UNSUPERVISED BAYESIAN ANALYSIS OF GENE EXPRESSION PATTERNS

Cécile Bazot [1], *Nicolas Dobigeon* [1], *Jean-Yves Tourneret* [1] *and Alfred O. Hero III* [2]

[1] University of Toulouse, IRIT/INP-ENSEEIHT, Toulouse, France
[2] University of Michigan, EECS Dept., Ann Arbor, USA
{cecile.bazot, nicolas.dobigeon, jean-yves.tourneret}@enseeiht.fr, hero@umich.edu

ABSTRACT

In this paper we introduce a new method for analyzing expression patterns from high throughput and complex data such as gene expression microarrays. These microarrays are collected under different conditions such as time, phenotype and treatment. The proposed method uses a Bayesian matrix decomposition, called Bayesian linear unmixing (BLU), to extract a set of characteristic gene signatures, or *factors*, and a set of coefficients, *factor scores*, that specify the relative contribution of each signature to a specific sample. BLU is related to Bayesian factor analysis but differs in an important respect: BLU constrains the factor loadings to be non-negative and the factor scores to be probability distributions over the factors. Thus BLU reduces the multiplexing of genes into different factors and can enhance interpretability of the factor loadings and factor scores. The unsupervised version of BLU presented in this paper also provides estimates of the number of factors. We illustrate the application of BLU to bioinformatics by analyzing gene expression microarray datasets.

Index Terms— Factor analysis, Bayesian inference, MCMC methods, gene expression data.

1. INTRODUCTION

Factor analysis methods such as principal component analysis (PCA) have been widely studied for discovering of patterns of differential expression in time course and/or multiple treatment biological experiments using gene microarray samples. These methods find a decomposition of the observation matrix whose rows are indexed by gene index and whose columns are indexed by sample index. This decomposition expresses each sample as a particular linear combination of fundamental gene expression signatures, called *factors* with appropriate proportions, called *factor scores*. The number of factors in the decomposition is usually much less than the number of samples. Traditional factor analysis methods such as PCA require the experimenter to specify the desired number of factors to be estimated. However, some recent Bayesian factor analysis methods are totally unsupervised in the sense that they also estimate the number of factors directly from the data [1, 2].

In this paper we propose a new Bayesian factor analysis method called unsupervised Bayesian linear unmixing (BLU), that incorporates a non-negativity constraint on the factors and the factor scores in addition to requiring the factor scores to sum to one. Thus each factor score corresponds to the proportion of a particular factor to a given sample. The advantage of this representation for gene expression analysis is twofold: (i) the factor scores correspond to the strengths of each gene contributing to that factor; (ii) for each gene chip the factor scores give the relative abundance of each factor present in the chip. Thus, for example, a gene having a large loading level (close to one) for a particular factor should have a small loading (close to zero) for all other factors. In this way, as opposed to other factor analysis methods, there is less multiplexing making it easier to associate specific genes to specific factors and vice versa.

Unsupervised Bayesian linear unmixing, traditionally used for hyperspectral image analysis, is one of many factor analysis methods. These methods include non-negative matrix factorization (NMF) [3], independent component analysis (ICA) [4], bi-clustering [5], PCA, penalized matrix decomposition (PMD) [2], and Bayesian factor regression modeling (BFRM) [1]. Contrary to BLU, the PCA, ICA, BFRM, bi-clustering and PMD methods do not account for non-negativity of the factor loadings and factor scores. On the other hand, NMF does not account for sum-to-one constraints on the columns of the factor score matrix. Unlike PCA or ICA, BLU does not impose orthogonality or independence on the components. These relaxed assumptions might better represent what is known about the preponderance of overlap and dependency in biological pathways. Finally, BLU naturally accommodates Bayesian estimation of the number of factors.

In this paper we provide comparative studies that establish the advantages of BLU over PCA, NMF and BFRM for time-varying gene expression analysis. We illustrate the application of unsupervised BLU to the analysis of a gene expression dataset. This set is the beverage data of Baty *et al* [6] in which 4 different beverages were administered on different days to 6 human individuals. Gene expression time courses were measured over a 12-hour period for each beverage and each individual.

The outline of the paper is as follows. Section 2 pro-

vides the observation model used in this paper. Section 3 presents the unsupervised BLU algorithm in the context of gene expression analysis. In Section 4.1 we establish performance advantages of unsupervised BLU for synthetic data with known ground truth. Section 4.2 is devoted to the application to the beverage gene expression dataset. Section 5 concludes the paper with comments about future works.

2. MATHEMATICAL MODEL

Consider a gene microarray represented by a vector $\mathbf{y}$ of G gene expression levels. For example, in an Affymetrix HU133 gene chip, the number G of genes indexing the elements of $\mathbf{y}$ may range from ten to twenty thousand depending on the type of chip description file (CDF) processing used to translate the oligonucleotide fragments to gene labels. The elements of $\mathbf{y}$ have units of hybridization abundance levels with non-negative values. In this context of gene expression data, the starting point for Bayesian linear unmixing is the linear mixing model (LMM)

$$\mathbf{y} = \sum_{r=1}^{R} \mathbf{m}_r a_r + \mathbf{n}, \tag{1}$$

where R is the number of distinct gene signatures that can be present in the chip, $\mathbf{m}_r = [m_{r,1}, \ldots, m_{r,G}]^T$ is a gene signature vector, $m_{r,g} \geq 0$ is the strength of the g-th gene in the r-th signature, and a_r is the relative contribution of the r-th signature vector to the sample $\mathbf{y}$, where $a_r \in [0, 1]$ and $\sum_{r=1}^{R} a_r = 1$. Here $\mathbf{n}$ denotes the residual error of the LMM. For a matrix of N data samples $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ the LMM can be written in matrix form

$$\mathbf{Y} = \mathbf{M}\mathbf{A} + \mathbf{N}, \tag{2}$$

with $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_R]$, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ and $\mathbf{N} = [\mathbf{n}_1, \ldots, \mathbf{n}_N]$. The matrices $\mathbf{M}$, $\mathbf{A}$ satisfy positivity and sum-to-one constraints defined by

$$m_{r,g} \geq 0, \quad a_{r,i} \geq 0, \quad [1, \ldots, 1]\mathbf{A} = [1, \ldots, 1], \tag{3}$$

where $m_{r,g}$ denotes the (r, g)-th element of matrix $\mathbf{M}$. The reader will note that except the constraints (3) the LMM (2) is identical to standard factor analysis model [7] for which the columns of $\mathbf{M}$ are called *factors*, the elements of each of these column vectors are called *factor loadings*, and the columns of $\mathbf{A}$ are called *factor scores*. The constraints (3) arise naturally when dealing with positive data for which one is seeking the relative contribution of positive factors that have the same numerical characteristics as the data, i.e., the signature $\mathbf{m}_r$ is itself interpretable as a vector of hybridization abundances.

The objective of linear unmixing is to simultaneously estimate the factor matrix $\mathbf{M}$ and the factor score matrix $\mathbf{A}$ from the available N data samples. The representation (2) is rank deficient since $\mathbf{A}$ has rank $N - 1$. This presents algorithmic challenges for solving the unmixing problem. Several algorithms have been proposed in the context of hyperspectral imaging to solve similar problems [8, 9]. Most of these algorithms perform unmixing in a two step procedure where $\mathbf{M}$ is estimated first using an *endmember extraction algorithm* (EEA) followed by a constrained linear least squares step to solve for $\mathbf{A}$. A common, but restrictive, assumption in these algorithms is that there exist samples in the dataset which are "pure" in the sense that they contain a single factor, say $\mathbf{m}_r$. Recently, this assumption has been relaxed by applying a single step hierarchical Bayesian approach, called Bayesian linear unmixing (BLU). The resulting algorithm required the number R of factors to be specified (see [10] for details). Here we extend BLU to a fully unsupervised algorithm that generates samples distributed according to the joint posterior distribution of R and the other model parameters, from which a Bayesian estimator of R can be derived. This unsupervised algorithm is then applied to extract expression patterns in gene microarray data.

3. UNSUPERVISED BAYESIAN LINEAR UNMIXING

The BLU algorithm studied in [10] generates an estimate of the posterior distribution of $\mathbf{M}$ and $\mathbf{A}$ given the number R of factors for appropriate prior distributions assigned to the mixing parameters in (1). Moreover, the residual errors in (1) are assumed to be independent identically distributed (i.i.d.) zero-mean Gaussian distributed residual errors ($\mathbf{n}_i \sim \mathcal{N}\left(\mathbf{0}_G, \sigma^2\mathbf{I}_G\right)$ for $i = 1, \ldots, N$, where $\mathbf{I}_G$ denotes the identity matrix of dimension $G \times G$).

Because of the constraints in (3), the data samples $\mathbf{y}_i$ ($i = 1, \ldots, N$) live in a lower-dimensional subset of $\mathbb{R}^K$ ($R - 1 \leq K \leq G$) that can be identified by a standard dimension reduction procedure, such as a PCA. Hence, instead of estimating the factor loadings $\mathbf{m}_r$ ($r = 1, \ldots, R$), we propose to estimate their corresponding projections on appropriate axes denoted as

$$\mathbf{t}_r = \mathbf{P}(\mathbf{m}_r - \bar{\mathbf{y}}) \tag{4}$$

where $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i$ is the empirical mean of the data and $\mathbf{P}$ is an appropriate projection matrix. A multivariate Gaussian distribution (MGD) truncated on a subset $\mathcal{T}_r$ is chosen as prior distribution for the projected factors $\mathbf{t}_r$. The subset $\mathcal{T}_r$ is defined so that the non-negativity constraint on $\mathbf{m}_r$ is ensured (see [10]). The mean vectors $\mathbf{e}_r$ of this truncated MGD are provided by an EEA dedicated to hyperspectral imagery and the variances s_r^2 are fixed to a large value. To summarize, the prior for $\mathbf{t}_r$ is

$$\mathbf{t}_r \sim \mathcal{N}_{\mathcal{T}_r}\left(\mathbf{e}_r, s_r^2\mathbf{I}_{R-1}\right) \tag{5}$$

where $\mathcal{N}_{\mathcal{T}_r}\left(\mathbf{e}_r, s_r^2\mathbf{I}_{R-1}\right)$ denotes the truncated MGD with mean $\mathbf{e}_r$ and covariance matrix $s_r^2\mathbf{I}_{R-1}$.

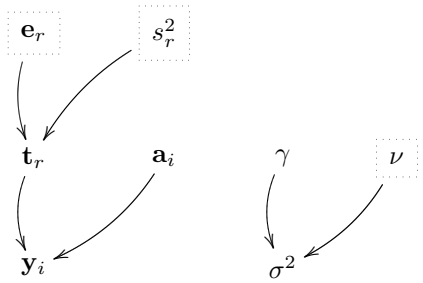

Fig. 1. DAG for the parameter priors and hyperpriors (the fixed parameters appear in dashed boxes).

The sum-to-one constraint for the factor scores $\mathbf{a}_i$, for each observed sample i ($i = 1, \ldots, N$), allows this vector $\mathbf{a}_i$ to be rewritten as

$$\mathbf{a}_i = \begin{pmatrix} \mathbf{a}_{i,1:R-1} \\ a_{i,R} \end{pmatrix} \text{ with } \mathbf{a}_{i,1:R-1} = [a_{i,1}, \ldots, a_{i,R-1}]^T,$$

and $a_{i,R} = 1 - \sum_{r=1}^{R-1} a_{i,r}$. We propose to assign uniform distributions over the simplex $\mathcal{S}$ as priors for the vectors $\mathbf{a}_{i,1:R-1}$, where the simplex $\mathcal{S}$ is defined by $\mathcal{S} = \{\mathbf{a}_{i,1:R-1} \mid \|\mathbf{a}_{i,1:R-1}\|_1 \leq 1 \text{ and } \mathbf{a}_{i,1:R} \succeq \mathbf{0}\}$, where $\|\cdot\|_1$ is the l_1 norm ($\|\mathbf{a}_i\|_1 = \sum_{r=1}^{R} |a_{i,r}|$) and $\mathbf{a}_i \succeq \mathbf{0}$ stands for the set of inequalities $\{a_{i,r} \geq 0\}_{r=1,\ldots,R}$. Finally, an inverse gamma prior has been chosen for the variance of the residual errors σ^2. The resulting hierarchical structure of the proposed BLU model is summarized in the directed acyclic graph (DAG) shown in Fig. 1.

The unsupervised version of BLU generates samples of R in addition to $\mathbf{M}$, $\mathbf{A}$ by assuming a birth/death process for R. This is achieved by an MCMC method that chooses a birth, death or switch move at each iteration (denoted as $^{(t)}$). The *birth* and *death* moves consist of increasing or decreasing by 1 the number R of factors using a reversible jump MCMC method (see [11] for more details), whereas the *switch* move does not change the dimension of R and requires the use of a Metropolis-Hastings acceptance procedure. For example, in the case of a *death* move, a factor and its corresponding factor scores are randomly removed. The factor matrix $\mathbf{M}$, the factor score matrix $\mathbf{A}$ and the noise variance σ^2 are then updated, conditionally upon the number of factors R, using Gibbs moves. After a sufficient number of iterations, the generated samples are used to approximate the maximum a posteriori (MAP) estimator of the number of factors $\widehat{R}$, and conditionally upon $\widehat{R}$ the joint MAP estimator $\left(\widehat{\mathbf{M}}, \widehat{\mathbf{A}}\right)$ of the factor and factor score matrices. Figure 2 summarizes the proposed BLU method.

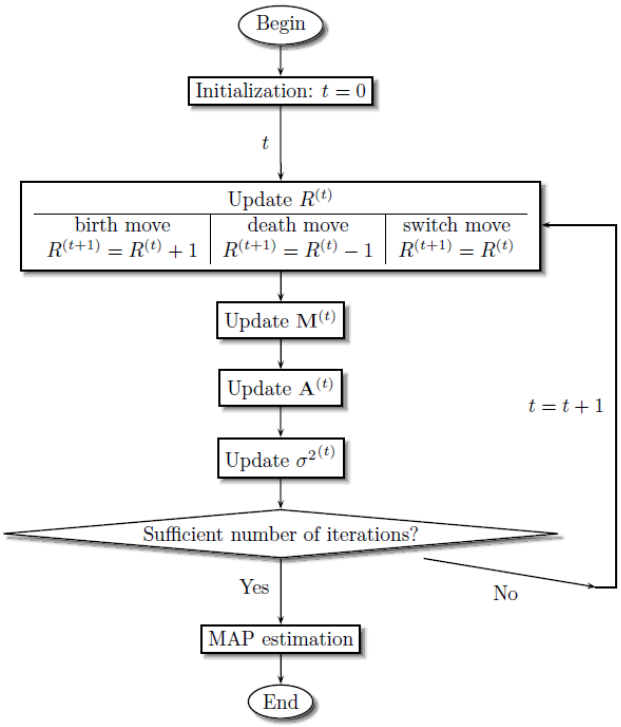

Fig. 2. Flow diagram of the BLU algorithm.

4. SIMULATION RESULTS

4.1. Synthetic data

To illustrate the performance of the proposed Bayesian unmixing algorithm, many simulations have been conducted on synthetic data. The experiments conducted in this paper correspond to the expression value of $G = 512$ genes, for $N = 128$ samples. Each sample is composed of exactly $R = 3$ different factors. These factors are mixed in random proportions (factor scores), and are corrupted by an i.i.d. noise sequence. The signal-to-noise ratio has been fixed to $\mathrm{SNR} = 20$ dB. The hidden mean vectors $\mathbf{e}_r$ ($r = 1, \ldots, R$) are chosen as the PCA projections of the factors, previously identified by the VCA algorithm [8]. The proposed algorithm has been run with $\mathrm{N_{mc}} = 5000$ Monte Carlo iterations with a burn-in period of $\mathrm{N_{bi}} = 500$ iterations.

The first step of the algorithm consists of estimating the number of factors R involved in the mixture, and hence determining the dimensions of the matrices $\mathbf{M}$ and $\mathbf{A}$. The estimated posterior distribution of R depicted in Fig. 3 is clearly in agreement with the actual value of R. This figure also shows that the proposed algorithm moves into spaces with different dimensions (corresponding to $R = 2$ and $R = 4$). The second step of the algorithm consists of estimating the unknown parameters ($\mathbf{M}$, $\mathbf{A}$ and σ^2) conditionally upon $\widehat{R}$. As an example, the posterior distribution of the factor scores obtained for the particular sample ♯30 is depicted in Fig. 4.

These posteriors are in good agreement with the actual values of the factor scores (dashed lines).

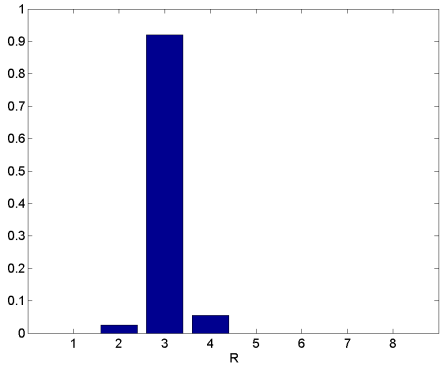

Fig. 3. Estimated posterior of R (synthetic data).

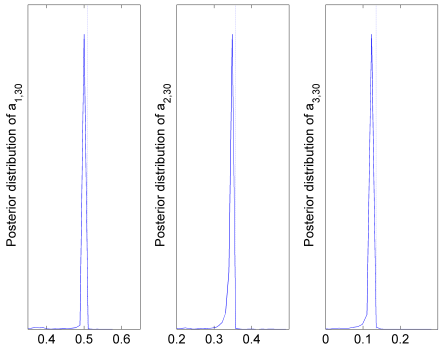

Fig. 4. Estimated posterior distributions for the factor scores $[a_{1,30}, a_{2,30}, a_{3,30}]^T$ conditionally upon $\widehat{R} = 3$.

The performance of the proposed BLU algorithm has been compared with other existing factor decomposition methods including NMF, PCA and BFRM by using different criteria

- the factor mean square errors (MSE)

$$\mathrm{MSE}_r^2 = \|\hat{\mathbf{m}}_r - \mathbf{m}_r\|^2, \; r = 1, \ldots, R$$

- the global MSE of factor scores

$$\mathrm{GMSE}_r^2 = \sum_{i=1}^{N} \left(\hat{a}_{i,r} - a_{i,r}\right)^2, \; r = 1, \ldots, R$$

- the reconstruction error

$$\mathrm{RE} = \sum_{i=1}^{N} \left\| \mathbf{y}_i - \sum_{r=1}^{R} \hat{\mathbf{m}}_r \hat{a}_{i,r} \right\|^2$$

- the computational time.

Simulation results are reported in Table 1. Note that the BFRM method can be run with or without specifying the number of factors. The NMF, PCA, and BFRM method have been run for the actual number of factors to be estimated, i.e. $R = 3$.

Table 1. Comparison measures between the proposed BLU algorithm and NMF, PCA and BFRM methods.

	BLU	NMF $R=3$	PCA $R=3$	BFRM $R=3$	BFRM
MSE_r^2	5.31	5.89	29.29	1209	2808
	7.22	2.75	55.30	1381	3113
	4.91	1.58	66.80	1641	N/A
GMSE_r^2	0.60	1.46	0.52	3616	79.15
	0.62	5.41	0.38	2188	77.74
	0.66	2.73	0.71	4883	N/A
RE	28.80	118.99	52.50	3.68×10^5	2.61×10^4
Time (in s)	647.1	1.5	0.1	28.2	283.0

The results obtained with synthetic data have illustrated the accuracy (and superiority) of the proposed Bayesian approach for the unsupervised unmixing of gene-expression data when compared to other existing factor decomposition methods.

4.2. Beverage data

This section shows some results obtained with a publicly available gene expression dataset from gene expression omnibus (GEO) [6] (through GEO Series accession number GSE3846) called the beverage dataset. This dataset consists of $N = 108$ processed affymetrix chips collected during an experiment where six subjects imbibed one of four different beverages. Peripheral blood microarray analysis was performed at five post-treatment time instants corresponding to 0, 1, 2, 4 and 12 hours. The four beverages were water, grape juice, red wine and alcohol. The experiment was repeated four times on each subject under treatment with a different beverage.

The proposed BLU algorithm was run with $\mathrm{N_{mc}} = 10000$ Monte Carlo iterations, including a burn-in period of $\mathrm{N_{bi}} = 1000$ iterations. Figure 5 shows that the MAP estimate of the number of factors is $\widehat{R} = 3$. These 3 factors are depicted in Fig. 6 where the $G = 22283$ genes have been reordered so that the dominant genes are grouped together in each factor. The 3 sharp peaks in the figure correspond to the gene index that has maximal loading for each of the factors. The factor scores are shown in Fig. 7, where they are displayed as an image whose columns (respectively rows) correspond to the 6 subjects (resp. the 5 time instants under the 4 experiments). Note that factor ♯2 is strongly associated with subjects ♯1 and ♯3. However, factor ♯3 seems to have close values for the different samples. Thus, this factor is not specific to a particular subject or treatment.

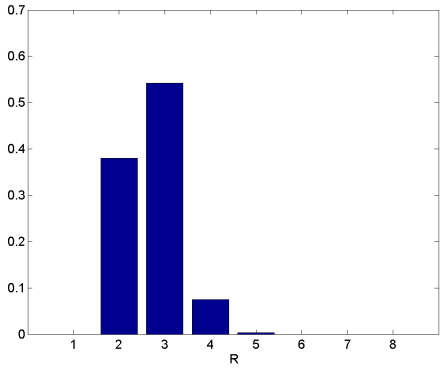

Fig. 5. Estimated posterior of R (beverage data).

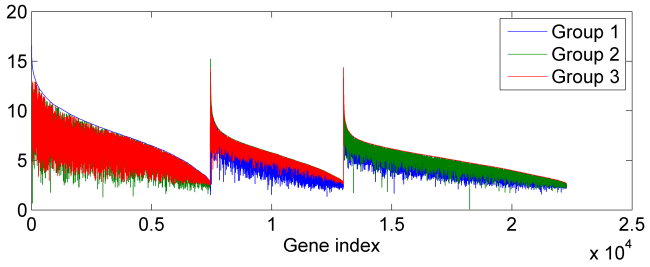

Fig. 6. Estimated factor loadings ranked by decreasing dominance.

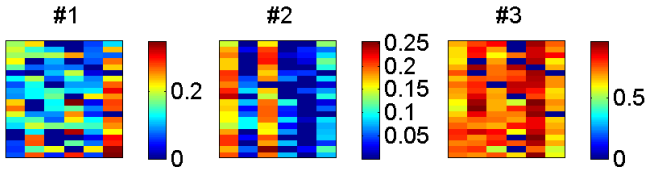

Fig. 7. Estimated factor scores for each of the $R = 3$ factors.

5. CONCLUSION

This paper studied an unsupervised Bayesian unmixing algorithm for gene expression microarrays. An interesting property of the proposed algorithm was to provide positive factor loadings and to ensure positivity as well as sum-to-one constraints for the factor scores. A reversible-jump MCMC algorithm was used to estimate the different unknown model parameters, including the number of factors involved in the mixture. Simulation results performed on synthetic and real data are very encouraging for gene expression analysis. Future works will be to extend the proposed model to models with temporal dependencies, to enforce sparsity on the factor scores [12], and to enforce the factors to be related to treatments instead of subjects.

6. REFERENCES

[1] C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modelling: Applications in gene expression genomics," *J. Amer. Stat. Assoc.*, vol. 103, no. 484, pp. 1438–1456, December 2008.

[2] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proc. 26th Annual Int. Conf. on Machine Learning*. ACM, 2009, pp. 777–784.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. of Neural Info. Process. Syst.*, 2000.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley, 2001.

[5] D. Dueck, Q. D. Morris, and B. J. Frey, "Multi-way clustering of microarray data using probabilistic sparse matrix factorization," *Bioinformatics*, vol. 21, pp. 144–151, 2005.

[6] F. Baty, M. Facompre, J. Wiegand, J. Schwager, and M. Brutsche, "Analysis with respect to instrumental variables for the exploration of microarray data structures," *BMC Bioinformatics*, vol. 7, no. 1, p. 422, 2006.

[7] M. West, "Bayesian factor regression models in the "large p, small n" paradigm," in *Bayesian Statistics*. Oxford University Press, 2003, pp. 723–732.

[8] J. M. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. and Remote Sensing*, vol. 43, no. 4, pp. 898–910, April 2005.

[9] M. E. Winter, "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data," in *Imaging Spectrometry V*, M. R. Descour and S. S. Shen, Eds., vol. 3753, no. 1. SPIE, 1999, pp. 266–275.

[10] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Trans. Signal Processing*, vol. 57, no. 11, pp. 4355–4368, Nov. 2009.

[11] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[12] C. Bazot, N. Dobigeon, J.-Y. Tourneret, and A. O. Hero, "Bernoulli-Gaussian model for gene expression analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, submitted.