Bayesian Analysis (2011) **6**, Number 2, pp. 197–202

Comment on Article by Hoff

Genevera I. Allen*

1 Introduction

This paper introduces and develops the array normal distribution by extending the matrix-variate normal to the tensor array setting using the Tucker product. Methods for maximum likelihood and Bayesian estimation of separable covariances are given. These contributions are noteworthy as statisticians are encountering increasing numbers of multi-dimensional data sets and methods are needed to model and analyze this array data. Tensor data is especially common in areas of bio-medical imaging, such as neuroimaging and microscopy. With functional magnetic resonance imaging, for example, three-dimensional images of the brain are taken every two to three seconds for many subjects. Often, the dimension of the location variables (voxels) measures in the hundred thousands and the time points measure in the thousands, forming an ultra high-dimensional array. Methods for understanding and modeling these large tensors are certainly needed, and the introduction of the array normal is an important first step in this process.

2 Separable Means

In some cases, having separable means as well as covariances may be useful for tensor data. This may be especially true when no sets of variables along each of the dimensions can be considered independent instances or repeated measures. Then, summing over repeated measures to estimate a mean matrix or mean array is infeasible. Allen and Tibshirani (2010b) modeled separable means for a single instance of a matrix-variate normal data matrix, giving the mean-restricted matrix normal distribution. A similar extension can be employed for the array normal by modeling a separate mean vector for each dimension.

Let $\mathbf{Y}$ be the observed array data, $\mathbf{Y} \in \Re^{m_1 \times \ldots \times m_K}$, and let $\mathbf{M}$ be the mean matrix $\mathbf{M} \in \Re^{m_1 \times \ldots \times m_K}$. Decompose $\mathbf{M} = \sum_{k=1}^{K} \mathbf{M}_k$ where $\mathbf{M}_k = \mathbf{1}_{m_1} \circ \ldots \circ \mathbf{1}_{m_{k-1}} \circ \mu_k \circ \mathbf{1}_{m_{k+1}} \circ \ldots \circ \mathbf{1}_{m_K}$, with $\mu_k \in \Re^{m_k}$, the mean vector of the k^{th} dimension of $\mathbf{Y}$. One can define the mean-restricted array normal as a simple extension of the array normal with the general mean matrix replaced by the structured mean, $\mathbf{M} = \sum_{k=1}^{K} \mathbf{M}_k$. The separable factor means provide a nice analog to the separable covariances of the array normal.

These separable mean parameters can be estimated in a step-wise procedure. (Note that as in the article by Hoff and in other work on the matrix-variate normal (Du-

*Department of Pediatrics-Neurology, Baylor College of Medicine and Department of Statistics, Rice University, Houston, TX, mailto:gallen@rice.edu

 DOI:10.1214/11-BA606A

tilleul 1999; Allen and Tibshirani 2010b), the means are estimated assuming identity covariances. Thus, mean parameters and covariance parameters are estimated in two stages.) Let $m_{(-k)} = \prod_{k' \neq k} m_{k'}$, and recall that $\mathbf{X}_{(k)}$ denotes the matricized array that is of dimension $m_k \times m_{(-k)}$. Then, beginning with $\mathbf{X}^{(1)} = \mathbf{Y}$, the following procedure repeated for each dimension, k, yields the mean MLEs: (i) Let $\hat{\mu}_k$ be the row mean of the matrix $\mathbf{X}^{(k)}_{(k)}$, (ii) define $\widehat{\mathbf{M}}_k = \mathbf{1}_{m_1} \circ \ldots \circ \mathbf{1}_{m_{k-1}} \circ \hat{\mu}_k \circ \mathbf{1}_{m_{k+1}} \circ \ldots \circ \mathbf{1}_{m_K}$, and (iii) set $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \widehat{\mathbf{M}}_k$. The overall mean matrix is $\widehat{\mathbf{M}} = \sum_{k=1}^{K} \widehat{\mathbf{M}}_k$ and the centered array is $\mathbf{X}^{(K+1)}$. Note that the individual means, $\hat{\mu}_k$, are only unique up to an additive constant, and hence the order in which they are estimated is unimportant. The overall mean matrix, $\widehat{\mathbf{M}}$, is unique, however, and is the maximum likelihood estimate. This can be seen by taking partial derivatives of the log-likelihood of the mean-restricted array normal with respect to the separable means, μ_k.

Modeling separable means for the array normal can be thought of as fitting a multi-factor ANOVA model in which there is only one replicate for each factor combination. This relationship may then be exploited to conduct inference on the presence of these factor level means. Further connections between this mean-restricted array normal with and without non-identity separable covariance structures and multi-factor ANOVA models should be investigated.

3 Regularizing Separable Tensor Concentration Matrices

The author presents methods for maximum likelihood and Bayesian estimation of the separable covariances of the array normal. Estimation of these covariances through direct penalized maximum likelihood estimation by placing penalties on the inverse covariance or concentration matrices may also be of interest. Encouraging sparsity in the concentration matrices for example, is related to covariance selection (Dempster 1972) and estimating Gaussian graphical models (Meinshausen and Buhlmann 2006). Penalized maximum likelihood estimation of concentration matrices has been well studied for the multivariate normal (Yuan and Lin 2007; Friedman et al. 2007; Rothman et al. 2008) and recently for the matrix-variate normal (Allen and Tibshirani 2010b). Results from the latter can be easily extended to the framework of the array normal distribution introduced in this paper.

Assume that $\mathbf{X}$ denotes the centered array, $\mathbf{X} = \mathbf{Y} - \widehat{\mathbf{M}}$, and consider the following penalized array normal log-likelihood, denoted by ℓ_p:

$$\ell_p(\mathbf{X} \mid \mathbf{\Sigma}^{-1}{}_1, \ldots \mathbf{\Sigma}^{-1}{}_K) \propto \sum_{k=1}^{K} \frac{m}{2m_k} \log| \mathbf{\Sigma}^{-1}{}_k | - \frac{1}{2} || \mathbf{X} \times \mathbf{\Sigma}^{-1/2} ||_F^2 - \sum_{k=1}^{K} \lambda_k P_k(\mathbf{\Sigma}^{-1}{}_k). \quad (1)$$

Here, λ_k are penalty parameters and $P_k()$ are matrix-convex penalties, $P_k() : \Re^{m_k \times m_k} \rightarrow \Re$, that is, $P_k()$ are matrix norms or convex functions of matrix norms. Some examples discussed in Allen and Tibshirani (2010b) are $P_k(\mathbf{\Sigma}^{-1}{}_k) = \sum_i \sum_{i'} | \mathbf{\Sigma}^{-1}{}_k(i,i') |$ or $P_k(\mathbf{\Sigma}^{-1}{}_k) = \sum_i \sum_{i'} (\mathbf{\Sigma}^{-1}{}_k(i,i'))^2 = \mathrm{tr}(\mathbf{\Sigma}_k^{-2}) = || \mathbf{\Sigma}^{-1}{}_k ||_F^2$, corresponding to L_1 and L_2 norm penalties.

If we define $\mathbf{\Sigma^{-1}}_{(-k)} = \mathbf{\Sigma^{-1}}_1 \otimes \ldots \otimes \mathbf{\Sigma^{-1}}_{k-1} \otimes \mathbf{\Sigma^{-1}}_{k+1} \otimes \ldots \otimes \mathbf{\Sigma^{-1}}_K$, then notice that the middle term of the penalized log-likelihood, ℓ_p, can be written in terms of $\mathbf{\Sigma^{-1}}_k$:

$$|| \mathbf{X} \times \mathbf{\Sigma}^{-1/2} ||_F^2 = \text{tr}\left(\mathbf{\Sigma^{-1}}_k \mathbf{X}_{(k)} \mathbf{\Sigma^{-1}}_{(-k)} \mathbf{X}_{(k)}^T \right).$$

From this, it is easy to see that ℓ_p is concave in each $\mathbf{\Sigma^{-1}}_k$ with the other concentration matrices fixed. Thus, an iterative block-wise maximization strategy may be employed that increases the penalized likelihood at each iteration and converges (Tseng 2001). This is similar to the approach outlined by Hoff for un-penalized covariance estimation.

At each step of this iterative estimation algorithm, one must solve the following subgradient equation:

$$\frac{\partial \ell_p}{\partial \mathbf{\Sigma^{-1}}_k} = \mathbf{\Sigma}_k - \frac{m_k}{m} \mathbf{X}_{(k)} \mathbf{\Sigma^{-1}}_{(-k)} \mathbf{X}_{(k)}^T - \nabla P_k(\mathbf{\Sigma^{-1}}_k) \frac{2\lambda_k m_k}{m} = 0. \quad (2)$$

Here, $\nabla P_k(\mathbf{\Sigma^{-1}}_k)$ is the subgradient of $P_k()$ with respect to $\mathbf{\Sigma^{-1}}_k$. For the L_1 penalty, $\nabla P_k(\mathbf{\Sigma^{-1}}_k) = \mathbf{\Gamma}(\mathbf{\Sigma^{-1}}_k)$ where $\mathbf{\Gamma}(i,i') = \text{sign}(\mathbf{\Sigma^{-1}}_k(i,i'))$ if $\mathbf{\Sigma^{-1}}_k(i,i') \neq 0$ or $\in [-1,1]$ otherwise. This subgradient equation can be solved by applying the graphical lasso algorithm (Friedman et al. 2007) to the second term with the penalty given by the coefficient of the third term. For the L_2 penalty, $\nabla P_k(\mathbf{\Sigma^{-1}}_k) = 2\,\mathbf{\Sigma^{-1}}_k$, yielding an eigenvalue problem. The solution for $\mathbf{\Sigma^{-1}}_k$ has the same eigenvalues as that of the second term of the gradient equation while the eigenvalues are regularized versions of the eigenvalues of the second term. For the matrix-variate normal with Frobenius norm penalties, Allen and Tibshirani (2010b) found an analytical solution for the concentration matrices that is a function of the singular value decomposition of the data matrix and forms the global solution to the penalized log-likelihood. The proof of the global nature of this solution relies on the uniqueness of the singular value decomposition. Given this, I conjecture that one can obtain a solution for the concentration matrices of the Frobenius norm penalized array normal that is a function of the Tucker decomposition of the array data (Tucker 1964, 1966). As the Tucker decomposition is not unique, however, this solution is unlikely to yield the global maximum of the penalized log-likelihood.

Regularizing the separable concentration matrices of the array normal presents many advantages. First, the L_1 penalty estimates network structures for variables along each dimension. Putting these together, one can estimate a Kronecker graph structure to represent and understand the relationships between elements in the array. As the L_2 penalized estimates have the same eigenvectors as the array maximum likelihood estimates described in the article, these estimates can be used to estimate non-singular covariances when the tensor is rank deficient. While I have presented a framework for regularizing separable tensor concentration matrices with convex penalties, this can also be cast as a Bayesian estimation problem with Wishart priors.

4 Conditional Distributions

Hoff provides results on conditional distributions for the array normal in Proposition 3.2 similar to the conditional distribution results for the matrix-variate normal given

in Allen and Tibshirani (2010b). These results are noteworthy and have numerous implications beyond the brief discussion provided in the article. The main advantage gained by calculating conditional distributions in this manner is computational, as they can also be calculated by vectorizing the array and applying the multivariate normal conditional distribution formulas. The separable covariances allow one to split these calculations along each dimension of the array, thus substantially reducing the order of computations. By using properties of the Schur complement, the computations can be reduced further, and the complexity depends on the minimum of the number of observed or number of missing elements per row of each $\mathbf{X}_{(k)}$ (Allen and Tibshirani 2010a).

These computationally efficient algorithms are needed with high-dimensional tensors for numerous items including missing data imputation and cross-validation. For Bayesian missing data imputation, one can use Gibbs samplers as outlined by Hoff and also by Allen and Tibshirani (2010a). For imputing the missing values to their conditional means, an alternating conditional expectations approach can be employed (Allen and Tibshirani 2010b). These conditional distributions are also important for cross-validation and can be used to assess the fit of certain covariance structures or to select penalty parameters. To perform cross-validation, a small fraction of elements can be deleted from the array and imputed; then, the prediction error of these deleted elements can be calculated for several folds. Following the reduction in computations given by using Schur complements, one can remove elements in the array according to a Latin hypercube sampling scheme that keeps the computational complexity fixed despite the dimensionality. This may be especially useful for assessing the fit of high-dimensional array normal models.

5 Conclusion

As multi-dimensional data with possible correlations among the variables of each dimension is becoming more prevalent, the introduction of the array normal distribution is an important contribution. Further development of the theoretical properties, alternative parametrization and estimation schemes, and strategies for efficient computation, manipulation and visualization of the array normal are needed. Thus, Hoff's work has paved the way for many open research questions related to modeling tensor data.

References

Allen, G. and Tibshirani, R. (2010a). "Transposable regularized covariance models with an applicaiton to missing data imputation. [Supplemental Materials]." *Annals of Applied Statistics*, 4(2): 764–790.
URL http://lib.stat.cmu.edu/aoas/314/supplement.PDF 200

Allen, G. I. and Tibshirani, R. (2010b). "Transposable regularized covariance models with an applicaiton to missing data imputation." *Annals of Applied Statistics*, 4(2): 764–790. 197, 198, 199, 200

Dempster, A. P. (1972). "Covariance Selection." *Biometrics*, 28(1): 157–175. 198

Dutilleul, P. (1999). "The MLE algorithm for the matrix normal distribution." *Journal of Statistical Computation and Simulation*, 64: 105–123. 197

Friedman, J., Hastie, T., and Tibshirani, R. (2007). "Sparse inverse covariance estimation with the lasso." *Biostatistics*, 9(3): 432–441. 198, 199

Meinshausen, N. and Buhlmann, P. (2006). "High-dimensional graphs and variable selection with the Lasso." *Annals of Statistics*, 34(3): 1436–1462. 198

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). "Sparse permutation invariant covariance estimation." *Electronic Jorunal of Statistics*, 2: 494–515. 198

Tseng, P. (2001). "Convergence of a block coordinate descent method for nondifferentiable minimization." *Journal of Optimization Theory and Applications*, 109(3): 475–494. 199

Tucker, L. R. (1964). "The extension of factor analysis to three-dimensional matrices." In Gulliksen, H. and Frederiksen, N. (eds.), *Contributions to Mathematical Psychology.*, 110–127. New York: Holt, Rinehart and Winston. 199

— (1966). "Some mathematical notes on three-mode factor analysis." *Psychometrika*, 31(3): 279–311. 199

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model." *Biometrika*, 94(1): 19–35. 198