Proceedings e report

114

SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

A Bayesian oblique factor model with extension to tensor data

Un modello fattoriale obliquo bayesiano con estensione a dati tensoriali

Michael Jauch, Paolo Giordani, and David Dunson

Abstract In this short paper, we discuss a novel way of constructing prior distributions for correlation matrices and an associated approach to inference. We construct a prior penalizing large correlations, which we incorporate into an oblique factor model and a Candecomp/Parafac model for three-way data. We argue that this choice of prior for the factor correlation matrix, combined with a shrinkage prior for elements of the factor loadings matrix, leads to interpretable solutions. At the meeting we will demonstrate this through applications to real data.
Abstract *In questo short paper discutiamo un nuovo modo di costruire distribuzioni a priori per matrici di correlazione ed i relativi aspetti inferenziali. La distribuzione a priori, costruita in maniera tale da penalizzare correlazioni elevate, viene inserita all'interno di un modello di analisi fattoriale obliqua e del modello Candecomp/Parafac per dati a tre vie. Riteniamo che questa scelta della a priori per la matrice di correlazione fattoriale, combinata con una a priori shrinkage per gli elementi della matrice dei loading fattoriali permette di ottenere soluzioni interpretabili. Al convegno dimostreremo il nostro assunto mediante applicazioni a dati reali*

Key words: oblique factor model, prior for correlation matrices, tensor decomposition, three-mode factor analysis

Michael Jauch
Department of Statistical Science, Duke University, e-mail: michael.jauch@duke.edu

Paolo Giordani
Department of Statistical Sciences, Sapienza University of Rome, e-mail: paolo.giordani@uniroma1.it

David B. Dunson
Department of Statistical Science, Duke University, e-mail: dunson@duke.edu

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

1 Introduction

Factor analysis aims to explain the covariance structure between observed variables as arising from a smaller number of unobserved latent factors. A Gaussian factor model with factor dimension S has the form

$$\boldsymbol{y}_i | \boldsymbol{B}, \boldsymbol{f}_i, \Sigma \sim N(\boldsymbol{B}\boldsymbol{f}_i, \boldsymbol{\Sigma}) \,, \quad \boldsymbol{f}_i \sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \tag{1}$$

where $\boldsymbol{y}_i$ is the centered vector of observed variables corresponding to the ith observation, $\boldsymbol{B}$ is the factor loadings matrix, $\boldsymbol{f}_i$ is the S-dimensional vector of latent factors for observation i, $\boldsymbol{\Omega}$ is the covariance matrix of the latent factors, and $\boldsymbol{\Sigma}$ is a diagonal positive definite matrix. Marginalizing out the latent factors yields

$$\boldsymbol{y}_i | \boldsymbol{B}, \Sigma \sim N(\boldsymbol{0}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{\Sigma}). \tag{2}$$

As is well-known, the factor model is not identifiable without further restrictions on $\boldsymbol{B}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}$. Identifiability assumptions are important in Bayesian computation as a means to ensure that estimation based on posterior samples is meaningful. See [14] for a discussion of identifiability of the oblique factor model. We impose the usual restriction that $\boldsymbol{\Omega}$ be a correlation matrix.

If $\boldsymbol{\Omega}$ is diagonal then the latent factors are uncorrelated (and thus independent) and we obtain the conventional orthogonal factor analysis model. If $\boldsymbol{\Omega}$ is not diagonal, the latent factors are correlated and we obtain the so-called oblique factor model. Many authors have argued that the restriction to uncorrelated factors is too strict. For example, discussing application of factor anaylsis in psychology, Thurstone [19] remarks

> It seems just as unnecessary to require that mental traits shall be uncorrelated in the general population as to require that height and weight be uncorrelated in the general population.

A large body of methodology has been developed for the oblique factor model. For a detailed discussion, see Chapter 12 of Harman [11].

In traditional applications of factor analysis, interest lies in interpreting the latent factors as distinct and scientifically meaningful quantities. Interpretation proceeds by examination of the factor loadings matrix, which relates the latent factors to the observed variables. Interpretation is made easier if the factor loadings matrix possesses a simple structure. An example of a simple structure is near sparsity, in which the factor loadings matrix has a small number of large entries and a large number of small entries. Typically, allowing correlated factors allows for a simpler structure in the factor loadings matrix. However, correlated factors present their own difficulties to interpretation. If two factors are highly correlated, then it becomes impossible to interpret them as distinct quantities. When allowing for correlated factors, we should recognize the tradeoff between factor correlation and complexity of the loadings matrix. We tolerate some of the former if it buys us less of the latter, and vice versa. This idea is captured in a quote from [11]:

> It is clear that a certain simplicity of interpretation is sacrificed upon relinquishing the standard of orthogonality. This disadvantage may be offset, however, if the linear descriptions

> of the variables in terms of correlated factors can be made simpler than in the case of uncorrelated ones. Generally this is possible.

We can address this tradeoff in a Bayesian oblique factor analysis setting through the choice of prior distributions for $\boldsymbol{B}$ and $\boldsymbol{\Omega}$. For those elements of the factor loadings matrix $\boldsymbol{B}$ which are not restricted to be zero for the sake of identifiability, we can take advantage of local-global shrinkage priors which will result in a nearly sparse estimate for $\boldsymbol{B}$ [15]. For the factor correlation matrix $\boldsymbol{\Omega}$, we need a prior on the set of correlation matrices which penalizes factor correlations. Defining such a prior distribution that lends itself to relatively simple and scalable inference is challenging. For a recent approach in the context of Bayesian factor analysis with correlated factors, see [9] which provides extensive references to earlier relevant works.

A main contribution of this short paper is what we believe to be a novel approach to constructing priors for correlation matrices. The construction is based on the observation that, for any $N \times P$ matrix $\boldsymbol{X}$ with unit norm columns, the product $\boldsymbol{X}^T\boldsymbol{X}$ is a correlation matrix. Assigning each column of $\boldsymbol{X}$ a probability distribution having support on the unit sphere then induces a probability distribution on correlation matrices. In the special case that each column of $\boldsymbol{X}$ is independent and uniformly distributed on the unit sphere, we obtain closed-form densities for the correlations which match priors discussed previously by [1, 9], and others. The proposed prior does indeed penalize correlation, and the penalty increases as N increases. In this short paper, we only discuss the special case of independent columns uniformly distributed on the unit sphere, but future work may consider other choices, leading to more flexible distributions for correlation matrices. Inference for parameters lying on the unit sphere can be performed using geodesic Monte Carlo [6], a scalable Markov chain Monte Carlo (MCMC) method which can accomodate parameters lying on manifolds embedded in Euclidean space.

We define a Bayesian oblique factor model using the aforementioned prior for the factor correlation matrix and a global-local shrinkage prior for elements of the factor loadings matrix. We discuss extension of the factor model to tensor valued data with an emphasis on the three-way case. For the conference presentation, we will show applications to real data.

2 A Bayesian model for oblique factor analysis

Suppose we have I observations of J variables. We let $\boldsymbol{y}_i$ be the vector of J centered variables corresponding to observation i. As before, we suppose that

$$\boldsymbol{y}_i = \boldsymbol{B}\boldsymbol{f}_i + \boldsymbol{e}_i\,, \quad \boldsymbol{f}_i \sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \quad i = 1, \ldots, I \tag{3}$$

where $\boldsymbol{B}$ is the $J \times S$ factor loadings matrix, $\boldsymbol{f}_i$ is the $S \times 1$ vector of latent factors for observation i, $\boldsymbol{\Omega}$ is a $S \times S$ correlation matrix, and $\boldsymbol{e}_i$ is a $J \times 1$ vector of errors. The errors are independent and identically-distributed $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_J^2)$ is a diagonal positive definite matrix. In matrix form, we have that

$$\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{B}^T + \boldsymbol{E} \tag{4}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_I)^T$ is the $I \times J$ data matrix, $\boldsymbol{F} = (\boldsymbol{f}_1, ..., \boldsymbol{f}_I)^T$ is the $I \times S$ matrix of latent factors, and $\boldsymbol{E} = (\boldsymbol{e}_1, ..., \boldsymbol{e}_I)^T$ is the $I \times J$ matrix of errors. To complete the Bayesian model specification, we need to define priors for our parameters.

2.1 The prior for $\boldsymbol{\Omega}$

As described in the introduction, let the matrix $\boldsymbol{X}$ have columns $\boldsymbol{x}_1, ..., \boldsymbol{x}_P \in \mathscr{S}_{N-1}$, the $N-1$-dimensional unit sphere. Suppose that the columns of $\boldsymbol{X}$ are independent and uniformly-distributed on $\mathscr{S}_{N-1}$. We then set $\boldsymbol{\Omega} = \boldsymbol{X}^T\boldsymbol{X}$.

Due to the simple construction for $\boldsymbol{\Omega}$, it is possible to derive closed-form expressions describing its distribution [8]. For instance, let ω be an arbitrary off-diagonal element of $\boldsymbol{\Omega}$. Then the density of ω is

$$p(\omega) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)} (1-\omega^2)^{\frac{N-3}{2}}, \quad \omega \in [-1, 1], \tag{5}$$

the even-order moments are

$$\mathbb{E}(\omega^{2m}) = \prod_{j=1}^{m} \frac{2j-1}{N+2j-2}, \quad m = 1, 2, 3, ... \tag{6}$$

and the odd-order moments are zero. As Fig. 1 makes evident,

$$\frac{\omega+1}{2} \sim \text{Beta}\left((N-1)/2, (N-1)/2\right) \tag{7}$$

and the prior places a penalty on correlations which increases with N.

The above properties make it clear that we have presented an alternate way of constructing a prior distribution for correlation matrices having the same marginal distributions for the correlations as the prior for correlation matrices discussed in [9] and the relevant references given there. However, our prior construction naturally leads to a wide variety of flexible generalizations (by choosing different distributions on the unit sphere) and allows for a different MCMC approach to inference based on Geodesic Monte Carlo [6].

2.2 Completing the prior specification

We would like to choose a prior for $\boldsymbol{B}$ favoring a simple, nearly sparse structure. A variety of global-local shrinkage priors [15, 4] have been proposed which satisfy this requirement and have desirable posterior concentration properties. These glocallocal shrinkage priors can typically be represented as scale mixtures of Gaussians, simplifying computation.

As mentioned in the introduction, identifiability assumptions are important in Bayesian computation because they ensure that estimation based on posterior samples is meaningful. We have already constrained $\boldsymbol{\Omega}$ to be a correlation matrix. The article by Peeters [14] gives three additional conditions on $\boldsymbol{B}$ which, under the usual regularity assumptions, guarantee identifiability of the oblique factor model. Decisions about how to satisfy those conditions should be made on a case by case basis.

We can assign conventional priors for variances to the diagonal elements of $\boldsymbol{\Sigma}$, e.g. inverse gamma or reference priors.

3 Extension to tensor data

When I observations of J variables are collected at K occasions we have a three-way array or tensor denoted by $\underline{\mathbf{Y}}$ of order $I \times J \times K$. Occasions may refer to time or in general to different conditions. Three-way tensor data are characterized by three

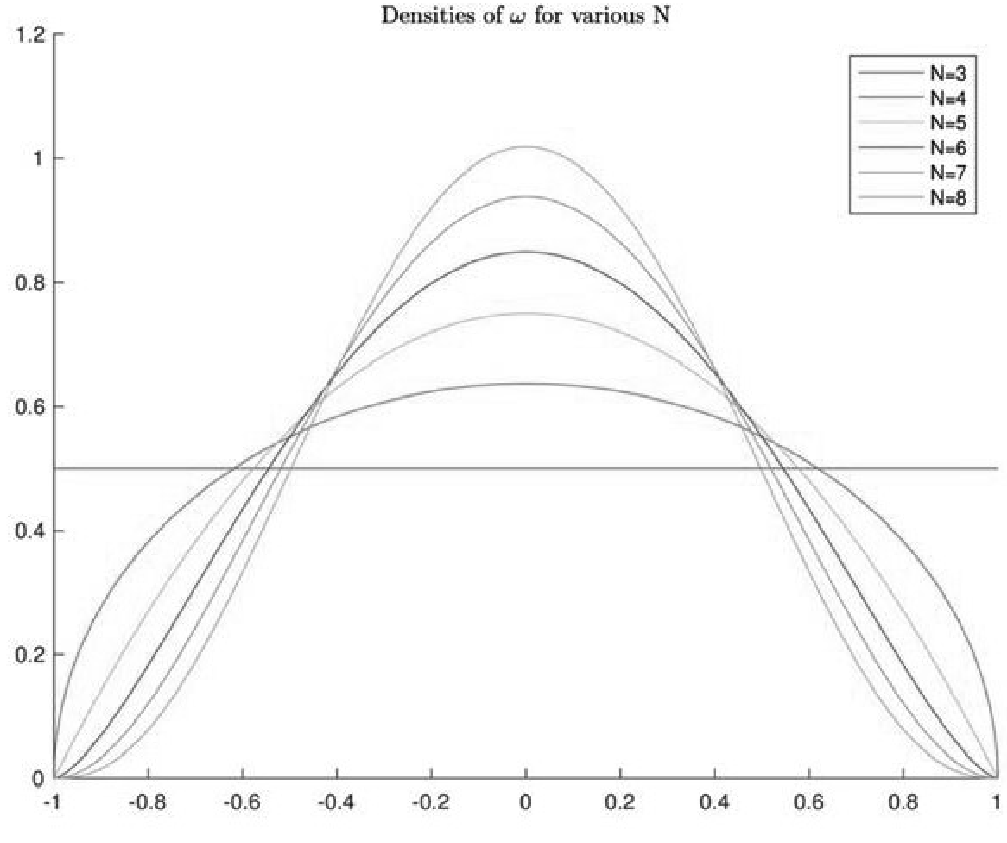

Fig. 1 Density of ω for various values of N. The densities are shifted and scaled $\text{Beta}\left((N-1)/2,(N-1)/2\right)$ densities.

modes, namely observation, variable and occasion modes. We let $\boldsymbol{y}_i$ be the vector corresponding to observation i. In contrast to the standard two-way case, $\boldsymbol{y}_i$ contains the scores of J centered variables at K occasions and thus has length JK.

In principle, the two-way factor model in (4) might still be applied for tensor data. In fact, it would be sufficient to juxtapose next to each other the observation-by-variable matrices collected at every occasion obtaining a matrix with rows given by $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$. Such a matrix, usually denoted by $\boldsymbol{Y}_A$, is the so-called observation mode matricization (or unfolding) of the tensor $\underline{\mathbf{Y}}$. However, in practice, the decomposition of $\boldsymbol{Y}_A$ through the factor model in (4) is inappropriate because the interactions among the modes cannot be modelled.

A more sensible strategy is the three-mode factor analysis model known as Candecomp [7] or Parafac [12] which we will refer to as Candecomp/Parafac or, more briefly, CP. The CP model can be expressed as

$$\boldsymbol{Y}_A = \boldsymbol{F}(\boldsymbol{C} \odot \boldsymbol{B})^T + \boldsymbol{E}_A \tag{8}$$

where $\boldsymbol{B}$ and $\boldsymbol{C}$ are the factor loadings matrices for the variables (of order $J \times S$) and for the occasions (of order $K \times S$), respectively. They capture the influences of the variables and occasions on the S latent factors. The symbol $\odot$ denotes the Khatri-Rao product, the Kronecker product between pairs of columns ($\boldsymbol{C} \odot \boldsymbol{B} = [\boldsymbol{c}_1 \otimes \boldsymbol{b}_1 | \cdots | \boldsymbol{c}_S \otimes \boldsymbol{b}_S]$, with $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_S)$ and $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_S)$). $\boldsymbol{E} = (\boldsymbol{e}_1, ..., \boldsymbol{e}_I)^T$ is the $I \times JK$ matricization of the tensor of errors $\underline{\mathbf{E}}$. In contrast with the two-way case, under mild conditions (see, e.g., [13]) the solution of the CP model is identified up to trivial scaling and simultaneous permutation of the columns of $\boldsymbol{F}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.

The CP model was originally proposed as an exploratory tool without probabilistic assumptions. The probabilistic version was developed in [5] and [2, 3]. Actually, such probabilistic counterparts were proposed for the so-called Tucker3 model [20], which we will refer to as the T3 model. The T3 model represents an alternative three-mode generalization of the two-way factor analysis model. It can be formulated as

$$\boldsymbol{Y}_A = \boldsymbol{F}\boldsymbol{G}_A(\boldsymbol{C} \otimes \boldsymbol{B})^T + \boldsymbol{E}_A. \tag{9}$$

In the T3 model, each mode has its own factors and different numbers of latent factors for each mode can be assumed. Hence, $\boldsymbol{F}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are matrices of order $I \times P$, $J \times Q$, and $K \times R$, respectively, with P, Q and R denoting the numbers of factors for each mode. The triple interactions among the factors of the three modes are captured by the $P \times Q \times R$ core tensor $\underline{\mathbf{G}}$, the generic element of which, g_{pqr}, expresses the strength of the interaction among factor p for the observation mode, factor q for the variable mode, and factor r for the occasion mode. Note that in (9) $\boldsymbol{G}_A$ denotes the observation mode matricization of $\underline{\mathbf{G}}$.

The CP and T3 models are closely related. If $P = Q = R = S$ and $\underline{\mathbf{G}}$ has a superidentity structure ($g_{pqr} = 1$ when $p = q = r$ and 0 otherwise), then it is easy to see that formulas (8) and (9) coincide. Therefore, the CP model can be seen as a constrained version of the T3 model where the same number of latent factors is assumed for all the modes and each factor of a certain mode interacts with exactly one factor of the other modes. This produces some relevant distinctions between the two models. The T3 model is more general than the CP model, but the solution is not identified. Equally well-fitting solutions can be found by rotating the factor matrices and compensating for such rotations in the core. On the other hand, the CP model has a more parsimonious structure and the solution, as mentioned, is identified. For this reason, we focus our attention on the CP model.

If the latent factors of $\boldsymbol{F}$ are correlated, the covariance structure of the data induced by the CP model takes the form (see also [17])

$$(\boldsymbol{C} \odot \boldsymbol{B})\boldsymbol{\Omega}(\boldsymbol{C} \odot \boldsymbol{B})^T + \boldsymbol{\Sigma}. \tag{10}$$

Hence, under the same assumptions adopted in the two-way case, the model for the ith observation in the tensor case is

$$\boldsymbol{y}_i|\boldsymbol{B},\boldsymbol{C},\boldsymbol{\Omega},\boldsymbol{\Sigma} \sim N(\boldsymbol{0},(\boldsymbol{C} \odot \boldsymbol{B})\boldsymbol{\Omega}(\boldsymbol{C} \odot \boldsymbol{B})^T + \boldsymbol{\Sigma}). \tag{11}$$

A Bayesian CP model can be developed as a natural generalization of the two-way model presented previously. The prior of Section 2.1 can again be used for the factor correlation matrix $\boldsymbol{\Omega}$. For the elements of $\boldsymbol{B}$ and $\boldsymbol{C}$, we can again use global-local shrinkage priors which favor a nearly sparse structure. The uniqueness of the CP solution up to scaling and simultaneous permutation of the columns of $\boldsymbol{B}$ and $\boldsymbol{C}$ implies that we only need to worry about column switching in the posterior samples, since the prior distributions for $\boldsymbol{B}$ and $\boldsymbol{C}$ fix their respective scales. Two solutions are a relabeling scheme in the style of [18] or simply fixing particular elements of $\boldsymbol{B}$ or $\boldsymbol{C}$ to be zero so that the columns can no longer be permuted. The Bayesian CP model enjoys the same advantages as the two-way oblique factor model. In particular, we hope to demonstrate that allowing (but penalizing) latent factor correlation and applying a shrinkage prior on the factor loadings leads to a general yet interpretable CP model.

4 Inference

For inference, we use geodesic Monte Carlo [6]. Geodesic Monte Carlo extends Hamiltonian Monte Carlo [16] to certain special manifolds which can be embedded in Euclidean space, such as the simplex, the unit sphere, or the Stiefel manifold. Like Hamiltonian Monte Carlo, geodesic Monte Carlo can generate distant Metropolis-Hastings proposals with a high probability of acceptance, ideally leading to a rapidly-mixing Markov chain with low autocorrelation. As described in [6], sometimes parallel tempering [10] is required to move between isolated modes of the posterior distribution.

References

1. Barnard, J., McCulloch, R., Meng, X.-L.: Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Stat. Sin. **10**, 1281–1311 (2000)
2. Bentler, P.M., Lee, S.Y.: Statistical aspects of a three-mode factor analysis model, Psychometrika **43**, 343–352 (1978)
3. Bentler, P.M., Lee, S.Y.: A statistical development of three-mode factor analysis, Brit. J. Math. Stat. Psy. **32**, 87–104 (1979)
4. Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B.: Dirichlet–Laplace priors for optimal shrinkage, J. Am. Stat. Assoc. **110**, 1479–1490 (2015)
5. Bloxom, B.: A note on invariance in three-mode factor analysis, Psychometrika **33**, 347–350 (1968)
6. Byrne, S., Girolami, M.: Geodesic Monte Carlo on embedded manifolds, Scand. J. Stat. **40**, 825–845 (2013)
7. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition, Psychometrika **35**, 283–319 (1970)
8. Cho, E.: Inner product of random vectors on S^n, J. Pure Appl. Math.: Adv. Appl. **9**, 63–68 (2013)
9. Conti, G., Frühwirth-Schnatter, S., Heckman, J.J., Piatek, R.: Bayesian exploratory factor analysis, J. Econom. **183**, 31–57 (2014)
10. Geyer, C.J.: Markov chain Monte Carlo maximum likelihood. In: Computing Science and Statistics, Proc. 23rd Symp. Interface, pp. 156163. Interface Foundation of North America (1991)
11. Harman, H.H.: Modern Factor Analysis. University of Chicago Press, Chicago (1967)
12. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis", UCLA Work. Pap. Phon. **16**, 1–84 (1970)
13. Kruskal, J.B.: Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra Appl. **18**, 95–138 (1977)
14. Peeters, C.F.W.: Rotational uniqueness conditions under oblique factor correlation metric, Psychometrika **77**, 288–292 (2012)
15. Polson, N.G., Scott, J.G., Clarke, B., Severinski, C.: Shrink globally, act locally: sparse bayesian regularization and prediction. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.) Bayesian Statistics 9, Oxford University Press (2012)
16. Neal, R.M.: MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G., Meng, X.-L. (eds.) Handbook of Markov Chain Monte Carlo, pp. 113–162, CRC Press (2011)
17. Stegeman, A., Lam, T.T.T.: Three-mode factor analysis by means of Candecomp/Parafac, Psychometrika **79**, 426–443 (2014)
18. Stephens, M.: Dealing with label switching in mixture models, J. R. Stat. Soc. Series B: Stat. Methodol. **62**, 795–809 (2000)
19. Thurstone, L.L.: Multiple-factor Analysis: A Development and Expansion of The Vectors of Mind. University of Chicago Press, Chicago (1947)
20. Tucker, L.R: Some mathematical notes on three-mode factor analysis, Psychometrika **31**, 279–311 (1966)