PSYCHOMETRIKA—VOL. 88, NO. 4, 1381–1406
DECEMBER 2023
https://doi.org/10.1007/s11336-022-09868-4

SPARSE AND SIMPLE STRUCTURE ESTIMATION VIA PRENET PENALIZATION

KEI HIROSE
KYUSHU UNIVERSITY
RIKEN CENTER FOR ADVANCED INTELLIGENCE PROJECT
YOSHIKAZU TERADA
OSAKA UNIVERSITY
RIKEN CENTER FOR ADVANCED INTELLIGENCE PROJECT

We propose a *prenet* (*pr*oduct-based *e*lastic *net*), a novel penalization method for factor analysis models. The penalty is based on the product of a pair of elements in each row of the loading matrix. The prenet not only shrinks some of the factor loadings toward exactly zero but also enhances the simplicity of the loading matrix, which plays an important role in the interpretation of the common factors. In particular, with a large amount of prenet penalization, the estimated loading matrix possesses a perfect simple structure, which is known as a desirable structure in terms of the simplicity of the loading matrix. Furthermore, the perfect simple structure estimation via the proposed penalization turns out to be a generalization of the k-means clustering of variables. On the other hand, a mild amount of the penalization approximates a loading matrix estimated by the quartimin rotation, one of the most commonly used oblique rotation techniques. Simulation studies compare the performance of our proposed penalization with that of existing methods under a variety of settings. The usefulness of the perfect simple structure estimation via our proposed procedure is presented through various real data applications.

Key words: multivariate analysis, quartimin rotation, penalized maximum likelihood estimation, perfect simple structure, sparse estimation.

Factor analysis investigates the correlation structure of high-dimensional observed variables by the construction of a small number of latent variables called common factors. Factor analysis can be considered as a soft clustering of variables, in which each factor corresponds to a cluster and observed variables are categorized into overlapping clusters. For interpretation purposes, it is desirable for the observed variables to be well-clustered (Yamamoto and Jennrich, 2013) or the loading matrix to be simple (Thurstone, 1947). In particular, the perfect simple structure (e.g., Bernaards and Jennrich, 2003; Jennrich, 2004), wherein each row of the loading matrix has at most one nonzero element, provides a non-overlapping clustering of variables in the sense that variables that correspond to nonzero elements of the jth column of the loading matrix belong to the jth cluster.

Conventionally, the well-clustered or simple structure of the loading matrix is found by rotation techniques. A number of rotation techniques have been proposed in the literature; for example, quartimin rotation (Carroll, 1953), varimax rotation (Kaiser, 1958), promax rotation (Hendrickson and White, 1964), simplimax rotation (Kiers, 1994), geomin rotation (Yates, 1987), and component loss criterion (Jennrich, 2004, 2006). The literature review of the rotation techniques is described in Browne 2001. The main purpose of the factor rotation is to get a good solution that is

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-022-09868-4.

Correspondence should be made to Kei Hirose, Institute of Mathematics for Industry, Kyushu University, Fukuoka, Japan. Email: hirose@imi.kyushu-u.ac.jp; URL: https://keihirose.com

© 2022 The Author(s)

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

as simple as possible. See, e.g., Thurstone (1935), Carroll (1953), Neuhaus and Wrigley (1954), Kaiser (1974), Bernaards and Jennrich (2003).

The problem with the rotation technique is that it cannot produce a sufficiently sparse solution in some cases (Hirose and Yamamoto, 2015), because the loading matrix must be found among a set of unpenalized maximum likelihood estimates. To obtain sparser solutions than the factor rotation, we employ a penalization method. It is shown that the penalization is a generalization of the rotation techniques and can produce sparser solutions than the rotation methods (Hirose and Yamamoto, 2015). Typically, many researchers use the L_1-type penalization, such as the lasso (Tibshirani, 1996), the adaptive lasso (Zou, 2006), and the minimax concave penalty (MCP) (Zhang, 2010); for example, Choi et al. (2011), Ning and Georgiou (2011), Srivastava et al. (2017), Hirose and Yamamoto (2015), Trendafilov et al. (2017), Hui et al. (2018). The L_1 penalization shrinks some of the parameters toward exactly zero; in other words, parameters that need not to be modeled are automatically disregarded. Furthermore, the degrees of sparsity are freely adjusted by changing the value of a regularization parameter. The L_1 penalization provides a good estimation accuracy, such as L_1 consistency and model selection consistency in high dimension (e.g., Zhao and Yu 2007; Wainwright, 2009; Bhlmann and van de Geer, 2011).

As described above, it is important to obtain a "good" loading matrix in the sense of simplicity and thus interpretability in the exploratory factor analysis. Although the L_1 penalization achieves the sparse estimation, the estimated loading matrix is not guaranteed to possess an interpretable structure. For example, a great amount of penalization leads to a zero matrix, which does not make sense from an interpretation viewpoint. Thus, the L_1 penalization cannot produce an interpretable loading matrix with a sufficient large regularization parameter. Even if a small value of regularization parameter is selected, the L_1 penalization cannot often approximate a true loading matrix when it is not sufficiently sparse; with the lasso, some of the factor loadings whose true values are close —but not very close—to zero are estimated as zero values, and this misspecification can often cause a significant negative effect on the estimation of other factor loadings (Hirose and Yamamoto, 2014). Therefore, it is important to estimate a loading matrix that is not only sparse but also interpretable. To achieve this, we need a different type of penalty.

In this study, we propose a *prenet* (*pr*oduct-based *e*lastic *net*) penalty, which is based on the product of a pair of parameters in each row of the loading matrix. A remarkable feature of the prenet is that a large amount of penalization leads to the perfect simple structure. The existing L_1-type penalization methods do not have that significant property. Furthermore, the perfect simple structure estimation via the prenet penalty is shown to be a generalization of the k-means clustering of variables. On the other hand, with a mild amount of prenet penalization, the estimated loading matrix is approximated by that obtained using the quartimin rotation, a widely used oblique rotation method. The quartimin criterion can often estimate a non-sparse loading matrix appropriately, so that the problem of the lasso-type penalization mentioned above is addressed. We employ the generalized expectation and maximization (EM) algorithm and the coordinate descent algorithm (e.g., Friedman et al., 2010) to obtain the estimator. The proposed algorithm monotonically decreases the objective function at each iteration.

In our proposed procedure, the regularization parameter controls the degrees of simplicity; the larger the regularization parameter is, the simpler the loading matrix is. The advantage of our proposed procedure is that we can change the degrees of simplicity according to the purpose of the analysis. This study focus on two different purposes of the analysis: (i) to find a loading matrix that fits the data and also is simple as much as possible and (ii) to conduct cluster analysis by estimating a perfect simple structure. The regularization parameter selection procedure differs depending on these two purposes. To achieve the purpose (i), we select the regularization parameter by the Akaike information criterion (AIC; Akaike, 1973; Zou et al., 2007) or the Bayesian information criterion (BIC; Schwarz, 1978; Zou et al., 2007). The purpose (ii) is attained by setting the regularization parameter to be infinity.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

We conduct the Monte Carlo simulations to compare the performance of our proposed method with that of L_1-type penalization and conventional rotation techniques. The Monte Carlo simulations investigate the performance in terms of both (i) and (ii); investigations of (i) and (ii) are detailed in Sects. 5.2 and 5.3, respectively. Our proposed method is applied to data from big five personality traits to study the performance for various sample sizes and impact of the regularization parameter on the accuracy. The analysis of big five personality traits aims at purpose (i). We also present the analyses of fMRI and electricity demand data, intended to purpose both (i) and (ii), in Section S2 of the supplemental material.

The rest of this article is organized as follows. Section 1 describes the estimation of the factor analysis model via penalization. In Sect. 2, we introduce the prenet penalty. Section 3 describes several properties of the prenet penalty, including its relationship with the quartimin criterion. Section 4 presents an estimation algorithm to obtain the prenet solutions. In Sect. 5, we conduct a Monte Carlo simulation to investigate the performance of the prenet penalization. Section 6 illustrates the usefulness of our proposed procedure through real data analysis. Extension and future works are discussed in Sect. 7. Some technical proofs and detail of our algorithm are shown in "Appendix." Supplemental materials include further numerical and theoretical investigation, including numerical analyses of resting-state fMRI and electricity demand data.

1. Estimation of Factor Analysis Model via Penalization

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a p-dimensional observed random vector with mean vector $\mathbf{0}$ and variance–covariance matrix $\boldsymbol{\Sigma}$. The factor analysis model is

$$\boldsymbol{X} = \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a $p \times m$ loading matrix, $\boldsymbol{F} = (F_1, \ldots, F_m)^T$ is a random vector of common factors, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$ is a random vector of unique factors. It is assumed that $E(\boldsymbol{F}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{F}\boldsymbol{F}^T) = \boldsymbol{\Phi}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{\Psi}$, and $E(\boldsymbol{F}\boldsymbol{\varepsilon}^T) = \boldsymbol{O}$, where $\boldsymbol{\Phi}$ is an $m \times m$ factor correlation matrix, and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix (i.e., strict factor model). The diagonal elements of $\boldsymbol{\Psi}$ are referred to as unique variances. Under these assumptions, the variance–covariance matrix of observed random vector $\boldsymbol{X}$ is $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$.

In many cases, the orthogonal factor model (i.e., $\boldsymbol{\Phi} = \boldsymbol{I}_m$) is used but is often oversimplified. Here, $\boldsymbol{I}_m$ is an identity matrix of order m. This paper covers the oblique factor model, which allows a more realistic estimation of latent factors than the orthogonal factor model in many cases (e.g., Fabrigar et al., 1999; Sass and Schmitt, 2010; Schmitt and Sass, 2011).

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n observations and $\boldsymbol{S} = (s_{ij})$ be the corresponding sample covariance matrix. Let $\boldsymbol{\theta} = (\mathrm{vec}(\boldsymbol{\Lambda})^T, \mathrm{diag}(\boldsymbol{\Psi})^T, \mathrm{vech}(\boldsymbol{\Phi})^T)^T$ be a parameter vector, where $\mathrm{vech}(\cdot)$ is a vector that consists of a lower triangular matrix without diagonal elements. We estimate the model parameter by minimizing the penalized loss function $\ell_\rho(\boldsymbol{\theta})$ expressed as

$$\ell_\rho(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \rho P(\boldsymbol{\Lambda}), \tag{2}$$

where $\ell(\boldsymbol{\theta})$ is a loss function, $P(\boldsymbol{\Lambda})$ is a penalty function, and $\rho > 0$ is a regularization parameter. As a loss function, we adopt the maximum likelihood discrepancy function

$$\ell_{\mathrm{DF}}(\boldsymbol{\theta}) = \frac{1}{2}\left\{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) - \log|\boldsymbol{\Sigma}^{-1}\boldsymbol{S}| - p\right\}, \tag{3}$$

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

where DF is an abbreviation for discrepancy function. Assume that the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are drawn from the p-dimensional normal population $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. The minimizer of $\ell_{\mathrm{DF}}(\boldsymbol{\theta})$ is the maximum likelihood estimate. It is shown that $\ell_{\mathrm{DF}}(\boldsymbol{\theta}) \geq 0$ for any $\boldsymbol{\theta}$, and $\ell_{\mathrm{DF}}(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} = \boldsymbol{S}$ when $\boldsymbol{\Sigma}$ and $\boldsymbol{S}$ are positive definite matrices. It is worth noting that our proposed penalty, described in Sect. 2, can be directly applied to many other loss functions, including a quadratic loss used for generalized least squares (Jöreskog and Goldberger, 1971).

When $\boldsymbol{\Phi} = \boldsymbol{I}_m$, the model has a rotational indeterminacy; both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}\boldsymbol{T}$ generate the same covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{T}$ is an arbitrary orthogonal matrix. Thus, when $\rho = 0$, the solution that minimizes (2) is not uniquely determined. However, when $\rho > 0$, the solution may be uniquely determined except for the sign and permutation of columns of the loading matrix when an appropriate penalty $P(\boldsymbol{\Lambda})$ is chosen.

2. Prenet Penalty

2.1. Definition

We propose a prenet penalty

$$P(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \sum_{j=1}^{m-1} \sum_{k>j} \left\{ \gamma |\lambda_{ij}\lambda_{ik}| + \frac{1}{2}(1-\gamma)(\lambda_{ij}\lambda_{ik})^2 \right\}, \tag{4}$$

where $\gamma \in (0, 1]$ is a tuning parameter. The most significant feature of the prenet penalty is that it is based on the product of a pair of parameters.

When $\gamma \to 0$, the prenet penalty is equivalent to the quartimin criterion (Carroll, 1953), a widely used oblique rotation criterion in factor rotation. As is the case with the quartimin rotation, the prenet penalty in (4) eliminates the rotational indeterminacy except for the sign and permutation of columns of the loading matrix and contributes significantly to the estimation of the simplicity of the loading matrix. The prenet penalty includes products of absolute values of factor loadings, producing factor loadings that are exactly zero.

2.2. Comparison with the Elastic Net Penalty

The prenet penalty is similar to the elastic net penalty (Zou and Hastie, 2005)

$$P(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \sum_{j=1}^{m} \left\{ \gamma |\lambda_{ij}| + \frac{1}{2}(1-\gamma)\lambda_{ij}^2 \right\}, \tag{5}$$

which is a hybrid of the lasso and the ridge penalties. Although the elastic net penalty is similar to the prenet penalty, there is a fundamental difference between these two penalties; the elastic net is constructed by the sum of the elements of parameter vector, whereas the prenet is based on the product of a pair of parameters.

Figure 1 shows the penalty functions of the prenet ($P(x, y) = \gamma|xy| + (1-\gamma)(xy)^2/2$) and the elastic net ($P(x, y) = \gamma(|x| + |y|) + (1-\gamma)(x^2 + y^2)/2$) when $\gamma = 0.1, 0.5, 0.9$. Clearly, the prenet penalty is a nonconvex function. A significant difference between the prenet and the elastic net is that although the prenet penalty becomes zero when either x or y attains zero, the elastic net penalty becomes zero only when both $x = 0$ and $y = 0$. Therefore, for a two-factor

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

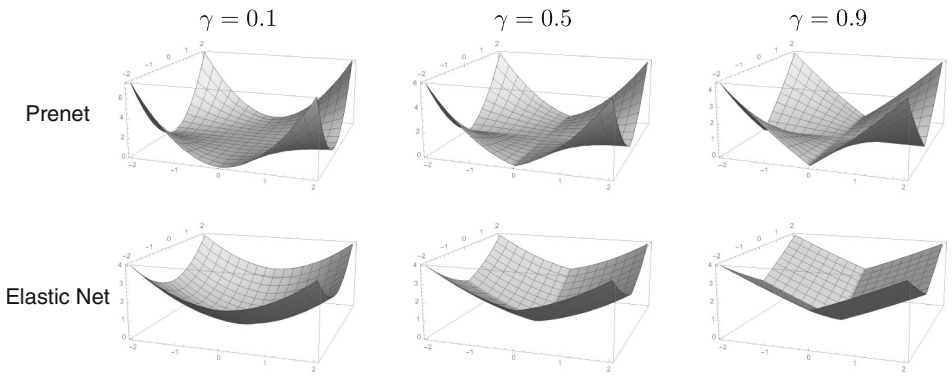


FIGURE 1.
Penalty functions of the prenet and the elastic net with various γ.

model, the estimate of either λ_{i1} or λ_{i2} can be zero with the prenet penalization, leading to a perfect simple structure. On the other hand, the elastic net tends to produce estimates in which both $|\lambda_{i1}|$ and $|\lambda_{i2}|$ are small.

The penalty functions in Fig. 1 also show that the prenet penalty becomes smooth as γ decreases. Thus, the value of γ controls the degrees of sparsity; the larger the value of γ, the sparser the estimate of the loading matrix. With an appropriate value of γ, the prenet penalty enhances both simplicity and sparsity of the loading matrix. Further investigation into the estimator against the value of γ is presented in Sect. 5.

3. Properties of the Prenet Penalty

3.1. Perfect Simple Structure

The model (1) does not impose an orthogonal constraint on the loading matrix $\mathbf{\Lambda}$. For unconstrained $\mathbf{\Lambda}$, most existing penalties, such as the lasso, shrink all coefficients toward zero when the tuning parameter ρ is sufficiently large; we usually obtain $\hat{\mathbf{\Lambda}} = \boldsymbol{O}$ when $\rho \to \infty$. However, the following proposition shows that the prenet penalty does not necessarily shrink all of the elements toward zero even when ρ is sufficiently large.

Proposition 1. *Assume that we use the prenet penalty with $\gamma \in (0, 1]$. As $\rho \to \infty$, the estimated loading matrix possesses the perfect simple structure, that is, each row has at most one nonzero element.*

Proof. As $\rho \to \infty$, $P(\hat{\mathbf{\Lambda}})$ must satisfy $P(\hat{\mathbf{\Lambda}}) \to 0$. Otherwise, the second term of (2) diverges. When $P(\hat{\mathbf{\Lambda}}) = 0$, $\hat{\lambda}_{ij}\hat{\lambda}_{ik} = 0$ for any $j \neq k$. Therefore, the ith row of $\mathbf{\Lambda}$ has at most one nonzero element. □

The perfect simple structure is known as a desirable property in the literature in factor analysis because it is easy to interpret the estimated loading matrix (e.g., Bernaards and Jennrich, 2003). When ρ is small, the estimated loading matrix can be away from the perfect simple structure but the goodness of fit to the model is improved.

Remark 1. For $\rho \to \infty$, the consistency of the loading matrix is shown when the true loading matrix possesses the perfect simple structure. For simplicity, we consider the orthogonal case. Assume that $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$, where the true $\mathbf{\Lambda}$ possesses the perfect simple structure. As $n \to \infty$,

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

the sample covariance matrix converges to the true covariance matrix almost surely; thus, the loss function (3) is minimized when $\boldsymbol{\Sigma} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \hat{\boldsymbol{\Psi}}$. When $\rho \to \infty$, $\hat{\boldsymbol{\Lambda}}$ must be the perfect simple structure. Therefore, we should consider the following problem:

$$\text{Find } (\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Psi}}) \text{ that satisfies } \begin{cases} \boldsymbol{\Sigma} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \hat{\boldsymbol{\Psi}}, \\ \hat{\boldsymbol{\Lambda}} \text{ is perfect simple structure.} \end{cases}$$

The solution to the above problem is $\hat{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}$ except for the sign and permutation of columns of the loading matrix if an identifiability condition for an orthogonal model (e.g., Theorem 5.1 in Anderson and Rubin, 1956) is satisfied.

3.2. Relationship with k-Means Clustering of Variables

The perfect simple structure corresponds to variables clustering, that is, variables that correspond to nonzero elements of the jth column of the loading matrix belong to the jth cluster. In this subsection, we investigate the relationship between the prenet with $\rho \to \infty$ and the k-means clustering of variables, one of the most popular cluster analyses.

Let $\boldsymbol{X}_0$ be an $n \times p$ data matrix. $\boldsymbol{X}_0$ can be expressed as $\boldsymbol{X}_0 = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*)$, where $\boldsymbol{x}_i^*$ is the ith column vector of $\boldsymbol{X}_0$. We consider the problem of the variables clustering of $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*$ by the k-means. Let C_j $(j = 1, \ldots, m)$ be a subset of indices of variables that belong to the jth cluster. The objective function of the k-means is

$$\sum_{j=1}^{m} \sum_{i \in C_j} \|\boldsymbol{x}_i^* - \boldsymbol{\mu}_j\|^2 = (n-1) \left(\sum_{i=1}^{p} s_{ii} - \sum_{j=1}^{m} \frac{1}{p_j} \sum_{i \in C_j} \sum_{i' \in C_j} s_{ii'} \right), \tag{6}$$

where $p_j = \#\{C_j\}$, $\boldsymbol{\mu}_j = \frac{1}{p_j} \sum_{i \in C_j} \boldsymbol{x}_i^*$, and recall that $s_{ii'} = \boldsymbol{x}_i^{*T} \boldsymbol{x}_{i'}^* / (n-1)$. Let $\boldsymbol{\Lambda} = (\lambda_{ij})$ be a $p \times m$ indicator variables matrix

$$\lambda_{ij} = \begin{cases} 1/\sqrt{p_j} & i \in C_j, \\ 0 & i \notin C_j. \end{cases} \tag{7}$$

Using the fact that $\boldsymbol{\Lambda}^T \boldsymbol{\Lambda} = \boldsymbol{I}_m$, the k-means clustering of variables using (6) is equivalent to (Ding et al., 2005)

$$\min_{\boldsymbol{\Lambda}} \|\boldsymbol{S} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\|_F^2, \quad \text{subject to (7)}, \tag{8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. We consider slightly modifying the condition on $\boldsymbol{\Lambda}$ in (7) to

$$\lambda_{ij}\lambda_{ik} = 0 \ (j \neq k) \text{ and } \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} = \boldsymbol{I}_m. \tag{9}$$

The modified k-means problem is then given as

$$\min_{\boldsymbol{\Lambda}} \|\boldsymbol{S} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\|_F^2 \text{ subject to (9)}. \tag{10}$$

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

The condition (9) is milder than (7); if $\mathbf{\Lambda}$ satisfies (7), we obtain (9). The reverse does not always hold; with (9), the nonzero elements for each column do not have to be equal. Therefore, the modified k-means in (10) may capture a more complex structure than the original k-means.

Proposition 2. *Assume that* $\mathbf{\Psi} = \alpha \boldsymbol{I}_p$, $\mathbf{\Phi} = \boldsymbol{I}_m$, *and* α *is given. Suppose that* $\mathbf{\Lambda}$ *satisfies* $\mathbf{\Lambda}^T \mathbf{\Lambda} = \boldsymbol{I}_m$. *The prenet solution with* $\rho \to \infty$ *is then obtained by (10).*

Proof. The proof appears in Appendix A.1. □

The proposition 2 shows that the prenet solution with $\rho \to \infty$ is a generalization of the problem (10). As mentioned above, the problem (10) is a generalization of the k-means problem in (8). Therefore, the perfect simple structure estimation via the prenet is a generalization of the k-means clustering of variables. We remark that the condition $\mathbf{\Psi} = \alpha \boldsymbol{I}_p$ in Proposition 2 implies the probabilistic principal component analysis (probabilistic PCA; Tipping and Bishop, 1999); the penalized probabilistic PCA via the prenet is also a generalization of the k-means clustering of variables.

3.3. Relationship with Quartimin Rotation

As described in Sect. 2, the prenet penalty is a generalization of the quartimin criterion (Carroll, 1953); setting $\gamma \to 0$ to the prenet penalty in (4) leads to the quartimin criterion

$$P_{\text{qmin}}(\mathbf{\Lambda}) = \sum_{i=1}^{p} \sum_{j=1}^{m-1} \sum_{k>j} (\lambda_{ij} \lambda_{ik})^2.$$

The quartimin criterion is typically used in the factor rotation. The solution of quartimin rotation method, say $\hat{\boldsymbol{\theta}}_q$, is obtained by two-step procedure. First, we calculate an unpenalized estimate, denoted by $\hat{\boldsymbol{\theta}}$. The estimate $\hat{\boldsymbol{\theta}}$, that satisfies $\ell(\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, is not unique due to the rotational indeterminacy. The second step is the minimization of the quartimin criterion with a restricted parameter space $\{\boldsymbol{\theta} | \ell(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})\}$. Hirose and Yamamoto (2015) showed that the solution of the quartimin rotation, $\hat{\boldsymbol{\theta}}_q$, can be obtained by

$$\min_{\boldsymbol{\theta}} P_{\text{qmin}}(\mathbf{\Lambda}), \text{ subject to } \ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) \tag{11}$$

under the condition that the unpenalized estimate of loading matrix $\hat{\mathbf{\Lambda}}$ is unique if the indeterminacy of the rotation in $\hat{\mathbf{\Lambda}}$ is excluded. It is not easy to check this condition, but several necessary conditions of the identifiability for the orthogonal model are provided (e.g., Theorem 5.1 in Anderson and Rubin, 1956.)

Now, we show a basic asymptotic result of the prenet solution, from which we can see that the prenet solution is a generalization of the quartimin rotation. Let (Θ, d) be a compact parameter space with distance d and $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that for any $(\text{vec}(\mathbf{\Lambda})^T, \text{diag}(\mathbf{\Psi})^T, \text{vech}(\mathbf{\Phi})^T)^T \in \Theta$ and any $\boldsymbol{T} \in \mathcal{O}(m)$, we have $(\text{vec}(\mathbf{\Lambda}\boldsymbol{T})^T, \text{diag}(\mathbf{\Psi})^T, \text{vech}(\mathbf{\Phi})^T)^T \in \Theta$, where $\mathcal{O}(m)$ is a set of $m \times m$ orthonormal matrices. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote independent $\mathbb{R}^p$-valued random variables with the common population distribution $\mathbb{P}$. Now, it is required that we can rewrite the empirical loss function and the true loss function as $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} q(\boldsymbol{X}_i; \boldsymbol{\theta})/n$ and $\ell_*(\boldsymbol{\theta}) = \int q(\boldsymbol{x}; \boldsymbol{\theta}) \, \mathbb{P}(d\boldsymbol{x})$, respectively. Let $\hat{\boldsymbol{\theta}}_\rho$ denote an arbitrary measurable prenet estimator which satisfies $\ell(\hat{\boldsymbol{\theta}}_\rho) + \rho P(\hat{\mathbf{\Lambda}}_\rho) = \min_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}) + \rho P(\mathbf{\Lambda})$.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

Condition 3.1. $q(\boldsymbol{x};\boldsymbol{\theta})$ fulfills the following conditions:

- For each $\boldsymbol{x} \in \mathbb{R}^p$, function $q(\boldsymbol{x};\boldsymbol{\theta})$ on Θ is continuous.
- There exists a $\mathbb{P}$-integrable function $g(\boldsymbol{x})$ such that for all $\boldsymbol{x} \in \mathbb{R}^p$ and for all $\boldsymbol{\theta} \in \Theta$ $|q(\boldsymbol{x};\boldsymbol{\theta})| \leq g(\boldsymbol{x})$.

Since $\ell(\boldsymbol{\theta})$ is the discrepancy function in Eq. (3), $q(\boldsymbol{x};\boldsymbol{\theta})$ becomes a logarithm of density function of normal distribution; in this case, Condition 3.1 is satisfied. The following proposition shows that the prenet estimator converges almost surely to a true parameter which minimizes the quartimin criterion when $\rho \to 0$ as $n \to \infty$.

Proposition 3. *Assume that Condition 3.1 is satisfied. We denote by Θ_q^* a set of true solutions of the following quartimin problem.*

$$\min_{\boldsymbol{\theta}\in\Theta} P_{\text{qmin}}(\boldsymbol{\Lambda}) \quad \textit{subject to} \quad \ell_*(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in\Theta} \ell_*(\boldsymbol{\theta}).$$

Let ρ_n $(n = 1, 2, \ldots)$ be a sequence that satisfies $\rho_n > 0$ and $\lim_{n\to\infty} \rho_n = 0$. Let the prenet solution with $\gamma \to 0$ and $\rho = \rho_n$ be $\hat{\boldsymbol{\theta}}_{\rho_n}$. Then we obtain

$$\lim_{n\to\infty} d(\hat{\boldsymbol{\theta}}_{\rho_n}, \Theta_q^*) = 0 \quad a.s.,$$

where $d(a, B) = \inf_{b\in B} d(a, b)$.

Proof. The proof is given in Appendix A.2. □

Remark 3.1. Proposition 3 uses a set of true solutions Θ_q^* instead of one true solution $\boldsymbol{\theta}_q^*$. This is because even if the quartimin solution does not have a rotational indeterminacy, it still has an indeterminacy with respect to sign and permutation of columns of the loading matrix.

Remark 3.2. The Geomin criterion (Yates, 1987) often produces a loading matrix similar to that obtained by the Quartimin criterion (Asparouhov and Muthén 2009). For the Geomin criterion, we add a small number to the loadings to address the identifiability problem (Hattori et al., 2017). Meanwhile, the prenet does not suffer from such a problem. The detailed discussion is described in Section S3 in the supplemental material.

4. Algorithm

It is well known that the solutions estimated by the lasso-type penalization methods are not usually expressed in a closed form because the penalty term includes an indifferentiable function. As the objective function of the prenet is nonconvex, it is not easy to construct an efficient algorithm to obtain a global minimum. Here, we use the generalized EM algorithm, in which the latent factors are considered to be missing values. The complete-data log-likelihood function is increased with the use of the coordinate descent algorithm (Friedman et al., 2010), which is commonly used in the lasso-type penalization. Although our proposed algorithm is not guaranteed to attain the global minimum, our algorithm decreases the objective function at each step. The update equation of the algorithm and its complexity are presented in Appendix B.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

4.1. *Efficient Algorithm for Sufficiently Large ρ*

The prenet tends to be multimodal for large ρ as is the case with the k-means algorithm. Therefore, we prepare many initial values, estimate the solutions for each initial value, and select a solution that minimizes the penalized loss function. In this case, it seems that we require heavy computational loads. However, we can construct an efficient algorithm for a sufficiently large ρ.

For sufficiently large ρ, the ith column of loading matrix $\mathbf{\Lambda}$ has at most one nonzero element, denoted by λ_{ij}. With the EM algorithm, we can easily find the location of the nonzero parameter when the current value of the parameter is given. Assume that the (i, j)th element of the loading matrix is nonzero and the (i, k)th elements $(k \neq j)$ are zero. Because the penalty function attains zero for sufficiently large ρ, it is sufficient to minimize the following function.

$$f(\lambda_{ij}) = \boldsymbol{\lambda}_i^T \boldsymbol{A}\boldsymbol{\lambda}_i - 2\boldsymbol{\lambda}_i^T \boldsymbol{b}_i = a_{jj}\lambda_{ij}^2 - 2\lambda_{ij} b_{ij}. \tag{12}$$

The minimizer is easily obtained by

$$\hat{\lambda}_{ij} = b_{ij}/a_{jj}. \tag{13}$$

Substituting (13) into (12) gives us $f(\hat{\lambda}_{ij}) = -\frac{b_{ij}^2}{a_{jj}}$. Therefore, the index j that minimizes the function $f(\lambda_{ij})$ is

$$j = \text{argmax}_k \frac{b_{ik}^2}{a_{kk}},$$

and λ_i is updated as $\hat{\lambda}_{ij} = b_{ij}/a_{jj}$ and $\hat{\lambda}_{ik} = 0$ for any $k \neq j$.

4.2. *Selection of the Maximum Value of ρ*

The value of $\rho_{\max}$, which is the minimum value of ρ that produces the perfect simple structure, is easily obtained using $\hat{\mathbf{\Lambda}}$ given by (13). Assume that $\hat{\lambda}_{ij} \neq 0$ and $\hat{\lambda}_{ik} = 0$ $(k \neq j)$. Using the update equation of λ_{ik} in (A.10) and the soft thresholding function in (A.12) of Appendix A, we show that the regularization parameter ρ must satisfy the following inequality to ensure that λ_{ik} is estimated to be zero.

$$\left| \frac{b_{ik} - a_{kj}\hat{\lambda}_{ij}}{a_{kk} + \rho\psi_i(1-\gamma)\hat{\lambda}_{ij}^2} \right| \leq \frac{\psi_i}{a_{kk} + \rho\psi_i(1-\gamma)\hat{\lambda}_{ij}^2} \rho\gamma|\hat{\lambda}_{ij}|.$$

Thus, the value of $\rho_{\max}$ is

$$\rho_{\max} = \max_i \max_{k \in C_i} \frac{|b_{ik} - a_{kj}\hat{\lambda}_{ij}|}{\gamma\psi_i|\hat{\lambda}_{ij}|},$$

where $C_i = \{k | k \neq j, \hat{\lambda}_{ij} \neq 0\}$.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

4.3. Estimation of the Entire Path of Solutions

The entire path of solutions can be produced with the grid of increasing values $\{\rho_1, \ldots, \rho_K\}$. Here, ρ_K is (5.2), and $\rho_1 = \rho_K \Delta \sqrt{\gamma}$, where Δ is a small value such as 0.001. The term $\sqrt{\gamma}$ allows us to estimate a variety of models even if γ is small.

The entire solution path can be made using a decreasing sequence $\{\rho_K \ldots, \rho_1\}$, starting with ρ_K. The proposed algorithm at ρ_K does not always converge to the global minimum, so that we prepare many initial values, estimate solutions for each initial value with the use of the efficient algorithm described in Sect. 4.1, and select a solution that minimizes the penalized log-likelihood function. We can use the warm start defined as follows: the solution at ρ_{k-1} is computed using the solution at ρ_k. The warm start leads to improved and smoother objective value surfaces (Mazumder et al., 2011).

One may use the warm start with increasing ρ; that is, the solution with $\rho = 0$ is obtained by the rotation technique with MLE, and then we gradually increase ρ, using the solution from the previous step. However, the decreasing sequence of ρ has a significant advantage over the increasing sequence; the decreasing sequence allows the application to the $n < p$ case. With an increasing order, the solution with $\rho = 0$ (MLE) is not available, and then the entire solution cannot be produced. Therefore, we adopt the warm start with decreasing sequence of ρ instead of an increasing sequence.

Another method to estimate the entire solution is to use the cold start with multiple random starts. Although the cold start does not always produce a smooth estimate as a function of ρ, it can sometimes find a better solution than the warm start when the loss function has multiple local minima. However, the cold start often requires heavier computational loads than the warm start.

When ρ is extremely small, the loss function becomes nearly flat due to rotational indeterminacy. However, in our experience, our proposed algorithm generally produces a smooth and stable estimate when the warm start is adopted. Even when the cold start is used, the estimate can often be stable for large sample sizes when ρ is not extremely but sufficiently small, such as $\rho = 10^{-4}$. However, when $n < p$, the maximum likelihood estimate cannot be obtained; therefore, the cold start often produces an unstable estimate with small ρ.

4.4. Selection of the Regularization Parameter ρ

The estimate of the loading matrix depends on the regularization parameter ρ. As described in the Introduction, this study focus on two different purposes of the analysis: (i) exploratory factor analysis and (ii) clustering of variables. When the purpose of the analysis is (ii), we simply set $\rho \to \infty$ to achieve the perfect simple structure estimation. When the purpose of the analysis is (i), ρ is selected by the AIC or the BIC (Zou et al., 2007);

$$\begin{aligned} \mathrm{AIC} &= -2n\ell(\hat{\boldsymbol{\theta}}) + 2p_0, \\ \mathrm{BIC} &= -2n\ell(\hat{\boldsymbol{\theta}}) + p_0 \log n, \end{aligned}$$

where p_0 is the number of nonzero parameters.

Our algorithm sometimes produces a loading matrix some of whose columns are zero vectors. In this case, the number of factors may be smaller than expected. The selection of the number of factors via the regularization is achieved by taking advantage of the zero column vectors estimation (Caner and Han, 2014; Hirose and Yamamoto, 2015).

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

5. Monte Carlo Simulations

The performance of our proposed method is investigated through Monte Carlo simulations. The prenet penalization has two different purposes of analysis: clustering of variables and exploratory factor analysis. In this section, we investigate the performance in terms of both purposes. The comparison of various exploratory factor analysis methods is described in Sect. 5.2, and the investigation of clustering of variables is presented in Sect. 5.3.

5.1. Simulation Models

In this simulation study, we use three simulation models as below.

Model (A):

$$\mathbf{\Lambda} = \begin{pmatrix} 0.8 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix} \otimes \mathbf{1}_{25},$$

where $\mathbf{1}_{25}$ is a 25-dimensional vector with each element being 1.

Model (B):

The size of the loading matrix of Model (B) is the same as that of Model (A), and the nonzero factor loadings share the same values. However, all zero elements in Model (A) are replaced by small random numbers from $U(-0.3, 0.3)$.

Model (C):

$$\mathbf{\Lambda} = (\mathbf{\Lambda}_1^T,\ \mathbf{\Lambda}_2^T)^T,$$

$$\mathbf{\Lambda}_1 = \begin{pmatrix} 0.79 & 0.00 & 0.00 & 0.49 & 0.50 & 0.00 & 0.68 & 0.29 & 0.66 & 0.33 & 0.00 & 0.00 & 0.62 \\ 0.00 & 0.77 & 0.00 & 0.53 & 0.00 & 0.50 & 0.32 & 0.66 & 0.00 & 0.00 & 0.62 & 0.34 & -0.64 \\ 0.00 & 0.00 & 0.76 & 0.00 & 0.52 & 0.48 & 0.00 & 0.00 & 0.34 & 0.66 & 0.33 & 0.62 & 0.00 \end{pmatrix}^T,$$

$$\mathbf{\Lambda}_2 = \begin{pmatrix} -0.62 & 0.62 & -0.62 & 0.00 & 0.00 & 0.43 & 0.47 & 0.00 & 0.42 & 0.43 & 0.00 & 0.36 & 0.26 \\ 0.64 & 0.00 & 0.00 & 0.67 & -0.67 & 0.58 & 0.00 & 0.50 & 0.57 & 0.00 & 0.50 & 0.38 & 0.43 \\ 0.00 & -0.61 & 0.61 & -0.63 & 0.63 & 0.00 & 0.57 & 0.48 & 0.00 & 0.57 & 0.46 & 0.38 & 0.38 \end{pmatrix}^T.$$

The simulation is conducted for both orthogonal and oblique models on (A) and an orthogonal model on (B) – (C). For Model (A), we write the orthogonal and oblique models as "Model (A-ORT)" and "Model (A-OBL)," respectively. Here, "ORT" and "OBL" are abbreviations for "orthogonal" and "oblique," respectively. The factor correlations for the oblique model are set to be 0.4. The unique variances are calculated by $\mathbf{\Psi} = \mathrm{diag}(\boldsymbol{I}_p - \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^T)$.

In Model (A), the loading matrix possesses the perfect simple structure. In such cases, the prenet is expected to perform well because it is able to estimate the perfect simple structure for large ρ (Proposition 1). Note that $p = 100$ on Model (A); therefore, maximum likelihood estimate cannot be available when $n < 100$.

The loading matrix of Model (B) is close to but not exactly a perfect simple structure. In this case, the prenet is expected to perform well when both ρ and γ are close to zero, thanks to Proposition 3. Meanwhile, the lasso would not perform well; a small illustrative example with intuitive description is presented in Section S1 in supplemental material.

In Model (C), the loading matrix is sparse but more complex than the perfect simple structure. The loading matrix is a rotated centroid solution of the Thurstone's box problem, reported in Thurstone (1947). We use `data(ThurstoneBox26)` in the `fungible` package in `R` to obtain

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

the loading matrix. With the original loading matrix, some of the unique variances can be larger than 1 with $\boldsymbol{\Psi} = \text{diag}(\boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)$; therefore, the elements of the original loading matrix are multiplied by 0.83. Furthermore, to enhance the sparsity, factor loadings whose absolute values are less than 0.1 are replaced by 0.

5.2. Accuracy Investigation

The model parameter is estimated by the prenet penalty with $\gamma = 1.0, 0.1, 0.01$, the lasso, the MCP (Zhang, 2010)

$$\rho P(\boldsymbol{\Lambda}; \rho; \gamma) = \sum_{i=1}^{p}\sum_{j=1}^{m} \rho \int_0^{|\lambda_{ij}|} \left(1 - \frac{x}{\rho\gamma}\right)_+ dx$$

with $\gamma = 3$, and the elastic net with $\gamma = 0.1$. The regularization parameter ρ is selected by the AIC and the BIC. We also compute a limit, $\lim_{\rho \to +0} \hat{\Lambda}_\rho$, where $\hat{\Lambda}_\rho$ is the estimate of the loading matrix obtained with a regularization parameter ρ. We note that $\lim_{\rho \to +0} \hat{\Lambda}_\rho$ corresponds to the factor rotation with MLE (Hirose and Yamamoto, 2015). In particular, the estimate with $\rho \to +0$ and $\gamma \to +0$ is equivalent to that obtained by the quartimin rotation with MLE, thanks to Proposition 3. We also conduct rotation techniques with MLE: varimax rotation (Kaiser, 1958) for the orthogonal model and promax rotation (Hendrickson and White, 1964) for the oblique model. When MLE cannot be found due to $n < p$, we conduct the lasso and obtain the approximation of the MLE with $\rho \to +0$.

The warm start is used for Models (A) and (B). The dimension of these models is $p = 100$, and then the warm start is stabler than the cold start for small ρ in our experience. Meanwhile, we adopt the cold start on Model (C) because Thurstone's box problem tends to possess multiple local minima.

In our implementation, the estimate of the elastic net sometimes diverges for the oblique model. Scharf and Nestler (2019a) reported the same phenomenon. To address this issue, we add a penalty $\zeta \log|\boldsymbol{\Phi}|$ to Eq. (2) with $\zeta = 0.01$. This penalty is based on Lee (1981), a conjugate prior for Wishart distribution from a Bayesian viewpoint. We remark that the prenet does not tend to suffer from this divergence issue even if $\zeta = 0$ in our experience. This is probably because the prenet does not shrink all loadings to zero, thanks to Proposition 1. For example, assume that $\hat{\sigma}_{ij} = \hat{\lambda}_{im}\hat{\lambda}_{jk}\hat{\phi}_{km}$ $(k \neq m)$. When the elastic net penalization is adopted, both $\hat{\lambda}_{im}$ and $\hat{\lambda}_{jk}$ are close to zero with a large ρ. When s_{ij} is large, $\hat{\phi}_{km}$ must be large to get $\sigma_{ij} \approx s_{ij}$. Accordingly, the value of $\hat{\phi}_{km}$ can be significantly large; it can be greater than 1. Meanwhile, the prenet may not suffer from this problem because either $\hat{\lambda}_{im}$ and $\hat{\lambda}_{jk}$ can become large.

For each model, $T = 1000$ data sets are generated with $N(\boldsymbol{0}, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})$. The number of observations is $n = 50$, 100, and 500. To investigate the performance of various penalization procedures, we compare the root mean squared error (RMSE) over $T = 1000$ simulations, which is defined by

$$\text{RMSE} = \frac{1}{T}\left(\sum_{s=1}^{T} \frac{\|\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}^{(s)}\|_F^2}{pm}\right)^{1/2},$$

where $\hat{\boldsymbol{\Lambda}}^{(s)}$ is the estimate of the loading matrix using the sth dataset. We also compare the rate of nonzero factor loadings for Models (A) and (B). Because the loading matrix is not identifiable

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

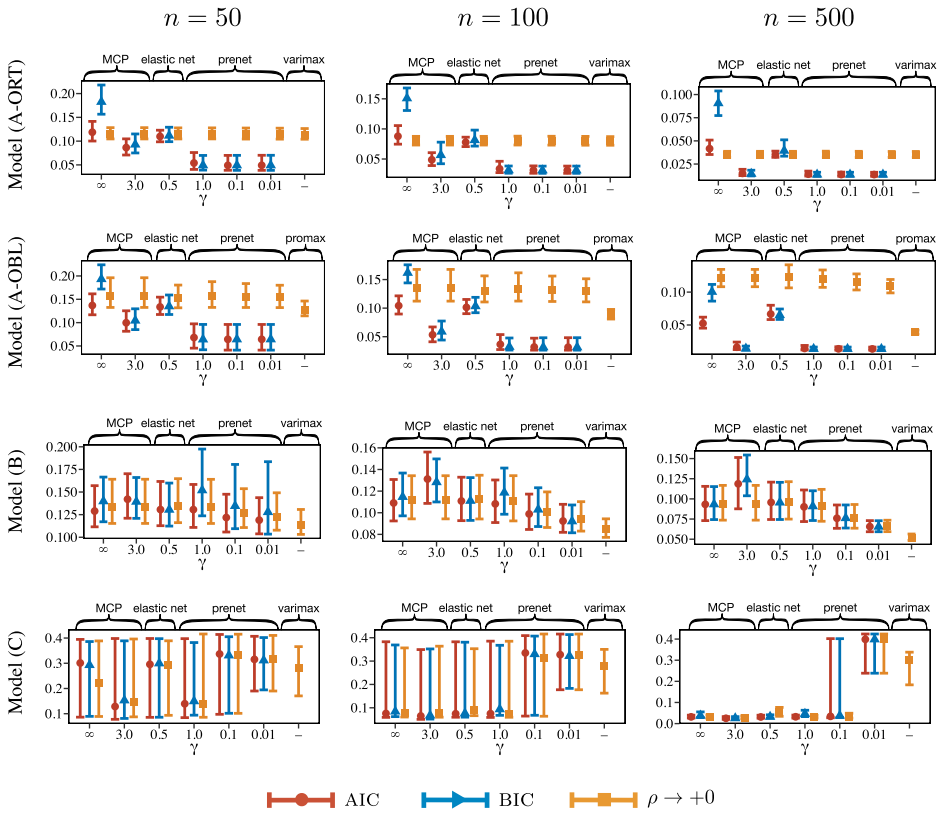


FIGURE 2.
RMSE of factor loadings. The upper and lower bars represent 95th and 5th percentiles, respectively. Here, "$\rho \to +0$" denotes a limit of the estimate of the factor loadings, $\lim_{\rho \to +0} \hat{\Lambda}_\rho$, which corresponds to the factor rotation.

due to permutation and sign of columns, we change the permutation and sign such that RMSE is minimized. It is not fair to compare the rate of nonzero loadings for $\rho \to +0$, because the estimated loading matrix cannot become sparse. Thus, we apply the hard-thresholding with a cutoff value being 0.1, the default of the `loadings` class in `R`.

The results for RMSE and the rate of nonzero factor loadings are depicted in Figs. 2 and 3, respectively. For reference, true positive rate (TPR) and false positive rate (FPR) of the loading matrix are depicted in Figures S1.4 and S1.5 in the supplemental material. The range of the error bar indicates 90% confidence interval; we calculate 5% and 95% quantiles over 1000 simulation results and use them as the boundaries of the error bar. We obtain the following empirical observations from these figures.

Model (A-ORT): The prenet penalization outperforms the existing methods in terms of RMSE when the regularization parameter is selected by the model selection criteria. It is also seen that the performance of the prenet is almost independent of γ. When $\rho \to +0$, all of the estimation procedures yield similar performances. When the ρ is selected by the model selection criteria, the rate of nonzero loadings of the prenet is almost 0.25, that is the true rate. Considering the TPR result in Figure S1.4 in the supplemental material, the prenet with AIC or BIC

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

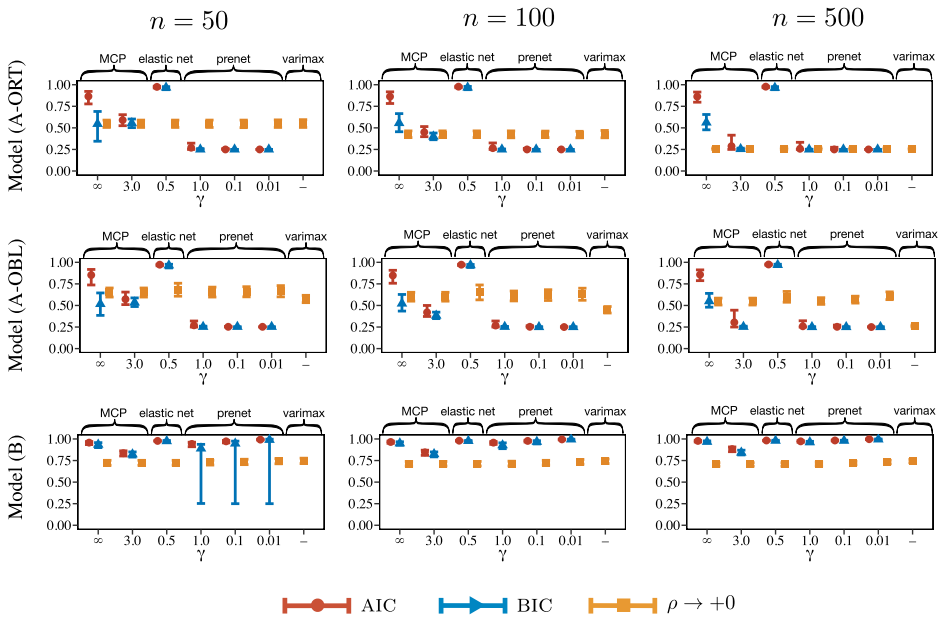


FIGURE 3.
Rate of nonzero factor loadings. The upper and lower bars represent 95th and 5th percentiles, respectively. Here, "$\rho \to +0$" denotes a limit of the estimate of the factor loadings, $\lim_{\rho \to +0} \hat{\Lambda}_\rho$, which corresponds to the factor rotation.

correctly estimates the true zero/nonzero pattern. Meanwhile, the MCP, elastic net, and lasso tend to select a denser model than the true one.

Model (A-OBL): The result for the oblique model is similar to that for the orthogonal model, but the oblique model tends to produce larger RMSEs than the orthogonal model for most cases. The $\rho \to +0$ produces larger RMSE than that with the regularization parameter ρ selected by model selection criteria. This is probably because the loss function becomes flat as $\rho \to +0$. Therefore, the regularization may help improve the accuracy. We note that the promax rotation, which corresponds to $\rho \to +0$, turns out to be stable.

Model (B): When n is large, the prenet with small γ and varimax rotation produces small RMSEs. Because the true values of cross-loadings (small loadings) are close to but not exactly zero, the L_1 type regularization that induces a sparse loading matrix does not work well. The prenet with $\gamma = 0.01$ achieves the sparse estimation but produces a loading matrix that is similar to the quartimin rotation, resulting in a nonsparse loading matrix. We also observe that the prenet with BIC sometimes results in too sparse loading matrix when $n = 50$.

Model (C): For small n, all methods result in large RMSE. For large n, the L_1 regularization methods, including the lasso, MCP, elastic net, and prenet with large γ yield small RMSE. However, the prenet with small γ and varimax rotation, which tend to estimate non-sparse loading matrix, produce large RMSE. Indeed, the average value of loading matrix in the supplemental material shows that the prenet with small γ is biased. Furthermore, the varimax rotation with true loading matrix does not approximate the true one. Therefore, when the loading matrix is sparse but does not have the perfect simple structure, the lasso-type penalization or prenet with $\gamma = 1$ would perform well.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

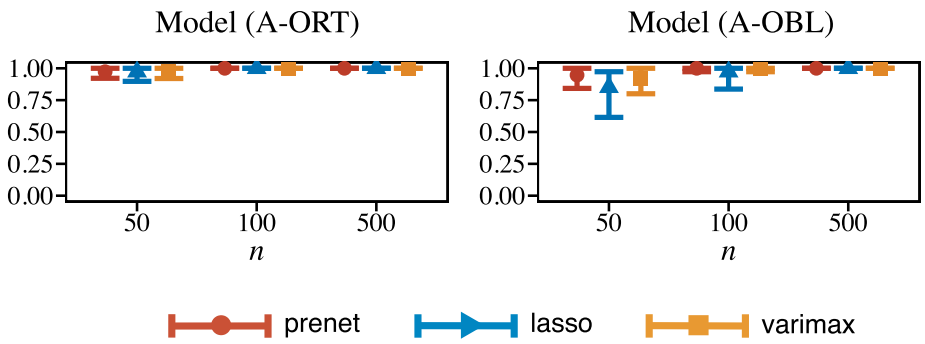


FIGURE 4.
Adjusted Rand Index (ARI) of the clustering results.

5.3. *Investigation of Clustering via Perfect Simple Structure Estimation*

As shown in Proposition 1, our proposed method allows the clustering of variables via the perfect simple structure estimation. We investigate the clustering accuracy on Model (A); the true loading matrix has the perfect simple structure, and then we know the true clusters. Figure 4 shows the Adjusted Rand Index (ARI) between true clusters and those obtained by prenet, lasso, and varimax. The range of the error bar indicates 90% confidence interval; we calculate 5% and 95% quantiles over 1000 simulation results and use them as the boundaries of the error bar.

The clustering via prenet is achieved by perfect simple structure estimation. The lasso and varimax cannot always estimate the perfect simple structure. Therefore, we estimate the clusters as follows: for ith row vector of $\hat{\mathbf{\Lambda}}$, say $\hat{\boldsymbol{\lambda}}_i = (\hat{\lambda}_{i1}, \ldots, \hat{\lambda}_{im})^T$, the ith variable belongs to jth cluster, where $j = \underset{j \in \{1,\ldots,m\}}{\arg\max} (|\hat{\lambda}_{ij}|)$. The regularization parameter for the lasso is $\rho \to +0$, which corresponds to a special case of the component loss criterion (Jennrich, 2004, 2006) with MLE.

The result of Fig. 4 shows that the prenet and varimax result in almost identical ARIs and are slightly better than the lasso when $n = 50$ on Model (A-ORT). All methods correctly detect the true clusters when $n = 100$ and $n = 500$. For Model (A-OBL), the prenet performs slightly better than the varimax when $n = 50$. As with the orthogonal model, the prenet and varimax correctly detect the true clusters when $n = 100$ and $n = 500$. The lasso performs worse than the other two methods for small sample sizes, suggesting that the prenet or varimax would be better if the clustering of variables is the purpose of the analysis.

6. Analysis of Big Five Personality Traits

We apply the prenet penalization to the survey data regarding the big five personality traits collected from Open Source Psychometrics Project (https://openpsychometrics.org/). Other real data applications (electricity demand and fMRI data) are described in Section S2 of the supplemental material. $n = 8582$ responders in the US region are asked to assess their own personality based on 50 questions developed by Goldberg (1992). Each question asks how well it describes the statement of the responders on a scale of 1–5. It is well known that the personality is characterized by five common factors; therefore, we choose $m = 5$. Several earlier researchers showed that the loading matrix may not possess the perfect simple structure due to the small cross-loadings (Marsh et al., 2010, 2013; Booth and Hughes, 2014); therefore, we do not aim at estimating the perfect simple structure with $\rho \to \infty$ in this analysis. We first interpret the estimated model and then investigate the performance of the prenet penalization with ρ selected by model selection

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

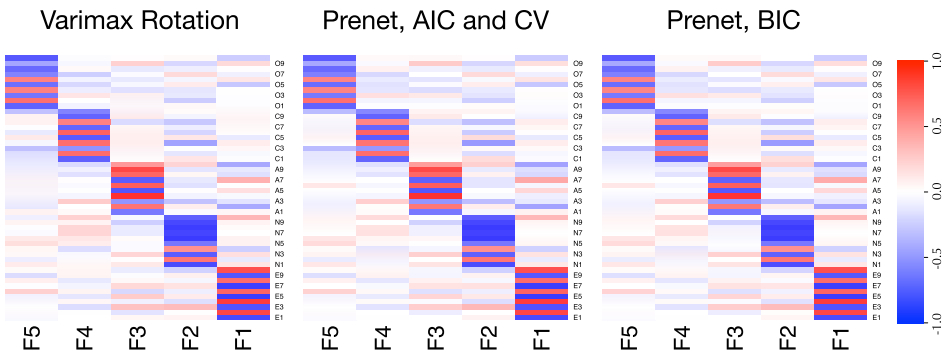


FIGURE 5.
Heatmaps of the loading matrices on big five personality traits data. Each cell corresponds to the factor loading, and the depth of color indicates the magnitude of the value of the factor loading.

TABLE 1.
Factor loadings of four items estimated by the prenet penalization with $\gamma = 0.01$. The regularization parameter, ρ, is selected by the BIC. The cross-loadings whose absolute values are larger than 0.3 are written in bold.

Item	F1	F2	F3	F4	F5
A2: Am not interested in other people's problems.	**0.341**	−0.062	−0.525	−0.020	0.070
A7: Have a soft heart.	**−0.317**	0.089	0.615	0.008	−0.010
A10: Do not have a good imagination.	**0.347**	−0.164	−0.375	0.116	0.082
C4: Change my mood a lot.	−0.083	**0.365**	0.033	−0.548	0.022

criteria for various sample sizes. The impact of the regularization parameter on the accuracy is also studied.

6.1. Interpretation of Latent Factors

We first apply the prenet penalization and the varimax rotation with maximum likelihood estimate and compare the loading matrices estimated by these two methods. With the prenet penalization, we choose a regularization parameter using AIC, BIC, and tenfold cross-validation (CV) with $\gamma = 1$. The regularization parameter selected by the AIC and CV is $\rho = 7.4 \times 10^{-4}$, and that selected by the BIC is $\rho = 2.9 \times 10^{-3}$. The heatmaps of the loading matrices are shown in Fig. 5. The result of Fig. 5 shows that these heatmaps are almost identical; all methods are able to detect the five personality traits appropriately. We also observe that the result is almost independent of γ. These similar results may be due to the large sample sizes.

We explore the estimates of cross-loadings whose absolute values are larger than 0.3; it would be reasonable to regard that these cross-loadings affect the items. There exists four items that include such large absolute values of cross-loadings, and factor loadings related to these four items are shown in Table 1.

The loading matrix is estimated by the prenet penalization with $\gamma = 0.01$. The regularization parameter ρ is selected by the BIC. The five factors represent "F1: extraversion," "F2: neuroticism," "F3: agreeableness," "F4: conscientiousness," and "F5: openness to experience." For reference, the complete loading matrix is shown in Tables S2.5 and S2.6 of the supplemental material.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

TABLE 2.
The number of times that the absolute values of four cross-loadings exceed 0.3. For regularization methods, ρ is selected by the BIC.

	Prenet ($\gamma = 1$)	Prenet ($\gamma = 0.01$)	Lasso	MCP	Varimax
A2	25	85	35	69	81
A7	10	72	18	45	61
A10	29	90	40	68	80
C4	20	94	49	78	94

The three items, A2, A7, and A10, are affected by "F1: extraversion" and "F3: agreeableness." The main and cross-loadings on the same item have opposite signs. We may make a proper interpretation of the factor loadings. For example, as for the question "A7: Have a soft heart," it is easy to imagine some people who have an extraversion cannot be kind. They are interested in a profit from a person rather than the situation that the person is in now; thus, they can become selfish to get the profit even if the person's feelings are hurt. Booth and Hughes (2014) also reported similar results of such cross-loadings. They mentioned that these cross-loadings were due to the overlap in content between extraversion and agreeableness.

Furthermore, we perform $M = 100$ subsampling simulation with $n = 500$ to investigate whether these cross-loadings can be found with small sample sizes. We compare the performance of four estimation methods: the prenet, lasso, MCP, and varimax rotation. For regularization methods, ρ is selected by the BIC. We set $\gamma = 1$ and $\gamma = 0.01$ for the prenet and $\gamma = 3$ for MCP.

Table 2 shows the number of times that the absolute values of these four cross-loadings exceed 0.3. The results show that the prenet with $\gamma = 0.01$ most frequently identifies these four cross-loadings among $M = 100$ simulations.

6.2. *RMSE Comparison*

We investigate the performance of the prenet in terms of estimation accuracy of the loading matrix through subsampling simulation. First, the dataset is randomly split into two datasets, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, without replacement. The sample sizes of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $n/2 = 4291$. The $\boldsymbol{X}_1$ is used for estimating a loading matrix with large sample sizes; we perform the varimax rotation with MLE and regard the estimated loading matrix as a true loading matrix, say $\boldsymbol{\Lambda}_{\text{true}}$. The true loading matrix is almost identical to the loading matrix obtained by the varimax with the entire dataset. We remark that the true loading matrix is also similar to the Model (B) of the Monte Carlo simulation described in Sect. 5.1.

The performance is investigated by subsampling the observations from $\boldsymbol{X}_2$ with $n = 100$ and $n = 500$. Figure 6 depicts RMSE and rate of nonzero loadings for n random subsampled data over 100 simulations. The RMSE is defined as

$$\text{RMSE} = \frac{1}{100}\left(\sum_{s=1}^{100} \frac{\|\boldsymbol{\Lambda}_{\text{true}} - \hat{\boldsymbol{\Lambda}}^{(s)}\|_F^2}{pm}\right)^{1/2},$$

where $\hat{\boldsymbol{\Lambda}}^{(s)}$ is the estimate of the loading matrix using the sth subsampled data. We apply the lasso, MCP with $\gamma = 3$, prenet with $\gamma = 1, 0.1, 0.01$, and the varimax rotation with MLE. The regularization parameter ρ is selected by the AIC, BIC, and tenfold CV. We also compute the loading matrix when $\rho \to +0$, which results in the solution of the factor rotation with MLE.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

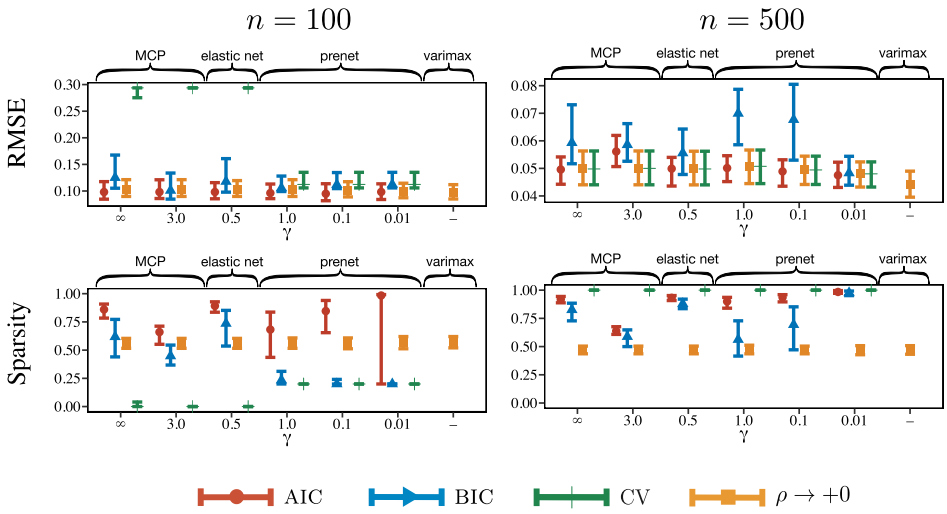


FIGURE 6.
RMSE and rate of nonzero loadings when $n = 100$ and 500. Here, "$\rho \to +0$" denotes a limit of the estimate of the factor loadings, $\lim_{\rho \to +0} \hat{\Lambda}_\rho$, which corresponds to the factor rotation.

The nonzero pattern of the loading matrix for $\rho \to +0$ is estimated by a hard-thresholdings with a cutoff value being 0.1. The range of the error bar indicates 90% confidence interval over 100 simulations.

We have the following empirical observations from Fig. 6:

- The smaller the number of observations is, the sparser the solution is. An earlier study has shown that the model selection criterion can select a parsimonious model with small sample sizes in general frameworks (Cudeck and Henly, 1991).
- The BIC results in larger RMSE and lower rate of nonzero loadings than other criteria, especially for small sample sizes. Therefore, the BIC tends to select sparse solutions, and some of the small nonzero factor loadings are estimated to be zero.
- When the lasso or MCP is applied, the CV results in poor RMSE when $n = 100$. This is because the estimated loading matrix is too sparse; it becomes (almost) zero matrix. When the prenet is applied, such a loading matrix cannot be obtained thanks to Proposition 1.
- With the prenet, small γ tends to estimate a dense loading matrix and produce good RMSE. A similar tendency is found in Model (B) of the Monte Carlo simulation, described in Sect. 5.2.

6.3. *Impact of Tuning Parameters*

We investigate the impact of the tuning parameters (ρ, γ) on the estimation of the loading matrix. Figure 7 depicts the heatmaps of the loading matrices for various values of tuning parameters on the MCP and the prenet penalization. We find the tuning parameters so that the degrees of sparseness (proportion of nonzero values) of the loading matrix are approximately 20%, 25%, 40%, and 50%. For the MCP, we set $\gamma = \infty$ (i.e., the lasso), 5.0, 2.0, and 1.01. For prenet penalty, the values of gamma are $\gamma = 1.0,\ 0.5,$ and 0.01. Each cell describes the elements of the factor loadings as with Fig. 5.

From Fig. 7, we obtain the empirical observations as follows.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

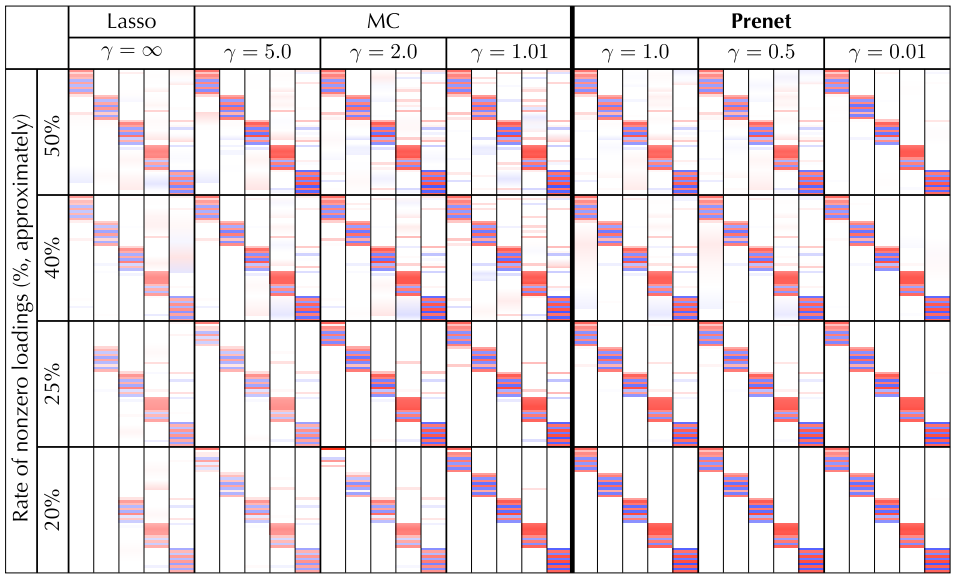


FIGURE 7.
Heatmaps of the loading matrices on big five personality traits data for various values of tuning parameters on the MCP and the prenet penalization.

- With the prenet penalization, the characteristic of five personality traits are appropriately extracted for any values of tuning parameters, which suggests that the prenet penalization is relatively robust against the tuning parameters.
- The prenet penalization is able to estimate the perfect simple structure when the degree of sparseness is 20%. On the other hand, with the MCP, we are not able to estimate the perfect simple structure even when γ is sufficiently small.
- With the lasso, the number of factors becomes less than five when the degrees of sparsity are 20% and 25%; the five personality traits are not able to be found. When the value of γ is not sufficiently large, the MCP produces five factor model.

7. Discussion

We proposed a prenet penalty, which is based on the product of a pair of parameters in each row of the loading matrix. The prenet aims at the estimation of not only sparsity but also the simplicity. Indeed, the prenet is a generalization of the quartimin criterion, one of the most popular oblique techniques for simple structure estimation. Furthermore, the prenet is able to estimate the perfect simple structure, which gave us a new variables clustering method using factor models. The clustering of variables opens the door to the application of the factor analysis to a wide variety of sciences, such as image analysis, neuroscience, marketing, and biosciences.

The prenet penalization has two different purposes of analysis: clustering of variables and exploratory factor analysis. The way of using the prenet penalization depends on the purpose of the analysis. When the purpose of the analysis is the clustering of variables, the regularization parameter is set to be $\rho \to \infty$ to achieve the perfect simple structure estimation. It is shown that the prenet performs better than the lasso and varimax in terms of clustering accuracy, as described in Sect. 5.3. Furthermore, the real data analyses in Section S2 in the supplemental material show the superiority of the prenet over the conventional clustering methods, such as the

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

k-means clustering. When the purpose of the analysis is exploratory factor analysis, the perfect simple structure estimation is not necessarily needed. In this case, the regularization parameter is selected by the model selection criteria. The numerical results show that the prenet penalization performs well when an appropriate value of γ is selected.

Over a span of several decades, a number of researchers have developed methods for finding a simple loading matrix (e.g., Kaiser, 1958; Hendrickson and White, 1964) in the Thurstonian sense (Thurstone, 1947). As the simple structure is a special case of sparsity, it seems the lasso-type sparse estimation is more flexible than the prenet. Indeed, the recent trend in the exploratory factor analysis literature is to find a loading matrix that possesses the sparsity rather than simplicity (Jennrich, 2004, 2006; Trendafilov, 2013; Scharf and Nestler, 2019b, 2019a).

Nevertheless, the lasso-type sparse estimation is not as flexible as expected. As mentioned in Model (B) in Monte Carlo simulation and Section S1 in the supplemental material, the lasso cannot often approximate the true loading matrix when the cross-loadings are not exactly but close to zero. Because some factor loadings are estimated to be *exactly* zero with the lasso, some other factor loadings turn out to be excessively large, which causes the difficulty in interpretation.

For this reason, we believe *both sparsity and simplicity* play important roles in the interpretation. The sparse estimation automatically produces the nonzero pattern of the loading matrix, which allows us to interpret the latent factors easily. In addition, simplicity is also helpful for the interpretation, as shown in Thurstone (1947). The prenet penalization is able to achieve simplicity and sparsity simultaneously. Indeed, a sparse loading matrix is estimated thanks to the penalty based on the absolute term in $\sum_{i,j,k} |\lambda_{ij}\lambda_{ik}|$. In addition, simplicity is also achieved because it generalizes the quartimin criterion that often produces a simple structure (Jennrich and Sampson, 1966). Furthermore, with a large value of the regularization parameter, the loading matrix enjoys the perfect simple structure. Meanwhile, the existing methods cannot always produce a loading matrix that is both sparse and simple. For example, the lasso produces a loading matrix that is sparse but not always simple.

The structural equation modeling (SEM) has been widely used in the social and behavioral sciences. The SEM covers a wide variety of statistical models, including the factor analysis model and the regression model. An analyst develops an assumption of causal relationship and determines whether the assumption is correct or not by testing the hypothesis or evaluating the goodness of fit indices. Recently, several researchers have proposed regularized structural equation models (Jacobucci et al., 2016; Huang et al., 2017; Huang, 2018). The analyst set lasso-type penalties to specific model parameters to conduct an automatic selection of the causal relationship, enhancing the flexibility in model specification. The application of the prenet to the SEM would be an interesting future research topic.

The lasso-type regularization extracts only the nonzero pattern of parameters. In some cases, the analyst needs to detect not only the nonzero pattern of parameters but also a more complex parameter structure. The penalty must be determined depending on the structure of the parameter. For example, when the analyst needs to estimate either θ_1 or θ_2 to be zero, the prenet penalty would be more useful than the lasso. More generally, when one of $\theta_1, \ldots, \theta_k$ is exactly or close to zero, we may use the Geomin-type penalty, $\prod_{j=1}^{k} |\theta_j|$. An application of a penalty that leads to structured sparsity would further enhance the flexibility of the analysis but beyond the scope of this research. We would like to take this as a future research topic.

Another interesting extension to the prenet penalization is the clustering of not only variables but also observations. This method is referred to as biclustering (e.g., Tan and Witten, 2014; Flynn and Perry, 2020). To achieve this, we may need an entirely new formulation along with an algorithm to compute the optimal solution. This extension should also be a future research topic.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

Supplemental Materials

Further numerical and theoretical investigations Analyses of resting-state fMRI data, electricity demand data, a loading matrix of big five data, and comparison with Geomin criterion.
R-package fanc R-package `fanc` containing code that performs our proposed algorithm.
Loadings Average of the estimated loading matrices for Monte Carlo simulations in Sect. 5 with excel files.

Acknowledgments

The authors would like to thank an associate editor and three reviewers for valuable comments and suggestions that improve the quality of the paper considerably. This research was supported in part by JSPS KAKENHI Grant (JP19K11862 and JP22H01139 to KH; JP20K19756, JP20H00601 to YT) and MEXT Project for Seismology toward Research Innovation with Data of Earthquake (STAR-E) Grant Number JPJ010217.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A. Proofs

A.1. Proof of Proposition 2

Because of Proposition 1, with the prenet, $\hat{\lambda}_{ij}\hat{\lambda}_{ik} = 0$ as $\rho \to \infty$. Thus, the prenet solution satisfies (9) as $\rho \to \infty$. We only need to show that the minimization problem of loss function $\ell_{\mathrm{ML}}(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ is equivalent to that of $\|\boldsymbol{S} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\|_F^2$. The inverse covariance matrix of the observed variables is expressed as

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I}_m)^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}.$$

Because $\boldsymbol{\Lambda}^T\boldsymbol{\Lambda} = \boldsymbol{I}_m$, we obtain

$$\boldsymbol{\Sigma}^{-1} = \alpha^{-1}\boldsymbol{I}_p - \frac{\alpha^{-2}}{\alpha^{-1}+1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T.$$

The determinant of $\boldsymbol{\Sigma}$ can be calculated as

$$|\boldsymbol{\Sigma}| = \alpha^{p-m}(1+\alpha)^m.$$

Then, the discrepancy function in (3) is expressed as

$$\frac{1}{2}\left\{\mathrm{tr}(\alpha^{-1}\boldsymbol{S}) - \frac{\alpha^{-2}}{\alpha^{-1}+1}\mathrm{tr}\left(\boldsymbol{\Lambda}^T\boldsymbol{S}\boldsymbol{\Lambda}\right) + p\log\alpha + m\log\left(1+\frac{1}{\alpha}\right) - \log|\boldsymbol{S}| - p\right\}.$$

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

Because α is given and $\|\boldsymbol{S}-\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\|_F^2 = -2\mathrm{tr}\left(\boldsymbol{\Lambda}^T\boldsymbol{S}\boldsymbol{\Lambda}\right)+C$. with constant value C, we can derive (10).

A.2. Proof of Proposition 3

Recall that $\hat{\boldsymbol{\theta}}$ is an unpenalized estimator that satisfies $\ell(\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}\in\Theta}\ell(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_q$ is a quartimin solution obtained by the following problem.

$$\min_{\boldsymbol{\theta}\in\Theta} P_{\mathrm{qmin}}(\boldsymbol{\Lambda}), \text{ subject to } \ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}).$$

First, we show that

$$\lim_{n\to\infty} d(\hat{\boldsymbol{\theta}}_q, \Theta_q^*) = 0 \ \ a.s. \tag{A.1}$$

From the assumptions, as the same manner of Chapter 6 in Pfanzagl (1994), we can obtain the following strong consistency.

$$\lim_{n\to\infty} d(\hat{\boldsymbol{\theta}}, \Theta_*) = 0 \ \text{ and } \ \lim_{n\to\infty} d(\hat{\boldsymbol{\theta}}_{\rho_n}, \Theta_*) = 0 \quad a.s., \tag{A.2}$$

where $\Theta_* := \{\boldsymbol{\theta}\in\Theta \mid \ell_*(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in\Theta}\ell_*(\boldsymbol{\theta})\}$. When $\lim_{n\to\infty} d(\hat{\boldsymbol{\theta}}, \Theta_*) = 0$, for all $\epsilon > 0$, by taking n large enough, we have

$$\|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_*\|_F < \epsilon \quad a.s.$$

for some $(\mathrm{vec}(\boldsymbol{\Lambda}_*)^T, \mathrm{diag}(\boldsymbol{\Psi}_*)^T, \mathrm{vech}(\boldsymbol{\Phi}_*)^T)^T \in \Theta_*$. From the uniform continuity of P_{qmin} on Θ and the fact that $\|\hat{\boldsymbol{\Lambda}}\boldsymbol{T} - \boldsymbol{\Lambda}_*\boldsymbol{T}\|_F = \|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_*\|_F$ for any $\boldsymbol{T}\in\mathcal{O}(m)$, we have

$$\sup_{\boldsymbol{T}\in\mathcal{O}(m)} |P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}\boldsymbol{T}) - P_{\mathrm{qmin}}(\boldsymbol{\Lambda}_*\boldsymbol{T})| < \epsilon \quad a.s. \tag{A.3}$$

Write $\hat{\boldsymbol{T}} := \arg\min_{\boldsymbol{T}\in\mathcal{O}(m)} P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}\boldsymbol{T})$ and $\boldsymbol{T}_* := \arg\min_{\boldsymbol{T}\in\mathcal{O}(m)} P_{\mathrm{qmin}}(\boldsymbol{\Lambda}_*\boldsymbol{T})$. We have

$$P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{T}}) - P_{\mathrm{qmin}}(\boldsymbol{\Lambda}_*\hat{\boldsymbol{T}}) \le P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{T}}) - P_{\mathrm{qmin}}(\boldsymbol{\Lambda}_*\boldsymbol{T}_*) \le P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}\boldsymbol{T}_*) - P_{\mathrm{qmin}}(\boldsymbol{\Lambda}_*\boldsymbol{T}_*).$$

From this, it follows that

$$|P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{T}}) - P_{\mathrm{qmin}}(\boldsymbol{\Lambda}_*\boldsymbol{T}_*)| \le \sup_{\boldsymbol{T}\in\mathcal{O}(m)} |P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}\boldsymbol{T}) - P_{\mathrm{qmin}}(\boldsymbol{\Lambda}_*\boldsymbol{T})|.$$

Thus, using (A.3), we obtain (A.1).
Next, as the similar manner of Proposition 15.1 in Foucart and Rauhut (2013), we prove $\lim_{n\to\infty} d(\hat{\boldsymbol{\theta}}_{\rho_n}, \Theta_q^*) = 0, a.s.$ By the definition of $\hat{\boldsymbol{\theta}}_{\rho_n}$, for any $\rho_n > 0$ we have

$$\ell(\hat{\boldsymbol{\theta}}_{\rho_n}) + \rho_n P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}_{\rho_n}) \le \ell(\hat{\boldsymbol{\theta}}_q) + \rho_n P_{\mathrm{qmin}}(\hat{\boldsymbol{\Lambda}}_q) \tag{A.4}$$

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

and

$$\ell(\hat{\boldsymbol{\theta}}_{\rho_n}) \geq \ell(\hat{\boldsymbol{\theta}}_q). \tag{A.5}$$

Combining (A.1), (A.4), (A.5), we obtain

$$P_{\text{qmin}}(\hat{\boldsymbol{\Lambda}}_{\rho_n}) \leq P_{\text{qmin}}(\hat{\boldsymbol{\Lambda}}_q) \to P_{\text{qmin}}(\boldsymbol{\Lambda}_q^*) \quad a.s.$$

for some $(\text{vec}(\boldsymbol{\Lambda}_q^*)^T, \text{diag}(\boldsymbol{\Psi}_q^*)^T, \text{vech}(\boldsymbol{\Phi}_q^*)^T)^T \in \Theta_q^*$. Therefore, we have

$$\lim_{n\to\infty} P_{\text{qmin}}(\hat{\boldsymbol{\Lambda}}_{\rho_n}) \leq P_{\text{qmin}}(\boldsymbol{\Lambda}_q^*) \quad a.s.$$

As shown in (A.2), $\lim_{n\to\infty} d(\hat{\boldsymbol{\theta}}_{\rho_n}, \Theta_*) = 0$ $a.s.$, and $\boldsymbol{\Lambda}_q^*$ is a minimizer of $P_{\text{qmin}}(\cdot)$ over Θ_*, so that the proof is complete.

B. Detail of the Algorithm

B.1. Update Equation of EM Algorithm for Fixed Tuning Parameters

We provide update equations of factor loadings and unique variances when ρ and γ are fixed. Suppose that $\boldsymbol{\Lambda}_{\text{old}}$, $\boldsymbol{\Psi}_{\text{old}}$, and $\boldsymbol{\Phi}_{\text{old}}$ are the current values of parameters. The parameter can be updated by minimizing the negative expectation of the complete-data penalized log-likelihood function with respect to $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ (e.g., Hirose and Yamamoto, 2014):

$$\begin{aligned} Q(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) = &\sum_{i=1}^{p} \left(\log \psi_i + \frac{s_{ii} - 2\boldsymbol{\lambda}_i^T \boldsymbol{b}_i + \boldsymbol{\lambda}_i^T \boldsymbol{A} \boldsymbol{\lambda}_i}{\psi_i} \right) + \log |\boldsymbol{\Phi}| + \text{tr}(\boldsymbol{\Phi}^{-1}\boldsymbol{A}) \\ &+ \rho P(\boldsymbol{\Lambda}) + C, \end{aligned} \tag{A.6}$$

where C is a constant and

$$\boldsymbol{A} = \boldsymbol{M}^{-1} + \boldsymbol{M}^{-1}\boldsymbol{\Lambda}_{\text{old}}^T \boldsymbol{\Psi}_{\text{old}}^{-1} \boldsymbol{S} \boldsymbol{\Psi}_{\text{old}}^{-1} \boldsymbol{\Lambda}_{\text{old}} \boldsymbol{M}^{-1}, \tag{A.7}$$

$$\boldsymbol{b}_i = \boldsymbol{M}^{-1}\boldsymbol{\Lambda}_{\text{old}}^T \boldsymbol{\Psi}_{\text{old}}^{-1} \boldsymbol{s}_i, \tag{A.8}$$

$$\boldsymbol{M} = \boldsymbol{\Lambda}_{\text{old}}^T \boldsymbol{\Psi}_{\text{old}}^{-1} \boldsymbol{\Lambda}_{\text{old}} + \boldsymbol{\Phi}_{\text{old}}^{-1}. \tag{A.9}$$

Here, $\boldsymbol{s}_i$ is the ith column vector of $\boldsymbol{S}$.

In practice, minimization of (A.6) is difficult, because the prenet penalty consists of nonconvex functions. Therefore, we use a coordinate descent algorithm to obtain updated loading matrix $\boldsymbol{\Lambda}_{\text{new}}$. Let $\tilde{\lambda}_i^{(j)}$ be a $(m-1)$-dimensional vector $(\tilde{\lambda}_{i1}, \tilde{\lambda}_{i2}, \ldots, \tilde{\lambda}_{i(j-1)}, \tilde{\lambda}_{i(j+1)}, \ldots, \tilde{\lambda}_{im})^T$. The parameter λ_{ij} can be updated by maximizing (A.6) with the other parameters $\tilde{\lambda}_i^{(j)}$ and with $\boldsymbol{\Psi}$ being fixed, that is, we solve the following problem.

$$\tilde{\lambda}_{ij} = \arg\min_{\lambda_{ij}} \frac{1}{2\psi_i} \left\{ a_{jj}\lambda_{ij}^2 - 2\left(b_{ij} - \sum_{k\neq j} a_{kj}\tilde{\lambda}_{ik} \right) \lambda_{ij} \right\}$$

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

$$
\begin{aligned}
&+\rho\left[\left\{\frac{1}{2}(1-\gamma)\sum_{k\neq j}\tilde{\lambda}_{ik}^2\right\}\lambda_{ij}^2+\left(\gamma\sum_{k\neq j}|\tilde{\lambda}_{ik}|\right)|\lambda_{ij}|\right]\\
&=\arg\min_{\lambda_{ij}}\frac{1}{2\psi_i}\left\{(a_{jj}+\beta)\lambda_{ij}^2-2\left(b_{ij}-\sum_{k\neq j}a_{kj}\tilde{\lambda}_{ik}\right)\lambda_{ij}\right\}+\rho\xi|\lambda_{ij}|\\
&=\arg\min_{\lambda_{ij}}\frac{1}{2}\left(\lambda_{ij}-\frac{b_{ij}-\sum_{k\neq j}a_{kj}\tilde{\lambda}_{ik}}{a_{jj}+\beta}\right)^2+\frac{\psi_i\rho\xi}{a_{jj}+\beta}|\lambda_{ij}|, \qquad \text{(A.10)}
\end{aligned}
$$

where

$$
\begin{aligned}
\beta&=\rho\psi_i(1-\gamma)\sum_{k\neq j}\tilde{\lambda}_{ik}^2,\\
\xi&=\gamma\sum_{k\neq j}|\tilde{\lambda}_{ik}|.
\end{aligned}
$$

This is equivalent to minimizing the following penalized squared error loss function.

$$
S(\tilde{\theta})=\arg\min_{\theta}\left\{\frac{1}{2}(\theta-\tilde{\theta})^2+\rho^*|\theta|\right\}. \qquad \text{(A.11)}
$$

The solution $S(\tilde{\theta})$ can be expressed in a closed form using the soft thresholding function:

$$
S(\tilde{\theta})=\text{sgn}(\tilde{\theta})(|\tilde{\theta}|-\rho^*)_+, \qquad \text{(A.12)}
$$

where $A_+=\max(A,0)$.
For given $\mathbf{\Lambda}_{\text{new}}$, the parameters of unique variances and factor correlations, say $\mathbf{\Psi}_{\text{new}}$ and Φ_{new}, are expressed as

$$
\begin{aligned}
\psi_i^{\text{new}}&=s_{ii}-2(\boldsymbol{\lambda}_i^{\text{new}})^T\boldsymbol{b}_i+(\boldsymbol{\lambda}_i^{\text{new}})^T\boldsymbol{A}\boldsymbol{\lambda}_i^{\text{new}}\quad(i=1,\ldots,p),\\
\Phi_{\text{new}}&=\arg\min_{\Phi}\{\log|\Phi|+\text{tr}(\Phi^{-1}\mathbf{A})\},
\end{aligned}
$$

where ψ_i^{new} is the ith diagonal element of $\mathbf{\Psi}_{\text{new}}$, and $\boldsymbol{\lambda}_i^{\text{new}}$ is the ith row of $\mathbf{\Lambda}_{\text{new}}$. The new value Φ_{new} may not be expressed in an explicit form, because all of the diagonal elements of Φ are fixed by 1. Thus, Φ_{new} is obtained by the Broyden–Fletcher–Goldfarb–Shanno optimization procedure.

B.2. Algorithm Complexity

The complexity for our proposed algorithm for the orthogonal case is considered. To update $\mathbf{\Lambda}$, we need a matrix $\boldsymbol{A}$ in (A.7). The matrix computation of $\boldsymbol{A}$ requires $O(p^2m)$ operation (Zhao et al., 2007). In the coordinate descent algorithm, we need to compute $\tilde{\theta}$ in (A.11) for each step, in which $O(m)$ operation is required. For simplicity, we consider the case where the number of cycles of the coordinate descent algorithm is one. We remark that our algorithm converges to the (local) optima for this case because both EM and coordinate descent algorithms monotonically decrease the objective function at each iteration. In this case, we need $O(m)$ to update λ_{ij} $(i=1,\ldots,p;$

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

$j = 1, \ldots, m$), and therefore, $O(pm^2)$ operation is required to update $\boldsymbol{\Lambda}$ in the coordinate descent algorithm. As a result, the algorithm complexity to update $\boldsymbol{\Lambda}$ is $O(p^2m) + O(pm^2) = O(p^2m)$. For the update of $\boldsymbol{\Psi}$, we need $O(pm^2)$ operation because the computation of $(\boldsymbol{\lambda}_i^{\text{new}})^T \boldsymbol{A} \boldsymbol{\lambda}_i^{\text{new}}$ in (A.13) requires $O(m^2)$ operation for $i = 1, \ldots, p$. Note that when the unpenalized maximum likelihood estimation is conducted, the order $O(pm)$ is achieved (Zhao et al., 2007); however, this order may not be achieved for the prenet penalization.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd International symposium on information theory* (pp. 267–81). Budapest: Akademiai Kiado.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 5).

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397–438.

Bernaards, C. A., & Jennrich, R. I. (2003). Orthomax rotation and perfect simple structure. *Psychometrika, 68*(4), 585–588.

Bhlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications* (1st ed.). Berlin: Springer.

Booth, T., & Hughes, D. J. (2014). Exploratory structural equation modeling of personality data. *Assessment, 21*(3), 260–271.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111–150.

Caner, M., & Han, X. (2014). Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. *Journal of Business & Economic Statistics, 32*(3), 359–374.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika, 18*(1), 23–38.

Choi, J., Zou, H., & Oehlert, G. (2011). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface, 3*(4), 429–436.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the" problem" of sample size: A clarification. *Psychological Bulletin, 109*(3), 512.

Ding, C. H. Q., He, X., & Simon, H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM* (Vol. 5, pp. 606–610). SIAM.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.

Flynn, C., & Perry, P. (2020). Profile likelihood biclustering. *Electronic Journal of Statistics, 14*(1), 731–768.

Foucart, S., & Rauhut, H. (2013). *A mathematical introduction to compressive sensing*. Berlin: Springer.

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

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment, 4*(1), 26.

Hattori, M., Zhang, G., & Preacher, K. J. (2017). Multiple local solutions and geomin rotation. *Multivariate Behavioral Research, 52*(6), 1–12.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17*(1), 65–70.

Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis, 79,* 120–132.

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing, 25*(5), 863–875.

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 71*(3), 499–522.

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika, 82*(2), 329–354.

Hui, F. K. C., Tanaka, E., & Warton, D. I. (2018). Order selection and sparsity in latent variable models via the ordered factor lasso. *Biometrics, 74*(4), 1311–1319.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 555–566.

Jennrich, R. I. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika, 69*(2), 257–273.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika, 71*(1), 173–191.

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31*(3), 313–313.

Jöreskog, K. G., & Goldberger, A. S. (1971). Factor analysis by generalized least squaresd. *ETS Research Bulletin Series, 1971,* i–32.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187–200.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36.

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press

Kiers, H. A. L. (1994). Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika, 59*(4), 567–579.
Lee, S.-Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika, 46*(2), 153–160.
Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*(3), 471–491.
Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology, 49*(6), 1194–1218.
Mazumder, R., Friedman, J., & Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association, 106,* 1125–1138.
Neuhaus, J. O., & Wrigley, C. (1954). The quartimax method: An analytical approach to orthogonal simple structure. *British Journal of Statistical Psychology, 7*(2), 81–91.
Ning, L., & Georgiou, T. T. (2011). Sparse factor analysis via likelihood and ℓ_1 regularization. In *50th IEEE conference on decision and control and European control conference* (pp. 5188–5192).
Pfanzagl, J. (1994). *Parametric statistical theory*. Berlin, Boston: De Gruyter.
Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*(1), 73–103.
Scharf, F., & Nestler, S. (2019a). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling, 26*(4), 576–590.
Scharf, F., & Nestler, S. (2019b). A comparison of simple structure rotation criteria in temporal exploratory factor analysis for event-related potential data. *Methodology, 15,* 43–60.
Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement, 71*(1), 95–113.
Schwarz, G. (1978). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics, 9,* 1135–1151.
Srivastava, S., Engelhardt, B. E., & Dunson, D. B. (2017). Expandable factor analysis. *Biometrika, 104*(3), 649–663.
Tan, K. M., & Witten, D. M. (2014). Sparse biclustering of transposable data. *Journal of Computational and Graphical Statistics, 23*(4), 985–1008.
Thurstone, L. L. (1935). *The vectors of mind. The vectors of mind*. Chicago: University of Chicago Press.
Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58,* 267–288.
Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*(3), 611–622.
Trendafilov, N. T. (2013). From simple structure to sparse components: A review. *Computational Statistics, 29*(3–4), 431–454.
Trendafilov, N. T., Fontanella, S., & Adachi, K. (2017). Sparse exploratory factor analysis. *Psychometrika, 82*(3), 778–794.
Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory, 55*(5), 2183–2202.
Yamamoto, M., & Jennrich, R. I. (2013). A cluster-based factor rotation. *British Journal of Mathematical and Statistical Psychology, 66*(3), 488–502.
Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. New York: State University of New York Press.
Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38,* 894–942.
Zhao, P., & Yu, B. (2007). On model selection consistency of lasso. *Journal of Machine Learning Research, 7*(2), 2541.
Zhao, J. H., Yu, P. L. H., & Jiang, Q. (2007). ML estimation for factor analysis: EM or non-EM? *Statistics and Computing, 18*(2), 109–123.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association, 101*(476), 1418–1429.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B, 67,* 301–320.
Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics, 35,* 2173–2192.

Manuscript Received: 12 FEB 2021
Final Version Received: 1 FEB 2022
Published Online Date: 23 MAY 2022

https://doi.org/10.1007/s11336-022-09868-4 Published online by Cambridge University Press