A BERNOULLI-GAUSSIAN MODEL FOR GENE FACTOR ANALYSIS

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

This paper investigates a Bayesian model and a Markov chain Monte Carlo (MCMC) algorithm for gene factor analysis. Each sample in the dataset is decomposed as a linear combination of characteristic gene signatures (also referred to as *factors*) following a linear mixing model. To enforce the sparsity of the relative contribution (called *factor score*) of each gene signature to a specific sample, constrained Bernoulli-Gaussian distributions are elected as prior distributions for these factor scores. This distribution allows one to ensure non-negativity and full-additivity constraints for the scores that are interpreted as concentrations. The complexity of the resulting Bayesian estimators is alleviated by using a Gibbs sampler which generates samples distributed according to the posterior distribution of interest. These samples are then used to approximate the standard *maximum a posteriori* (MAP) or *minimum mean square error* (MMSE) estimators. The accuracy of the proposed Bayesian method is illustrated by simulations conducted on synthetic and real data.

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

1. INTRODUCTION AND PROBLEM STATEMENT

Factor analysis methods aim at finding a decomposition of an observation matrix $\mathbf{Y} \in \mathbb{R}^{G\times N}$ whose rows (resp. columns) correspond to genes (resp. samples). Typically, in gene expression analysis, the number N of samples is much less than the number G of genes. Each observed sample vector $\mathbf{y}_i$ ($i = 1, \ldots, N$) is assumed to satisfy a linear mixing model (LMM)

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

where $\mathbf{m}_r = [m_{r,1}, \ldots, m_{r,G}]^T$ denotes the rth gene signature vector, also called *factor*, $a_{i,r}$ is the contribution (or *factor score*) of the rth signature vector in the ith observed sample, R is the number of gene signatures present in the chip and $\mathbf{n}_i$ denotes a residual error. In this paper, the R factors $\{\mathbf{m}_r\}_{r=1,\ldots,R}$ are assumed to belong to a library $\mathbf{M}$ of K gene signature vectors, therefore $\mathbf{M} \in \mathbb{R}^{G\times K}$ with $K > R$. Considering N samples, the model can be rewritten with matrix notations

$$\mathbf{Y} = \mathbf{M}\mathbf{A} + \mathbf{N}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ represents the factor score matrix, $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_K]$ the factor loading matrix and $\mathbf{N} = [\mathbf{n}_1, \ldots, \mathbf{n}_N]$.

The proposed method studies the problem of gene factor analysis in a fully unsupervised framework, i.e, it estimates the factor score proportions and the gene signatures jointly. Note that the number of factors is determined directly from the data. The advantage of the proposed method compared to other factor analysis methods, such as the nonparametric Bayesian factor analysis (NPBFA) [1] or the Bayesian factor regression modeling (BFRM) [2], is that it incorporates non-negativity constraints on the factor components $m_{k,1}, \ldots, m_{k,G}$ and the factor scores $a_{i,1}, \ldots, a_{i,K}$, as well as a full-additivity constraint for the factor scores, i.e.,

$$\begin{aligned} & a_{i,k} \geq 0,\ i = 1, \ldots, N, \quad k = 1, \ldots, K, \\ & \sum_{k=1}^{K} a_{i,k} = 1,\ i = 1, \ldots, N, \\ & m_{k,g} \geq 0,\ k = 1, \ldots, K, \quad g = 1, \ldots, G. \end{aligned} \tag{2}$$

Such constraints are natural for non-negative data, such as gene expression measured as an abundance of molecular binding, and can often lead to more straightforward interpretation of the factor loadings and scores. Note that the constraints (2) were used in [3] for hyperspectral image unmixing. However, the approach adopted in this paper differs from [3] since it enforces a sparsity constraint on the factor scores.

As in other Bayesian factor analysis methods, the residual error vector $\mathbf{n}_i = [n_{i,1}, \ldots, n_{i,G}]^T$ is assumed to be an independent and identically distributed (i.i.d.) zero-mean Gaussian sequence with covariance matrix $\Sigma = \sigma^2 \mathbf{I}_G$

$$\mathbf{n}_i | \sigma^2 \sim \mathcal{N}\left(\mathbf{0}_G, \sigma^2 \mathbf{I}_G\right) \tag{3}$$

where $\mathbf{I}_G$ is the identity matrix of dimension $G \times G$ and $\mathcal{N}(\mathbf{m}, \Sigma)$ denotes the Gaussian distribution with mean vector $\mathbf{m}$ and covariance matrix Σ. The problem addressed in this paper consists of estimating the factor loadings $\mathbf{m}_1, \ldots, \mathbf{m}_K$ and the factor scores $\mathbf{a}_1, \ldots, \mathbf{a}_N$ jointly from the observed sample vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$.

This paper is organized as follows. Section 2 presents the Bayesian factor analysis (BFA) model. Section 3 studies a Gibbs sampler used for generating samples distributed according to the posterior distribution associated to the BFA model. We illustrate the proposed factor analysis method on both synthetic and real data, presented in Section 4 and Section 5 respectively. Conclusions are given in Section 6.

2. BAYESIAN MODEL

This section introduces the Bayesian model used to estimate the unknown factor score vectors $\{\mathbf{a}_i\}_{i=1,\ldots,N}$ and the factor signature $\{\mathbf{m}_r\}_{r=1,\ldots,R}$ under the constraints specified in (2). This model is based on the likelihood of the observations and on prior distributions for the unknown parameters.

This work was partially supported by the National Science Foundation grant CCF 0830490.

2.1. Likelihood function

The LMM defined in (1) and the statistical properties of the residual error $\mathbf{n}_i$ (3) lead to a conditionally Gaussian distribution for the ith observed sample, i.e., $\mathbf{y}_i|\mathbf{M}, \mathbf{a}_i, \Sigma \sim \mathcal{N}(\mathbf{M}\mathbf{a}_i, \Sigma)$. Thus, the likelihood function of $\mathbf{Y}$ can be expressed as

$$f(\mathbf{Y}|\mathbf{M}, \mathbf{A}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{GN/2}} \exp\left[-\frac{\sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{M}\mathbf{a}_i\|^2}{2\sigma^2}\right] \tag{4}$$

where $\|\cdot\|$ stands for the l_2-norm.

2.2. Parameter priors

2.2.1. Factor loading prior

Due to the constraints in (2), the observed data $\mathbf{y}_i$, $i = 1, \ldots, N$, lie in a simplex defined in a lower-dimensional subspace of $\mathbb{R}^{G-1}$ (of dimension $K-1$) denoted as $\mathcal{V}_{K-1}$, with $R \leq K \leq G$. This subspace can be identified with a standard dimension reduction method such as principal component analysis (PCA). Following the approach in [3], instead of estimating the factor loadings $\mathbf{m}_k$ ($k = 1, \ldots, K$) in the observation space $\mathbb{R}^G$, we propose to estimate their projections $\mathbf{t}_k$ ($k = 1, \ldots, K$) onto $\mathcal{V}_{K-1}$. Let $\bar{\mathbf{y}}$ be the empirical mean of the observed vectors and $\mathbf{P}$ the $(K-1) \times G$ projection matrix onto $\mathcal{V}_{K-1}$, e.g., composed of appropriate eigenvectors of the empirical covariance matrix of $\mathbf{Y}$. The prior distributions for the projected factors

$$\mathbf{t}_k = \mathbf{P}(\mathbf{m}_k - \bar{\mathbf{y}}) \tag{5}$$

are modeled as multivariate normal distributions $\mathcal{N}_{\mathcal{T}_k}\left(\mathbf{e}_k, s_k^2\mathbf{I}_{k-1}\right)$ truncated on the set $\mathcal{T}_k$. The truncation on the set $\mathcal{T}_k$ (defined in [3]) ensures that all the components of the factor signatures are positive

$$\{m_{k,g} \geq 0, \forall g = 1, \ldots, G\} \quad \Leftrightarrow \quad \{\mathbf{t}_k \in \mathcal{T}_k\}. \tag{6}$$

The mean vectors $\mathbf{e}_k$ are fixed using available prior knowledge or, as in [3], provided by an endmember extraction algorithm dedicated to hyperspectral image analysis. The variance s_k reflects the degree of confidence given to this prior information (it will be fixed to a large value in this paper). Assuming that the projected factors are *a priori* independent, the joint prior distribution for $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_K]$ is $f(\mathbf{T}) = \prod_{k=1}^{K} f(\mathbf{t}_k)$.

2.2.2. Factor score prior

Consider a Gaussian distribution, with hidden mean zero and hidden variance parameter α^2. This distribution is truncated on the interval $[0, \mu^+]$ denoted as $\mathcal{N}_{[0,\mu^+]}(0, \alpha^2)$ and its probability density function (pdf) can be expressed as [4]

$$\varphi_{[0,\mu^+]}(x) = \frac{C}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{x^2}{2\alpha^2}\right) \mathbf{1}_{[0,\mu^+]}(x) \tag{7}$$

where μ^+ is the right truncation bound, $\mathbf{1}_{\mathbb{E}}(x)$ is the indicator function defined on $\mathbb{E}$ (i.e., $\mathbf{1}_{\mathbb{E}}(x) = 1$ if $x \in \mathbb{E}$ and $\mathbf{1}_{\mathbb{E}}(x) = 0$ if $x \notin \mathbb{E}$). In (7), $C = \left[\Phi\left(\frac{\mu^+}{\alpha}\right) - \frac{1}{2}\right]^{-1}$ is a normalization constant, where Φ denotes the cumulative density function (cdf) of the standard normal distribution. As it can be seen later, the truncation on the interval $[0, \mu^+]$ will be used to ensure the non-negativity and full-additivity constraints of the factor scores.

From the LMM in (1), one can notice that only $R < K$ factors among the K contained in the library $\mathbf{M}$ are actually involved in the mixture. In other words, most of the coefficients $a_{i,k}$, $k = 1, \ldots, K$ equal zero. Consequently, a distribution that enforces sparsity should be chosen as prior for the factor scores $a_{i,k}$. Following the approach in [5], the distribution in (7) is coupled with an atom at zero, leading to a prior density mixture. More precisely, denote as $\mathbf{a}_{i,1:k-1}$ the vector composed of the first $k-1$ components of the factor score vector $\mathbf{a}_i$. The following truncated Bernoulli-Gaussian distribution[1] is chosen as prior distribution for the factor scores $a_{i,1}$ and $a_{i,k}$ ($k = 2, \ldots, K-1$)

$$\begin{array}{ll} a_{i,1} & \sim (1-w_i)\delta(a_{i,1}) + w_i\mathcal{N}_{[0,1]}(0, \alpha^2), \\ a_{i,k}|\mathbf{a}_{i,1:k-1} & \sim (1-w_i)\delta(a_{i,k}) + w_i\mathcal{N}_{[0,\mu_{i,k}^+]}(0, \alpha^2), \end{array} \tag{8}$$

where $\delta(\cdot)$ is the Dirac function and w_i is an unknown hyperparameter which provides the prior probability of having a non-zero factor score. To ensure the additivity constraint, the right truncation point $\mu_{i,k}^+$ is fixed to $\mu_{i,k}^+ = 1 - \sum_{j=1}^{k-1} a_{i,j}$ ($k = 2, \ldots, K-1$) whereas the last factor score is set to $a_{i,K} = \mu_{i,K}^+ \triangleq 1 - \sum_{k=1}^{K-1} a_{i,k}$. Finally, the prior distribution for the score vector $\mathbf{a}_i$ whose last element $a_{i,K}$ has been fixed to $\mu_{i,K}^+$ can be expressed as the recursion

$$f(\mathbf{a}_i) = f(a_{i,1}) \left[\prod_{k=2}^{K-1} f(a_{i,k}|\mathbf{a}_{i,1:k-1})\right] \delta\left(a_{i,K} - \mu_{i,K}^+\right).$$

Assuming the score vectors $\mathbf{a}_i$ (for $i = 1, \ldots, N$) are *a priori* independent from sample to sample, the joint prior distribution for the factor score matrix $\mathbf{A}$ is $f(\mathbf{A}) = \prod_{i=1}^{N} f(\mathbf{a}_i)$.

2.2.3. Noise variance prior

A conjugate inverse-Gamma distribution with parameters $\nu/2$ and $\gamma/2$ is chosen as prior distribution for the noise variance

$$\sigma^2|\nu, \gamma \sim \mathcal{IG}\left(\frac{\nu}{2}, \frac{\gamma}{2}\right). \tag{9}$$

The shape parameter ν will be fixed to $\nu = 2$ whereas the scale parameter γ will be an adjustable hyperparameter (as in [3, 5]).

2.3. Hyper-parameter priors

Let $\mathbf{\Psi} = \{\boldsymbol{w}, \gamma\}$ be the hyperparameter vector associated to the model defined above, with $\boldsymbol{w} = [w_1, \ldots, w_N]^T$. The accuracy of the proposed Bayesian estimation algorithm depends on the values of these hyperparameters. The approach investigated here is to assign these hyperparameters appropriate priors (also referred to as *hyperpriors*) following hierarchical Bayesian inference.

More precisely, a uniform distribution on the set $[0, 1]$ is chosen as prior distribution for the mean proportion of non-zero score coefficients, i.e., $w_i \sim \mathcal{U}([0, 1])$. Following [3, 5], a non-informative Jeffreys' prior is chosen as the prior distribution for the hyperparameter γ, i.e., $f(\gamma) \propto \frac{1}{\gamma}\mathbf{1}_{\mathbb{R}^+}(\gamma)$. Assuming that all the individual hyperparameters of this Bayesian model are statistically independent, the full posterior distribution of the hyperparameter vector $\mathbf{\Psi}$ is

$$f(\mathbf{\Psi}) = f(\boldsymbol{w})f(\gamma) \propto \frac{1}{\gamma} \prod_{i=1}^{N} \mathbf{1}_{[0,1]}(w_i)\, \mathbf{1}_{\mathbb{R}^+}(\gamma) \tag{10}$$

where $\propto$ stands for "proportional to".

[1] Note that the dependence upon the hyperparameters w_i and α^2 is implicit in the notation.

2.4. Posterior distribution

The joint posterior distribution of the unknown parameter vector $\boldsymbol{\Theta} = \{\mathbf{T}, \mathbf{A}, \sigma^2\}$ and the hyperparameter vector $\boldsymbol{\Psi} = \{\boldsymbol{w}, \gamma\}$ can be computed as

$$f(\boldsymbol{\Theta}, \boldsymbol{\Psi}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}|\boldsymbol{\Psi}) f(\boldsymbol{\Psi}) \tag{11}$$

where $f(\mathbf{Y}|\boldsymbol{\Theta})$ and $f(\boldsymbol{\Psi})$ have been respectively defined in the equations (4) and (10). Assuming *a priori* independence between the individual parameters, the following prior is obtained

$$f(\boldsymbol{\Theta}|\boldsymbol{\Psi}) = f(\mathbf{T}) f(\mathbf{A}|\boldsymbol{w}, \alpha^2) f(\sigma^2|\nu, \gamma). \tag{12}$$

3. GIBBS SAMPLER

This section proposes a Gibbs sampling strategy for generating random samples (denoted by $\cdot^{(\ell)}$, where ℓ is the iteration index), asymptotically distributed according to the joint posterior distribution defined in (11). This Markov chain Monte Carlo (MCMC) technique consists of generating sequences $\{\mathbf{T}^{(\ell)}\}_{\ell=1,\ldots}$, $\{\mathbf{A}^{(\ell)}\}_{\ell=1,\ldots}$, $\{\sigma^{2^{(\ell)}}\}_{\ell=1,\ldots}$ and $\{\boldsymbol{w}^{(\ell)}\}_{\ell=1,\ldots}$ according to the conditional posterior distributions, as detailed below (the interested reader is invited to consult [3] for more details regarding the MCMC implementation).

3.1. Sampling from $f(\mathbf{T}|\mathbf{A}, \sigma^2, \mathbf{Y})$

Sampling from the joint conditional $f(\mathbf{T}|\mathbf{A}, \sigma^2, \mathbf{Y})$ is achieved by updating each column of $\mathbf{T}$ using Gibbs moves. Let denote $\mathbf{T}_{\backslash k}$ the matrix $\mathbf{T}$ whose kth column has been removed. The posterior distribution of $\mathbf{t}_k$ is the following truncated multivariate Gaussian distribution

$$\mathbf{t}_k|\mathbf{T}_{\backslash k}, \mathbf{a}_k, \sigma^2, \mathbf{Y} \sim \mathcal{N}_{\mathcal{T}_k}\left(\boldsymbol{\tau}_k, \boldsymbol{\Gamma}_k\right) \tag{13}$$

where

$$\begin{aligned} \boldsymbol{\Gamma}_k &= \left[\textstyle\sum_{i=1}^{N} a_{i,k}^2 \mathbf{P}\Sigma^{-1}\mathbf{P}^T + \frac{1}{s_k^2}\mathbf{I}_K\right]^{-1}, \\ \boldsymbol{\tau}_k &= \boldsymbol{\Gamma}_k \left[\textstyle\sum_{i=1}^{N} a_{i,k} \mathbf{P}\Sigma^{-1}\boldsymbol{\epsilon}_{i,k} + \frac{1}{s_k^2}\mathbf{e}_k\right], \\ \boldsymbol{\epsilon}_{i,k} &= \mathbf{y}_i - a_{i,k}\bar{\mathbf{y}} - \textstyle\sum_{j\neq k} a_{i,k}\mathbf{m}_j. \end{aligned} \tag{14}$$

For more details on how we generate realizations from this truncated distribution, see [3].

3.2. Sampling from $f(\mathbf{A}|\boldsymbol{w}, \sigma^2, \mathbf{Y})$

Similarly, straightforward computations lead to the following posterior distribution of each element of $\mathbf{A}$

$$a_{i,k}|w_i, \sigma^2, \mathbf{a}_{i,\backslash k}, \mathbf{y}_i \sim (1-\widetilde{w}_{i,k})\delta\left(a_{i,k}\right) + \widetilde{w}_{i,k}\mathcal{N}_{]0,\mu_{i,k}^{+}[}(\mu_{i,k}, \eta_{i,k}^2) \tag{15}$$

where $\mathbf{a}_{i,\backslash k}$ denotes the score vector $\mathbf{a}_i$ whose kth element has been removed and

$$\begin{cases} \widetilde{w}_{i,k} &= \frac{u_{i,k}}{u_{i,k}+(1-w_i)}, \\ u_{i,k} &= w_i \frac{\eta_{i,k}}{\alpha} \exp\left(\frac{\mu_{i,k}^2}{2\eta_{i,k}^2}\right)\left[\Phi\left(\frac{\mu_{i,k}^{+}-\mu_{i,k}}{\eta_{i,k}}\right) - \Phi\left(\frac{-\mu_{i,k}}{\eta_{i,k}}\right)\right], \\ \eta_{i,k}^2 &= \left(\frac{\|\mathbf{m}_k\|^2}{\sigma^2} + \frac{1}{\alpha^2}\right)^{-1}, \\ \mu_{i,k} &= \eta_{i,k}^2\left(\frac{\mathbf{m}_k^T\boldsymbol{\epsilon}_{\backslash k}}{\sigma^2}\right), \\ \boldsymbol{\epsilon}_{\backslash k} &= \mathbf{y}_i - \sum_{j=1,j\neq k}^{K}\mathbf{m}_j a_{i,j}. \end{cases}$$

Therefore, factor scores will be sampled from this Bernoulli-truncated Gaussian distribution with parameters $(\widetilde{w}_{i,k}, \mu_{i,k}, \eta_{i,k}^2, \mu_{i,k}^{+})$.

3.3. Sampling from $f(\boldsymbol{w}|\mathbf{A})$

Generating samples distributed according to $f(\boldsymbol{w}|\mathbf{A})$ can be achieved using N Gibbs moves using $(i = 1, \ldots, N)$

$$w_i|\mathbf{a}_i \sim \mathcal{B}\left(1 + n_{1,i}, 1 + n_{0,i}\right) \tag{16}$$

where $n_{1,i} = \sharp\{k|a_{i,k} \neq 0\}$, and $n_{0,i} = K - n_{1,i}$.

3.4. Sampling from $f(\sigma^2|\mathbf{M}, \mathbf{A}, \mathbf{Y})$

Using (9) and (4), one can show that the conditional distribution $f(\sigma^2|\mathbf{M}, \mathbf{A}, \mathbf{Y})$ is the following inverse-Gamma distribution

$$\sigma^2|\mathbf{M}, \mathbf{A}, \mathbf{Y} \sim \mathcal{IG}\left(\frac{GN}{2}, \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{y}_i - \mathbf{M}\mathbf{a}_i\|^2\right). \tag{17}$$

4. SIMULATION RESULTS ON SYNTHETIC DATA

To illustrate the performance of the proposed unsupervised Bayesian algorithm for gene expression factor analysis, simulations are conducted on a synthetic dataset consisting of $N = 128$ observed samples and $G = 256$ gene expression levels. Each sample is composed of exactly $R = 4$ factors, selected from $K = 9$ possible biological pathways. The factor scores have been randomly generated according to a Dirichlet distribution $\mathcal{D}(1, \ldots, 1)$ and the observed vectors are corrupted by an i.i.d. noise sequence.

The factor loading library $\mathbf{M}$ is unknown and must be estimated. The hidden mean vectors $\mathbf{e}_k$ $(k = 1, \ldots, K)$ required for evaluating the loading prior introduced in Section 2.2.1 are chosen as the PCA projections of signatures previously identified by vertex component analysis (VCA) [6]. The SNR has been fixed in this simulation to SNR= 20 dB.

The MMSE estimates of the factor score vectors $\mathbf{a}_i$ $(i = 1, \ldots, N)$ and the projected factor loading vectors $\mathbf{t}_k$ $(k = 1, \ldots, K)$ are approximated using the generated samples, as in [7]. The associated MSEs for factor score vectors and factor loading vectors are defined as

$$\mathrm{GMSE}_r^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{a}_{i,r} - a_{i,r}\right)^2, \mathrm{MSE}_r^2 = \|\widehat{\mathbf{m}_r} - \mathbf{m}_r\|^2 \tag{18}$$

for $r = 1, \ldots, R$. The results are reported in Table 1 where the proposed BeG method is compared to the unsupervised NPBFA method [1], the Bayesian factor regression modeling (BFRM) proposed by Carvalho *et al.* [2], the non-negative matrix factorization (NMF) [8] and the PCA algorithm. The NMF and PCA methods have been run for the actual number of factors to be estimated, i.e. $R = 4$. Moreover, due to the constraints of positivity and additivity, the linear mixing solution of the BeG model is unique up to a permutation of the factors, whereas for the other factor decomposition methods, a re-scaling is also needed. The proposed Bayesian method exhibits significantly better performance for these examples. This improved accuracy can be attributed to the fact that the proposed method incorporates the non-negativity and sum-to-one constraints.

Table 1. MSEs of the estimates for the BeG, NPBFA, BFRM, NMF and PCA methods.

		BeG	NPBFA	BFRM	NMF	PCA
MSE^2 $(\times 10^3)$	Factor 1	0.372	1.352	1.827	3.438	2.186
	Factor 2	0.288	1.558	1.434	3.303	0.364
	Factor 3	0.016	1.237	1.946	4.281	0.413
	Factor 4	0.012	2.645	N/A	6.580	0.381
GMSE^2	Factor 1	6.955	40.013	198.735	19.709	4.191
	Factor 2	8.065	35.282	183.638	34.913	0.022
	Factor 3	5.410	23.687	191.699	17.254	7.069
	Factor 4	14.293	26.766	N/A	18.802	5.119

5. SIMULATION RESULTS ON REAL DATA

This section illustrates the proposed algorithm on a public dataset described in [9]. This dataset consists of the gene expression levels of $N = 108$ Affymetrix chips collected on six subjects, at five time points: 0, 1, 2, 4 and 12 hours after the subjects have imbibed one of four different beverages (alcohol, grape juice, water and red wine).

The BeG factor analysis is applied on the data with $K = 9$ factors whose loading spectra are shown in Fig. 1. Choosing $K = 9$ allows the dimensionality of the problem to be significantly reduced while keeping a sufficient cumulative energy. The figure shows the loading coefficients plotted over the $G = 22283$ gene indices after reordering these indices so as to group together the dominant genes in each factor. Specifically, the k-th sharp peak in the figure occurs at the gene index that has maximal loading in factor k and genes to the right of this gene index up to the $(k+1)$-st peak also dominate in the k-th factor, but to a lesser degree. The 9 factors discovered are dominated by groups of genes of sizes $[5160, 3, 10045, 1455, 1315, 224, 147, 3352, 582]$.

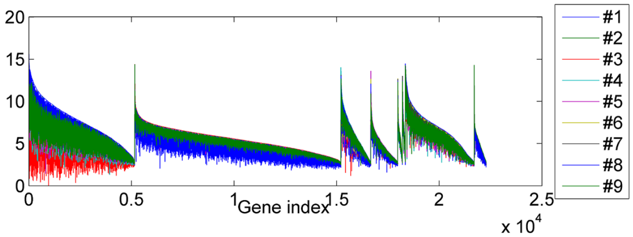

Fig. 1. Factor loadings ranked by decreasing dominance.

The factor scores are shown in Fig. 2 for each of the 9 factors. For each factor, these scores are rendered as an image whose columns index the 6 subjects and whose rows index the 5 time points under each of the beverage treatments. Note from the images that several of the factors are strongly associated with particular individuals, e.g., factor 5 (subject 5), factor 6 (subjects 1, 2), factor 7 (subjects 4, 6), and factor 9 (subjects 1, 3). Other factors are more globally associated with treatment and time.

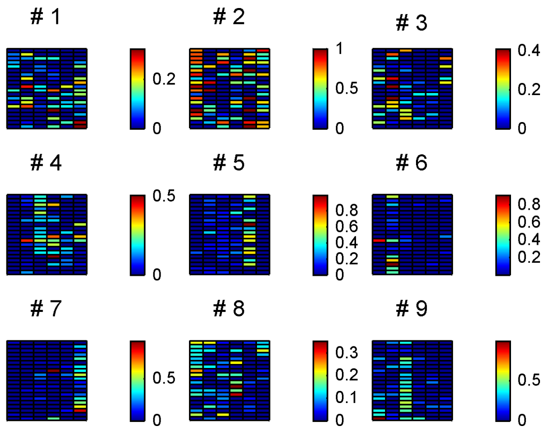

Fig. 2. Factors scores for each of the 9 factors.

Figure 3 shows how the factor scores can be used as features to visualize the samples. In the figure the score vector for each sample is mapped to a coordinate in the plane using euclidean multidimensional scaling (MDS). Each sample is embedded with a color and a size denoting the beverage treatment and the time stamp of each sample. Note the interesting structure of the data in this MDS domain - the late-time courses of wine (blue) and grape juice (red) are separated from each other possibly indicating dichotomous gene response over the population of subjects.

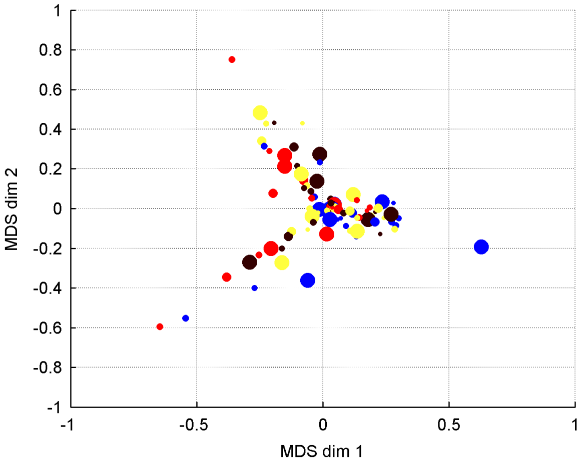

Fig. 3. Chip cloud after demixing: alcohol = black, grape juice = red, water = yellow, wine = blue.

6. CONCLUSIONS

This paper presented a Bayesian estimation algorithm for gene-expression data. To ensure the positivity and the full-additivity of the abundance vector, a constrained Bernoulli-Gaussian distribution was chosen as a prior distribution for the factor scores. Due to the complexity of the posterior distribution, a Gibbs sampler algorithm was proposed to generate samples distributed according to this posterior. Then, the MMSE and MAP estimator were computed using these generated samples. The simulation results conducted on synthetic and real data illustrated the performance of the proposed Bayesian algorithm.

7. REFERENCES

[1] B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G. S. Ginsburg, A. O. Hero, J. Lucas, D. Dunson, and L. Carin, "Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies," *BMC Bioinformatics*, vol. 11, no. 1, p. 552, 2010.

[2] 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.

[3] 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.

[4] C. P. Robert, "Simulation of truncated normal variables," *Statistics and Computing*, vol. 5, no. 2, pp. 121–125, June 1995.

[5] N. Dobigeon, A. O. Hero, and J.-Y. Tourneret, "Hierarchical Bayesian sparse image reconstruction with application to MRFM," *IEEE Trans. Image Processing*, vol. 18, no. 9, pp. 2059–2070, Sept. 2009.

[6] 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.

[7] N. Dobigeon and J.-Y. Tourneret, "Bayesian orthogonal component analysis for sparse representation," *IEEE Trans. Signal Processing*, vol. 58, no. 5, pp. 2675–2685, May 2010.

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

[9] 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.