mathcal{M}^*$ with probability tending to 1 because $\pi^{(r^*,s^*)}$ will be larger than any $\pi^{(r,s)}$ where (r,s) is such that $\mathcal{M}^{(r,s)} \neq \mathcal{M}^*$.

5. Data analysis

5·1. *Set-up and comparison metrics*

We compared our method with those of Caner & Han (2014), Hirose & Yamamoto (2015), Ročková & George (2016) and Witten et al. (2009). The first competitor was developed to estimate the rank of Λ, and the last three competitors were developed to estimate Λ. We used two versions of Ročková and George's method. The first version uses the expectation-maximization algorithm developed in Ročková & George (2016), and the second version adds an extra step in every iteration of the algorithm that rotates the loadings matrix using the varimax criterion.

We evaluated the performance of the methods for estimating Λ on simulated data using the root mean square error, proportion of true positives, and proportion of false discoveries:

$$\text{mean square error} = \sum_{d=1}^{p}\sum_{j=1}^{k}(|\lambda^*_{dj}| - |\hat{\lambda}_{dj}|)^2/(pk), \quad \text{true positive rate} = |\hat{\mathcal{M}} \cap \mathcal{M}^*|/|\mathcal{M}^*|,$$

$$\text{false discovery rate} = |\hat{\mathcal{M}} \backslash \mathcal{M}^*|/|\hat{\mathcal{M}}|,$$

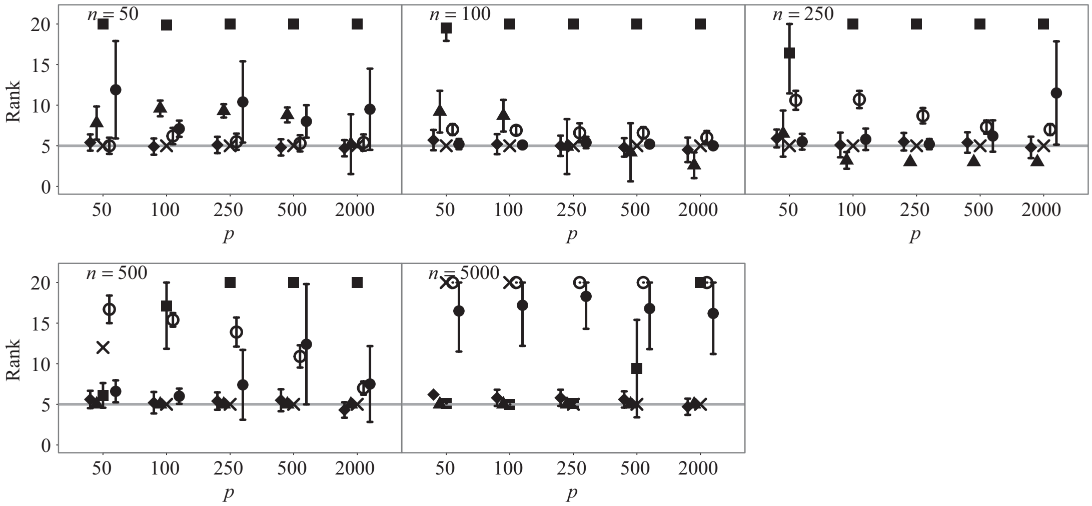

Fig. 1. Rank estimate averaged across simulation replications for the methods of Caner & Han (2014) (crosses), Hirose & Yamamoto (2015) (squares), Ročková & George (2016) varimax-free version (empty circles), Ročková & George (2016) varimax version (filled circles) and Witten et al. (2009) (diamonds), as well as our estimation algorithm (triangles). In each panel the horizontal grey line represents the true number of factors; error bars represent Monte Carlo errors.

where Λ^* and $\hat{\Lambda}$ are the true and estimated loadings matrices and $\mathcal{M}^*$ and $\hat{\mathcal{M}}$ are the true and estimated locations of nonzero loadings. We assume that $\lambda^*_{dj} = 0$ for any d and $j = k^*+1,\ldots,k$. Since λ^*_{dj} and $\hat{\lambda}_{dj}$ could differ in sign, mean square error compared their magnitudes.

5·2. *Simulated data analysis*

The simulation settings were based on examples in Kneip & Sarda (2011). The number of dimensions varied among $p = 50, 100, 250, 500, 2000$. The rank of every simulated loadings matrix was fixed at $k^* = 5$. The magnitudes of nonzero loadings in a column were equal and decreased as 10, 8, 6, 4 and 2 from the first to the fifth column. The signs of the nonzero loadings were chosen such that the columns of any loadings matrix were orthogonal, with a small fraction of overlapping nonzero loadings between adjacent columns:

$$\lambda^*_{dj} = \begin{cases} 2(6-j), & 1+(j-1)\frac{p}{k^*} \leqslant d \leqslant j\frac{p}{k^*}, \quad 1 \leqslant j \leqslant k^*, \\ -2(6-j), & 1+j\frac{p}{k^*} \leqslant d \leqslant (j+1)\frac{p}{k^*}, \quad 1 \leqslant j \leqslant k^*-1, \\ -2(6-j), & (j-1)\frac{p}{k^*} \leqslant d \leqslant j\frac{p}{k^*} - 1, \quad 2 \leqslant j \leqslant k^*, \\ 0, & \text{otherwise.} \end{cases}$$

The error variances σ^2_d increased linearly from 0·01 to 1 for $d = 1,\ldots,p$. With varying sample sizes $n = 50, 100, 250, 500, 5000$, data were simulated using model (1) for all combinations of n and p. The simulation set-up was replicated ten times and all five methods were applied in every replication by fixing the upper bound on the number of factors at 20. The δ-ρ grid had dimensions 20×20, and $\log_{10}\delta$ increased linearly from $\log_{10} 2$ to $\log_{10} 10$ while $\log_{10}\rho$ increased linearly from -3 to 3 when $n > p$ and from -2 to 6 when $n \leqslant p$.

All five methods had the same computational complexity of $O(p \log p)$ for one iteration, but their runtimes differed depending on their implementations, with the method of Witten et al. (2009) being the fastest. Figure 1 shows that Hirose and Yamamoto's method and both versions of Ročková and George's method significantly overestimated k^* for large p. The method of Witten

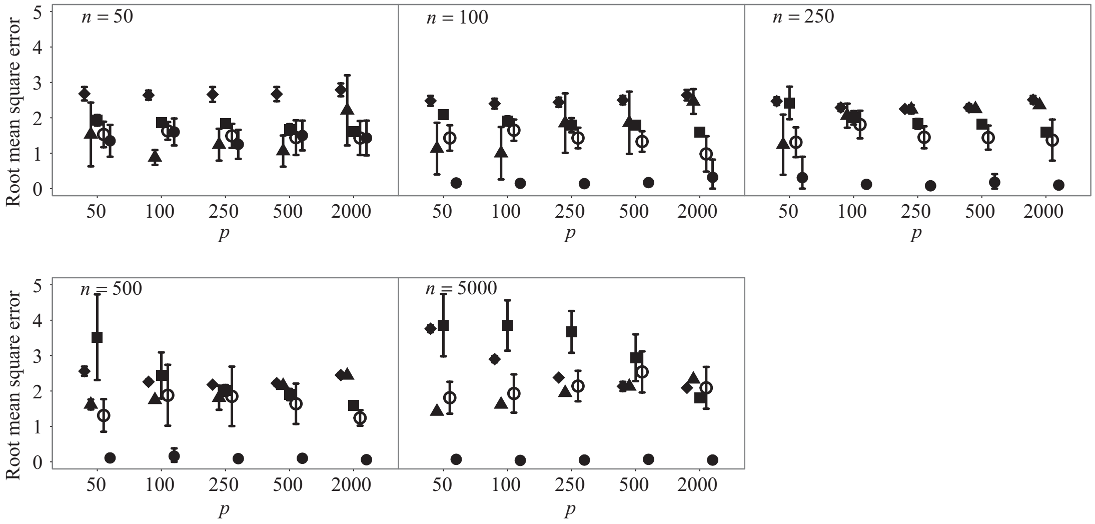

Fig. 2. Root mean square error averaged across simulation replications for the methods of Hirose & Yamamoto (2015) (squares), Ročková & George (2016) varimax-free version (empty circles), Ročková & George (2016) varimax version (filled circles) and Witten et al. (2009) (diamonds), as well as our estimation algorithm (triangles). Error bars represent Monte Carlo errors.

et al. slightly overestimated k^* across all settings. Caner and Han's method showed excellent performance and accurately estimated k^* across all simulation settings, except when $n = 5000$ and $p = 50$ or 100. When n was larger than 500, Assumption A4 was satisfied and our method accurately estimated k^* as 5 in every setting, performing better than Caner and Han's method when $n = 5000$.

The four methods for estimating Λ differed significantly in their root mean square errors, true positive rates, and false discovery rates; see Figs 2–4. Hirose and Yamamoto's method had the highest false discovery rates and the lowest true positive rates across most settings. Both versions of Ročková and George's method estimated an overly dense Λ across most settings, resulting in high true positive rates and high false discovery rates. The extra rotation step in the second version of Ročková and George's method resulted in excellent mean square error performance; however, varimax rotation is a post-processing step. A similar step to reduce the mean square error could be added to our method, for example by including a step to rotate the Λ^0 in step 3 of Algorithm 1 using the varimax criterion. When n and p were small, the method of Witten et al. achieved the lowest false discovery rates while our method achieved the highest true positive rates. When n and p were larger than 250 and 100, respectively, Assumption A4 was satisfied and our method simultaneously achieved the highest true positive rates and lowest false discovery rates while maintaining competitive mean square errors relative to the rotation-free methods.

5·3. *Microarray data analysis*

We used gene expression data on ageing in mice from the AGEMAP database (Zahn et al., 2007). There were 40 mice aged 1, 6, 16 and 24 months in this study. Each age group included five male and five female mice. Tissue samples were collected from 16 different tissues, including the cerebrum and cerebellum, for every mouse. Gene expression levels in every tissue sample were measured on a microarray platform. After normalization and removal of missing data, gene expression data were available for all 8932 probes across 618 microarrays. We used a factor model to estimate the effect of latent biological processes on gene expression variation.

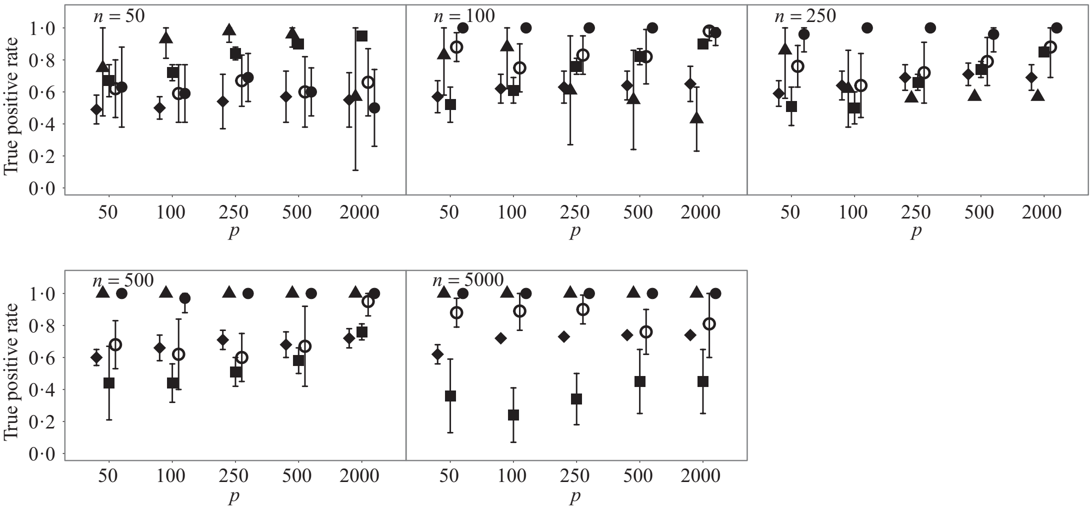

Fig. 3. True positive rate averaged across simulation replications for the methods of Hirose & Yamamoto (2015) (squares), Ročková & George (2016) varimax-free version (empty circles), Ročková & George (2016) varimax version (filled circles) and Witten et al. (2009) (diamonds), as well as our estimation algorithm (triangles). Error bars represent Monte Carlo errors.

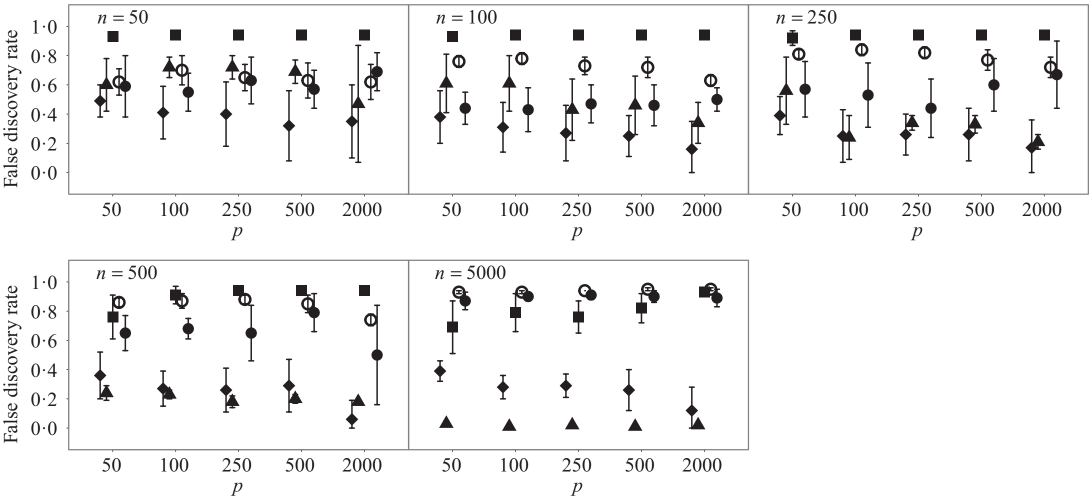

Fig. 4. False discovery rate averaged across simulation replications for the methods of Hirose & Yamamoto (2015) (squares), Ročková & George (2016) original version (empty circles), Ročková & George (2016) varimax version (filled circles) and Witten et al. (2009) (diamonds), as well as our estimation algorithm (triangles). Error bars represent Monte Carlo errors.

AGEMAP data were centred before analysis following Perry & Owen (2010). Gene expression measurements were represented by $Y \in \mathbb{R}^{n\times p}$, where $n = 618$ and $p = 8932$. Further, age$_i$ represented the age of mouse i and gender$_i$ was 1 if mouse i was female and 0 otherwise. Least-squares estimates of the intercept, age effect and gender effect in the linear model $y_{id} = \beta_{0d} + \beta_{1d}\,\text{age}_i + \beta_{2d}\,\text{gender}_i + e_{id}$ $(i = 1,\ldots,n)$, with idiosyncratic error e_{id}, were represented as $\hat\beta_{0d}$, $\hat\beta_{1d}$ and $\hat\beta_{2d}$. Using these estimates for $d = 1,\ldots,p$, the mean-centred data were defined as

$$\hat y_{id} = y_{id} - \hat\beta_{0d} + \hat\beta_{1d}\,\text{age}_i + \hat\beta_{2d}\,\text{gender}_i \quad (i = 1,\ldots,n;\ d = 1,\ldots,p).$$

Four mice were randomly held out, and all tissue samples for these mice in $\hat{Y}$ were used as test data. The remaining samples were used as training data. This set-up was replicated ten times. All four methods were applied to the training data in every replication by fixing the upper bound on the number of factors at 10. The δ-ρ grid had dimensions 20×20, and $\log_{10} \delta$ increased linearly from $\log_{10} 2$ to $\log_{10} 10$ while $\log_{10} \rho$ increased linearly from -3 to 6.

The results for all five methods were stable across all ten folds of crossvalidation. Caner and Han's method, Hirose and Yamamoto's method, both versions of Ročková and George's method, the method of Witten et al. and our method selected 10, 10, 10, 4 and 1, respectively, as the number of latent biological processes k^* across all folds. Our result matched the result of Perry & Owen (2010), who confirmed the presence of one latent variable using rotation tests. Our simulation results and the findings in Perry & Owen (2010) strongly suggest that our method accurately estimated k^* and the other methods overestimated k^*.

We also estimated the factors for the test data. With $\hat{y}_i$ denoting test datum i and UDV^{T} denoting the singular value decomposition of Λ, the factor estimate of test datum i was $n_{\mathrm{T}}^{-1/2} U^{\mathrm{T}} \hat{y}_i$, where n_{T} denotes the number of samples in the training data. Perry & Owen (2010) found that factor estimates for the tissue samples from cerebrum and cerebellum, respectively, had bimodal densities. We used the density function in R with default settings to obtain kernel density estimates of the factors. Hirose and Yamamoto's method and both versions of Ročková and George's method estimated the number of factors as 10, which made the results challenging to interpret. The method of Witten et al. recovered bimodal densities in all four factors for both tissue samples, but it was unclear which of these four factors corresponded to the factor estimated by Perry & Owen (2010). Our method estimated the number of factors to be 1 and recovered the bimodal density in both tissue samples.

Acknowledgement

This work was supported by the U.S. National Institute of Environmental Health Sciences, National Institutes of Health, and National Science Foundation. We are grateful to the referees, associate editor, and editor for their comments and suggestions.

Supplementary material

Supplementary material available at *Biometrika* online includes derivation of the expectation-maximization algorithm, proofs of Lemmas 1 and 2 and Theorems 1–3, supporting figures for the results in § 5·3, and the R code used for data analysis.

Appendix

Assumptions

Assumptions A0–A4 follow from the theoretical set-up for high-dimensional factor models in Kneip & Sarda (2011). Assumption A5 is based on results in Zou & Li (2008) for variable selection.

Assumption A0. Let $y_i = w_i + e_i$, $E(y_i) = 0$, $\mathrm{var}(y_i) = \Omega^*$, $E(w_i) = 0$, $\mathrm{var}(w_i) = \Lambda^* \Lambda^{*\mathrm{T}}$, $E(e_i) = 0$ and $\mathrm{var}(e_i) = \Sigma^*$ $(i = 1, \ldots, n)$.

Assumption A1. There exist finite positive constants D_0, D_3 and $D_1 \leqslant D_2$ such that $E(y_{id}^2) \leqslant D_0$, $E(e_{id}^4) \leqslant D_3$ and $0 < D_1 \leqslant (\Sigma^*)_{dd} \leqslant D_2$ $(i = 1, \ldots, n;\ d = 1, \ldots, p)$.

Assumption A2. There exists a constant $C_0 \in (8,\infty)$ such that $\sum_{i=1}^n w_{ij}w_{il}/n$, $\sum_{i=1}^n e_{ij}e_{il}/n$, $\sum_{i=1}^n w_{ij}e_{il}/n$ and $\sum_{i=1}^n y_{ij}y_{il}/n$ are (C_0/n)-sub-Gaussian for every $j,l \in \{1,\ldots,p\}$. A random variable X is c-sub-Gaussian if $\text{pr}\{|X - E(X)| > t\} \leqslant 2\exp\{-t^2/(2c)\}$ for any $t > 0$.

Assumption A3. Let $b_1 > \cdots > b_{k^*} > 0$ be the eigenvalues of $\Lambda^*\Lambda^{*\text{T}}$; then there exists a v_0 such that $0 < v_0 \leqslant 1$, $pv_0 > 6D_2$, $\min_{j,l\leqslant k^*, j\neq l} |b_j/p - b_l/p| \geqslant v_0$, and $\min_{j\leqslant k^*} b_j/p \geqslant v_0$.

Assumption A4. The sample size n and dimension $p \geqslant n$ are large enough that $C_0(\log p/n)^{1/2} \geqslant D_0/p$ and $v_0 \geqslant 6\{D_2/p + C_0(\log p/n)^{1/2}\}$.

Assumption A5. Let k be the upper bound on k^* and let δ, ρ, $\alpha_j(\delta)$ and $\eta_j(\rho)$ $(j = 1,\ldots,k)$ be defined as in Lemma 2. Then $k = O(\log p)$, $\alpha_j(\delta) \to \infty$, $n^{-1/2}\alpha_j(\delta) \to 0$ and $(np)^{1/2}\eta_j(\rho) \to c_j > 0$ $(j = 1,\ldots,k)$ as $n \to \infty$, $n \leqslant p \to \infty$ and $\log p/n \to 0$.

Assumption A6. The elements of the set $\mathcal{M}^*$ are fixed and do not change as n or p increases to ∞.

Model (2) is recovered upon substituting $w_i = \Lambda^* z_i$ into Assumption A0. Assumption A1 ensures that Ω^* is positive definite. Assumption A2 ensures that the empirical covariances are good approximations of the true covariances. Specifically, for any $t > 0$,

$$\sup_{1\leqslant j,l\leqslant p}\left|\frac{1}{n}\sum_{i=1}^n w_{ij}w_{il} - \text{cov}(w_{ij},w_{il})\right| \leqslant t, \quad \sup_{1\leqslant j,l\leqslant p}\left|\frac{1}{n}\sum_{i=1}^n e_{ij}e_{il} - \text{cov}(e_{ij},e_{il})\right| \leqslant t,$$

$$\sup_{1\leqslant j,l\leqslant p}\left|\frac{1}{n}\sum_{i=1}^n w_{ij}e_{il}\right| \leqslant t, \quad \sup_{1\leqslant j,l\leqslant p}\left|\frac{1}{n}\sum_{i=1}^n y_{ij}y_{il} - \text{cov}(y_{ij},y_{il})\right| \leqslant t$$

hold simultaneously with probability at least $A_t(n,p) = 1 - 8p^2\exp\{-nt^2/(2C_0)\}$. If $t_0 = C_0(\log p/n)^{1/2}$, then $A_{t_0}(n,p) \to 1$ as $n,p \to \infty$ and $\log p/n \to 0$. Assumption A3 guarantees identifiability of Λ^0 when p is large and $v_0 \gg 1/p$. Assumption A4 is required to ensure that $p^{-1/2}(\hat{\zeta}_d)^{1/2}\hat{\psi}_{dj}$ is a root-n-consistent estimator of $p^{-1/2}\lambda_{dj}$ as $n \to \infty$, $n \leqslant p \to \infty$ and $\log p/n \to 0$.

One additional assumption is required to relate $\text{EBIC}_\gamma(\mathcal{M})$ and $\pi_{\mathcal{M}} = \text{pr}(\mathcal{M} \mid Y)$.

Assumption A7. Let $p = O(n^\kappa)$ for a fixed constant $\kappa \geqslant 1$ such that $\gamma > 1 - 1/(2\kappa)$.

Assumption A7 and equation (4.6) in Theorem 3 of Kneip & Sarda (2011) imply that $\hat{\zeta}_l > 0$ for any l such that $1 \leqslant l \leqslant 2k < p$, because $(\log p)^{3/2}/n^{1/2} \to 0$ as $n \to \infty$.

References

Ahn, S. C. & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–27.

Armagan, A., Dunson, D. B. & Lee, J. (2013). Generalized double Pareto shrinkage. *Statist. Sinica* **23**, 119–43.

Bai, J. & Li, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40**, 436–65.

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

Bhattacharya, A. & Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98**, 291–306.

Caner, M. & Han, X. (2014). Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. *J. Bus. Econ. Statist.* **32**, 359–74.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Am. Statist. Assoc.* **103**, 1438–56.

Chen, J. & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–71.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Friedman, J. H., Hastie, T. J. & Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.

Hirose, K. & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statist. Comp.* **25**, 863–75.

Jolliffe, I. T., Trendafilov, N. T. & Uddin, M. (2003). A modified principal component technique based on the LASSO. *J. Comp. Graph. Statist.* **12**, 531–47.

Kneip, A. & Sarda, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *Ann. Statist.* **39**, 2410–47.

Knowles, D. & Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Ann. Statist.* **5**, 1534–52.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica* **77**, 1447–79.

Perry, P. O. & Owen, A. B. (2010). A rotation test to verify latent structure. *J. Mach. Learn. Res.* **11**, 603–24.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Ročková, V. & George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Am. Statist. Assoc.* **111**, 1608–22.

Shen, H. & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Mult. Anal.* **99**, 1015–34.

Witten, D. M., Tibshirani, R. J. & Hastie, T. J. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–34.

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., Hastie, T. J. & Tibshirani, R. J. (2006). Sparse principal component analysis. *J. Comp. Graph. Statist.* **15**, 265–86.

Zou, H. & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509–33.

[*Received on* 10 *August* 2016. *Editorial decision on* 18 *March* 2017]