The Annals of Applied Statistics
2013, Vol. 7, No. 4, 2431–2457
DOI: 10.1214/13-AOAS643

MULTI-WAY BLOCKMODELS FOR ANALYZING COORDINATED HIGH-DIMENSIONAL RESPONSES[1]

BY EDOARDO M. AIROLDI, XIAOPEI WANG AND XIAODONG LIN

Harvard University, University of Cincinnati and Rutgers University

We consider the problem of quantifying temporal coordination between multiple high-dimensional responses. We introduce a family of multi-way stochastic blockmodels suited for this problem, which avoids preprocessing steps such as binning and thresholding commonly adopted for this type of data, in biology. We develop two inference procedures based on collapsed Gibbs sampling and variational methods. We provide a thorough evaluation of the proposed methods on simulated data, in terms of membership and blockmodel estimation, predictions out-of-sample and run-time. We also quantify the effects of censoring procedures such as binning and thresholding on the estimation tasks. We use these models to carry out an empirical analysis of the functional mechanisms driving the coordination between gene expression and metabolite concentrations during carbon and nitrogen starvation, in *S. cerevisiae*.

1. Introduction. In recent years, the biology community at large has engaged in an effort to characterize *coordinated* mechanisms of cellular regulation, to enable a *systems-level* understanding of cellular functions. Reference databases, such as the yeast genome database (SGD), catalog the many regulatory roles of genes and proteins with links to the originating literature [Cherry et al. (1997), Kanehisa and Goto (2000)]. Recent work spans approaches that leverage these databases to integrate genomic information across multiple studies and technologies about the same regulatory mechanism, for example, transcription [Cope et al. (2004), Franks et al. (2012)], as well as approaches to integrate genomic information across levels of regulation, for example, epigenetic markers, chromatin modifications, transcription and translation [Troyanskaya et al. (2003), Lu et al. (2009), Markowetz et al. (2009)].

We consider the problem of quantifying temporal coordination between gene expression and metabolite concentrations in yeast [Brauer et al. (2006, 2008)]. More generally, we are interested in statistical methods to analyze multiple coordinated high-dimensional measurements about a system organism, where correlation among pairs of measurements is believed to indicate coordinated functional

Received December 2011; revised December 2012.

[1]Supported in part by NSF Grants DMS-11-06980, IIS-10-17967 and CAREER IIS-11-49662, and by NIH Grant R01 GM096193 all to Harvard University, by a faculty research grant from Rutgers Business School, and by a Taft fellowship to the University of Cincinnati.

Key words and phrases. High dimensional data, variational inference, molecular biology, yeast.

and regulatory roles. We develop methods for analyzing experiments on regulation dynamics that involve the following: (1) data collections about multiple stages of regulations (transcriptional and metabolic) that offer complementary views of the cellular response (to Nitrogen and Carbon starvation), quantified in terms of high-dimensional measurements; and (2) data collected according to a specific coordinated temporal design, whereby the experiments at different stages of regulation are conducted on cell cultures with matching conditions (nutrient limitations, environmental stress and chemical compounds present) over time. Coordinated time courses about complementary stages of regulation arguably provide the best opportunity to characterize coordinated regulation dynamics, quantitatively.

A popular approach to study coordinated cellular responses in biology involves Bayesian networks [Bradley et al. (2009), Troyanskaya et al. (2003)]. This approach requires *binning* real-valued measurements into discrete categories. A deterministic alternative to explore coordination is the cross-associations algorithm [Chakrabarti et al. (2004)], which instead requires *thresholding* the matrix of correlations between pairs of genes and metabolites into binary on–off relations. While binning and thresholding are accepted data preprocessing steps in the computational biology literature, they raise serious statistical issues [Blocker and Meng (2013)]. On the one hand, the lack of appropriate and principled alternatives, together with the sizable amount of data typical in a coordinated study of cellular responses, for example, genome-wide expression and hundreds of metabolites, make preprocessing necessary. These preprocessing steps reduce the computational burden of the analysis with Bayesian networks and cross-associations. On the other hand, however, these preprocessing steps are essentially censoring mechanisms that may compromise the patterns of variation and covariation in the original data, when the discovery in such patterns, local and global, is the primary goal of the analysis [Turnbull (1976), Vardi (1985)].

In this paper we develop a family of blockmodels to analyze a correlation matrix among sets of temporally paired measurements on two distinct populations of objects. Our work extends a recent block modeling approach that leverages the notion of *structural equivalence* [Snijders and Nowicki (1997), Nowicki and Snijders (2001)] to the analysis of coordinated measurements on two populations. For more details on blockmodels see Goldenberg et al. (2009). Section 2 introduces two-way (and multi-way) stochastic blockmodels for a function of the high-dimensional responses, such as their correlation. These simple models explicitly allow different objects in the two (or more) populations to be associated with multiple blocks, say, of correlation, to different degrees, and does not require binning or thresholding. Estimation and inference using variational methods is outlined in Section 2.4. Details of variational and MCMC inference are provided in the supplement [Airoldi, Wang and Lin (2013b)]. Section 3 develops a thorough evaluation of the proposed methods on simulated data, including a comparative evaluation of the MCMC and variational inference procedures in terms of the following: (1) membership and blockmodel matrix estimation, (2) predictions out-of-sample, and (3) run-time.

We assess the effects of thresholding on inference in Section 3.7. In Section 4 we analyze two recently published collections of time-course data to explore the functional mechanisms underlying the coordination of transcription and metabolism during carbon and nitrogen starvation, in *S. cerevisiae*. We compare the results with published results on the same data using binning and Bayesian networks, and to new results we obtain using thresholding and cross-associations.

2. Multi-way stochastic blockmodels. In this section we introduce multi-way stochastic blockmodels and the associated inference procedures. This family of models generalizes mixed membership stochastic blockmodels for analyzing interactions within a single population [Airoldi et al. (2008)] to interactions between two or more populations. Multi-way stochastic blockmodels models enable the discovery of interactions between latent groups across different populations, and provide estimates of the group memberships for each subject. We develop two inference strategies: one based on collapsed Gibbs sampling [Liu (1994)], the other based on variational Expectation–Maximization (vEM) [Jordan et al. (1999), Airoldi (2007)].

2.1. *Two-way blockmodels.* Consider a two-way interaction table between two sets of nodes $\mathcal{N}_1$ and $\mathcal{N}_2$ of size N_1 and N_2, respectively. These two sets of nodes represent elements of two distinct populations. An observation $Y(j,k)$, $j = 1,\ldots,N_1$, $k = 1,\ldots,N_2$, denotes the strength of the interaction between the jth element of $\mathcal{N}_1$ and the kth element of $\mathcal{N}_2$.

As a running example, we consider the coordinated time course data we analyze in Section 4. The data consists of N_1 time series of gene expression levels and of N_2 time series of metabolite concentrations, before and after Nitrogen and Carbon starvation for a total of seven time points, in yeast [Brauer et al. (2006), Bradley et al. (2009)]. We posit a model for the $N_1 \times N_2$ matrix of Fisher-transformed correlations of time courses for each gene–metabolite pair or for any of its submatrices obtained by selecting subsets of genes and metabolites of special interest to biologists. The goal of the analysis is to reveal interactions between gene functions and metabolic pathways, operationally defined as sets of genes and sets of metabolites, respectively, with similar correlation patterns.

In the context of this application, we posit that each gene can participate in up to K_1 functions, that is, latent row groups, and that each metabolite can participate in up to K_2 metabolic pathways, that is, latent column groups.[2] Latent Dirichlet vectors $\vec{\pi}_j$ and $\vec{p}_k$ capture the relative fractions of time gene j and metabolite k participate in the different cellular functions and pathways, or latent groups. The distribution of the correlation, or, more generally, interaction, $Y(j,k)$, is then

[2]We refer to gene functions and metabolic pathways as defined in the yeast genome database and the Kyoto encyclopedia of genes and genomes.

a function of the interactions among the latent groups, fully specified by a $K_1 \times K_2$ matrix B, together with the latent memberships of the gene and metabolite involved. The data generating process, given α, β, B and σ, is as follows:

$$\vec{\pi}_j \sim \text{Dirichlet}(\alpha), \tag{2.1}$$

$$\vec{p}_k \sim \text{Dirichlet}(\beta), \tag{2.2}$$

$$Y(j,k) \sim \text{Normal}\big(\vec{\pi}_j' B \vec{p}_k, \sigma^2\big), \tag{2.3}$$

where indices $j = 1, \ldots, N_1$ and $k = 1, \ldots, N_2$ run over genes and metabolites, respectively, vectors $\vec{\pi}_j$ and $\vec{p}_k$ are K_1- and K_2-dimensional, respectively, and elements of the blockmodel mean matrix $B_{gh} \in \mathbb{R}$.

While the observations $Y(j,k)$ in the motivation application are Fisher-transformed correlations, real-valued with real-valued mean matrix B, the proposed models are more flexible. For instance, we develop a two-way block model for binary observations in Section 2.2, that is used in Section 3.7 for quantifying the effects of censoring the data matrix Y.

For inference purposes, we consider an augmented data generating process, in which we introduce latent indicator vectors $\vec{D}_{j\to k}$ and $\vec{E}_{j\leftarrow k}$ that denote the single memberships of gene j and metabolite k for the correlation $Y(j,k)$. The latent indicators $\{D, E\}$ do not have a clear biological interpretation, but serve to improve computational tractability of the inference; they lead to optimization problems that have analytical solutions. The trade-offs of such a strategy have been explored elsewhere [e.g., see Airoldi et al. (2008)]. From a statistical perspective, introducing $\{D, E\}$ amounts to a specific representation of the interactions in terms of random effects.

2.2. *Extension to non-Gaussian responses.* In the data generating process above, Y is generated from a Normal distribution and the blockmodel's elements take real values. Extending the proposed model to other distributions to account for data Y that live in a different space is straightforward. And because of the hierarchical structure of the model, only a minor portion of the inference and estimation strategies detailed in Section 2.4 will need to be modified appropriately, as a consequence.

We will consider one such extension to binary observations $Y(j,k)$—namely, correlations after thresholding—in Section 3.7 to assess the effects of preprocessing on the accuracy in estimating the blockmodel. The data generating process in Section 2.1 is modified as follows. The blockmodel's elements now take values in the unit interval, since they capture the probability that there is a correlation above threshold between members of any pair of blocks, $B_{gh} \in [0, 1]$. For each pair (j,k), $j = 1, \ldots, N_1$, $k = 1, \ldots, N_2$, we sample the pairwise binary observation $Y(j,k) \sim \text{Bernoulli}(\vec{D}'_{j\to k} B \vec{E}_{j\leftarrow k})$. Variational Bayes and MCMC inference also remain mostly unchanged. New updating equations for the elements of B will be needed; see equation (2.11) and the supplement [Airoldi, Wang and Lin (2013b)].

2.3. *Extension to multi-way blockmodels.* The two-way blockmodel introduced above can also be extended for analyzing multi-way interactions between three or more populations.

Consider a three-way interaction table $Y(i_1, i_2, i_3)$ observed on three populations $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$, where $i_1 \in \mathcal{N}_1$, $i_2 \in \mathcal{N}_2$ and $i_3 \in \mathcal{N}_3$. Assume that there are K_1, K_2 and K_3 latent groups existing in $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}_3$, respectively. We can treat the three way interaction observed in Y as a result of three way group interactions. Namely, $Y(i_1, i_2, i_3)$ can be fully characterized by $B(g_1, g_2, g_3)$, with items $\{i_1, i_2, i_3\}$ belonging to group $\{g_1, g_2, g_3\}$, respectively. Therefore, inferences procedures for this three-way blockmodel can be developed in a similar fashion as those for the two-way blockmodel. Note that although the ideas for generalizations to higher order tables remain the same, keeping track of indices during inference becomes tedious.

2.4. *Parameter estimation and posterior inference.* The main inference task is to estimate the matrix B and the mixed membership vectors $\vec{\pi}$ and $\vec{p}$. Given the observed data $Y = Y(j,k)$, latent variable $X = \{\vec{\pi}_j, \vec{p}_k, \vec{D}_{j\to k}, \vec{E}_{j\leftarrow k}\}$ and the parameters $\Theta = \{\alpha, \beta, \sigma^2, B\}$, the complete data likelihood $p(Y, X|\Theta)$ can be written as

$$
\begin{aligned}
&p\big(Y, X|\alpha, \beta, B, \sigma^2\big) \\
&\quad = \prod_j p_1(\vec{\pi}_j|\alpha) \prod_k p_1(\vec{p}_k|\beta) \\
&\qquad \times \prod_{j,k} p_0\big(Y(j,k)|\vec{D}_{j\to k}, \vec{E}_{j\leftarrow k}, B, \sigma^2\big) p_2(\vec{D}_{j\to k}|\vec{\pi}_j) p_2(\vec{E}_{j\leftarrow k}|\vec{p}_k),
\end{aligned} \tag{2.4}
$$

where p_0 is a Normal distribution with mean $\mu = \vec{D}'_{j\to k} B \vec{E}_{j\leftarrow k}$ and variance σ^2, p_1 is a Dirichlet distribution, and p_2 is a Multinomial distribution with $n = 1$. The posterior distribution of the latent variable X is

$$
p(X|Y, \Theta) = \frac{p(Y, X|\Theta)}{p(Y|\Theta)}, \tag{2.5}
$$

where the marginal distribution $p(Y|\Theta)$ has the following form:

$$
\begin{aligned}
p(Y|\Theta) &= \int_X p(Y, X|\Theta)\, dX \\
&= \sum_{\vec{D}} \sum_{\vec{E}} \Big\{ \int\int \prod_j p_1(\vec{\pi}_j|\alpha) \prod_k p_1(\vec{p}_k|\beta) \\
&\qquad \times \prod_{j,k} p_2(\vec{D}_{j\to k}|\vec{\pi}_j) p_2(\vec{E}_{j\leftarrow k}|\vec{p}_k)\, d\vec{\pi}\, d\vec{p} \\
&\qquad \times \prod_{j,k} p_0\big(Y(j,k)|\vec{D}_{j\to k}, \vec{E}_{j\leftarrow k}, B, \sigma^2\big) \Big\}.
\end{aligned}
$$

There does not exist an explicit solution to the maximization of $p(Y|\Theta)$. Therefore, we propose an iterative procedure based on variational Bayes for parameter estimation. In comparison, we also develop a MCMC scheme based on collapsed Gibbs sampling to achieve the desired statistical inferences.

2.4.1. *Variational expectation–maximization.* To achieve variational inference, we introduce free variational parameters $\vec{\nu}_j$ and $\vec{\xi}_k$ to approximate $\vec{\pi}_j$ and $\vec{p}_k$, free variational variables $\vec{\phi}_{j\to k}$ and $\vec{\eta}_{j\leftarrow k}$ to approximate $\vec{D}_{j\to k}$ and $\vec{E}_{j\leftarrow k}$, and latent distribution $q(X)$ to approximate the true posterior distribution $p(X|Y,\Theta)$. By Jensen's inequality, we have the following likelihood lower bound:

$$\log p(Y|\Theta) \geq E_q\big[\log p(Y,X|\Theta)\big] - E_q\big[\log q(X)\big]. \tag{2.6}$$

A coordinate ascend algorithm can be applied to obtain a local maximizer of this lower bound, which results in the updates (2.7)–(2.11). Detailed derivations are left in the supplementary material [Airoldi, Wang and Lin (2013b)]. The resulting variational EM algorithm is given in Algorithm 1:

$$\begin{aligned}\phi_{j\to k,g} \propto{}& \exp\left(\psi(\nu_{j,g}) - \psi\left(\sum_g \nu_{j,g}\right)\right)\\ &\times \prod_h \big(\sigma^2 \cdot e^{(Y(j,k)-B(g,h))^2/\sigma^2}\big)^{-1/2\eta_{j\leftarrow k,h}},\end{aligned} \tag{2.7}$$

$$\begin{aligned}\eta_{j\leftarrow k,h} \propto{}& \exp\left(\psi(\xi_{k,h}) - \psi\left(\sum_h \xi_{k,h}\right)\right)\\ &\times \prod_g \big(\sigma^2 \cdot e^{(Y(j,k)-B(g,h))^2/\sigma^2}\big)^{-1/2\phi_{j\to k,g}},\end{aligned} \tag{2.8}$$

$$\nu_{j,g} = \sum_k \phi_{j\to k,g} + \alpha, \tag{2.9}$$

$$\xi_{k,h} = \sum_j \eta_{j\leftarrow k,h} + \beta, \tag{2.10}$$

$$B(g,h) = \frac{\sum_{j,k} \phi_{j\to k,g}\eta_{j\leftarrow k,h} Y(j,k)}{\sum_{j,k} \phi_{j\to k,g}\eta_{j\leftarrow k,h}}. \tag{2.11}$$

3. Evaluating inference and effects of preprocessing. Here we use simulated data to compare the performance of variational and MCMC inference procedures for the two-way block model along multiple dimensions: estimation accuracy of mixed membership vectors, accuracy of predictions out-of-sample, estimation accuracy of the blockmodel interaction matrix B and run-time. This extensive comparative evaluation provides a practical guideline for choosing the proper inference procedure in a real setting, especially when analyzing large tables. In addition, we quantify the effect of censoring on the inference in terms of estimation error.

Variational EM $\left(Y(j,k)_{j=1,k=1}^{N_1,N_2}, \alpha, \beta, \sigma^2 \right)$

initialize $\vec{\phi}_{j\to k} := 1/K_1$ for all j and k
initialize $\vec{\eta}_{j\leftarrow k} := 1/K_2$ for all j and k
initialize $\vec{\nu}_j := N_2/K_1 + \alpha$ for all j
initialize $\vec{\xi}_k := N_1/K_2 + \beta$ for all k
initialize $B(g,h)$ for all g and h as the data mean plus a random noise
repeat
    E step: update $\vec{\phi}_{j\to k}$ for all j and k using equation (2.7) and normalize to sum to 1
        update $\vec{\eta}_{j\leftarrow k}$ for all j and k using equation (2.8) and normalize to sum to 1
        update $\vec{\nu}_j$ for all j using equation (2.9)
        update $\vec{\xi}_k$ for all k using equation (2.10)
    M step: update $B(g,h)$ for all g,h using equation (2.11)
until convergence;
return $(\vec{\phi}, \vec{\eta}, \vec{\nu}, \vec{\xi}, B)$

Algorithm 1: The variational EM algorithm. The E steps 6–9 are also repeated until convergence to achieve the most stabilized mutual updates for the set of free parameters $\vec{\phi}$, $\vec{\eta}$, $\vec{\nu}$, $\vec{\xi}$.

3.1. *Design of experiments.* In the past decade, variational EM (vEM) has become a practical alternative to MCMC when dealing with large data sets, despite its lack of theoretical guarantees [Jordan et al. (1999), Airoldi (2007), Joutard et al. (2008)]. The relative merits between vEM and MCMC have been established empirically for a number of models [e.g., see Blei and Jordan (2006), Braun and McAuliffe (2010)]. We designed simulations with the goal of exploring the trade-off between estimation accuracy and computational burden that vEM helps manage in the context of estimation and posterior inference with the proposed model.

Briefly, vEM is an optimization approach, no sampling is involved, which requires key choices about the following: (1) error tolerance for both the approximate E step and the M step, and (2) how to design multiple initializations and how many to use. MCMC is a sampling approach, which requires key choices about the following: (1) convergence criteria, (2) burn-in, (3) thinning to reduce autocorrelation, and (4) multiple chains. For the variational EM approach, we set the overall error tolerance at 1e–5, the maximum number of iterations for the variational E steps at 10, and 10 random initializations. For the MCMC approach, we investigated the convergence using Gelman–Rubin and Raftery–Lewis for the median, autocorrelation using trace plots and partial autocorrelation functions. Based on these studies, we chose to use 1000 iterations for burn-in, 6000 iterations and a 10 to 1 thinning ratio, which results in 500 draws for each chain, and we used 10 chains. For both approaches, we use the true Dirichlet parameters α, β and the true variance $\sigma^2 = 0.01$. Overall, this seems a fair comparison.

The data are generated using the procedures described in Section 2.1 with the following specifications. The $B(g,h)$ follows a Normal distribution $B(g,h) \sim \text{Normal}(\mu_B(g,h), \sigma_B^2(g,h))$, where $\mu_B(g,h) = 0$ and $\sigma_B^2(g,h) = 1$. Three sets of block sizes are considered: $(K_1, K_2) = (2,3)$, $(4,6)$ and $(6,9)$. The corresponding table sizes are $(N_1, N_2) = (10,15)$, $(50,75)$ and $(100,150)$, respectively. The Dirichlet parameters are set to be $\alpha = \beta = 0.2$ or $\alpha = \beta = 0.05$. In all the experiments, we set $\sigma^2 = 0.01$.

3.2. *Mixed membership estimation.* Here we evaluate the competing estimation procedures on recovering mixed membership vectors. We report results on the accuracy of the first and second largest membership components. It is well known that mixture models and mixed membership models suffer from identifiability issues, that is, their likelihood is uniquely specified up to permutations of the labels [Titterington, Smith and Makov (1985)]. We evaluate the performance for a fixed permutation, obtained empirically by sorting the membership vectors for the vEM and by using a standard Procrustes transform for the MCMC [Stephens (2000)]. We note that vEM converged quickly to a (local) optimum, thus involving a considerably more mitigated label switching issue than the collapsed Gibbs sampler. This is an advantage, especially given that the empirical vEM estimation error reported in Table 1 is comparable to that of the more principled MCMC sampler.

To quantify accuracy, we identify the locations of the largest two components in the estimated vector of probabilities, $\vec{\pi}_j$, and take those to be the first and second choice of group memberships for the jth row. These assignments are compared, via zero-one loss, with the true memberships: if there is a match, we note the accuracy as 1, otherwise 0. The recorded row accuracy is the average over all the rows and the ten experiments. The column accuracy is defined in a similar fashion.

The results for the estimated first and second memberships are summarized in Table 1. The results for the first membership suggest that estimation is well behaved in the proposed model; the true membership can be recovered with a fairly high successful rate under different experimental settings. As expected, the estimation accuracy decreases with the increase on the block size. The lowest pair reported in the table are 0.485 and 0.357 for $K_1 = 6$ and $K_2 = 9$, still much better than random assignments where the accuracy would be $1/6$ and $1/9$, respectively. For the second membership, we only consider elements with an estimated second membership probability greater than a threshold. In this study, the thresholds are $\frac{1}{10K_1}$ and $\frac{1}{10K_2}$ for row and column memberships, respectively. It is clear that the variational Bayes approach performs much better than MCMC in estimating the second membership. One explanation can be that the second membership is more ambiguous than the first membership, requiring a large number of iterations for MCMC to converge.

Another factor that affects model performances is the Dirichlet parameters α and β. Judging from the table, the accuracy when $\alpha = \beta = 0.05$ is generally higher than those of $\alpha = \beta = 0.2$. This result is reasonable since a smaller α and β value

TABLE 1

Comparisons on row and column estimation accuracy of estimates for the first highest membership (regular font) and second highest membership (italic font) obtained with variational EM and MCMC. Standard errors are quoted inside parenthesis

		$K_1 = 2$ and $K_2 = 3$		$K_1 = 4$ and $K_2 = 6$		$K_1 = 6$ and $K_2 = 9$	
(N_1, N_2)	α/β	**Row**	**Column**	**Row**	**Column**	**Row**	**Column**
vEM							
(10, 15)	0.2	0.970 (0.067)	0.667 (0.031)	0.620 (0.063)	0.587 (0.069)	0.470 (0.048)	0.520 (0.076)
		0.970 (0.067)	*0.522 (0.075)*	*0.233 (0.152)*	*0.179 (0.077)*	*0.210 (0.129)*	*0.060 (0.058)*
	0.05	0.980 (0.042)	0.967 (0.085)	0.870 (0.125)	0.807 (0.066)	0.780 (0.063)	0.567 (0.085)
		0.980 (0.042)	*0.533 (0.233)*	*0.233 (0.179)*	*0.190 (0.110)*	*0.317 (0.123)*	*0.133 (0.112)*
(50, 75)	0.2	0.784 (0.122)	0.751 (0.146)	0.680 (0.034)	0.471 (0.039)	0.426 (0.053)	0.416 (0.041)
		0.784 (0.122)	*0.694 (0.130)*	*0.304 (0.097)*	*0.175 (0.033)*	*0.194 (0.054)*	*0.136 (0.031)*
	0.05	0.980 (0.000)	0.849 (0.074)	0.620 (0.104)	0.575 (0.058)	0.634 (0.046)	0.483 (0.053)
		0.980 (0.000)	*0.662 (0.132)*	*0.239 (0.118)*	*0.216 (0.058)*	*0.210 (0.073)*	*0.149 (0.047)*
(100, 150)	0.2	0.960 (0.000)	0.823 (0.106)	0.601 (0.077)	0.670 (0.076)	0.485 (0.048)	0.357 (0.029)
		0.960 (0.000)	*0.612 (0.247)*	*0.261 (0.055)*	*0.237 (0.063)*	*0.194 (0.029)*	*0.137 (0.022)*
	0.05	0.946 (0.092)	0.743 (0.132)	0.769 (0.055)	0.707 (0.057)	0.553 (0.084)	0.479 (0.052)
		0.946 (0.092)	*0.520 (0.227)*	*0.361 (0.057)*	*0.236 (0.064)*	*0.217 (0.060)*	*0.135 (0.028)*
MCMC							
(10, 15)	0.2	0.922 (0.148)	0.730 (0.102)	0.678 (0.015)	0.665 (0.012)	0.669 (0.008)	0.521 (0.007)
		0.922 (0.148)	*0.504 (0.167)*	*0.306 (0.053)*	*0.204 (0.031)*	*0.207 (0.011)*	*0.157 (0.004)*
	0.05	0.841 (0.121)	0.901 (0.120)	1.000 (0.000)	0.878 (0.031)	0.884 (0.005)	0.825 (0.007)
		0.841 (0.121)	*0.409 (0.138)*	*0.520 (0.122)*	*0.413 (0.091)*	*0.227 (0.052)*	*0.161 (0.022)*
(50, 75)	0.2	0.871 (0.121)	0.671 (0.097)	0.711 (0.095)	0.659 (0.084)	0.682 (0.106)	0.562 (0.051)
		0.871 (0.121)	*0.437 (0.186)*	*0.380 (0.039)*	*0.300 (0.065)*	*0.301 (0.093)*	*0.231 (0.026)*
	0.05	0.994 (0.013)	0.676 (0.113)	0.775 (0.176)	0.753 (0.129)	0.824 (0.088)	0.839 (0.054)
		0.994 (0.013)	*0.452 (0.131)*	*0.383 (0.135)*	*0.319 (0.142)*	*0.357 (0.090)*	*0.365 (0.074)*
(100, 150)	0.2	0.971 (0.032)	0.653 (0.150)	0.682 (0.119)	0.633 (0.083)	0.735 (0.069)	0.614 (0.078)
		0.968 (0.034)	*0.420 (0.223)*	*0.332 (0.080)*	*0.255 (0.054)*	*0.310 (0.074)*	*0.235 (0.059)*
	0.05	0.830 (0.208)	0.773 (0.138)	0.810 (0.140)	0.772 (0.127)	0.780 (0.046)	0.750 (0.064)
		0.829 (0.208)	*0.463 (0.203)*	*0.354 (0.151)*	*0.277 (0.088)*	*0.285 (0.053)*	*0.249 (0.046)*

corresponds to a higher likelihood of a dominating component, which is easier to identify than more ambiguous memberships.

The membership accuracy computed through variational Bayes aligns with those calculated from MCMC, and even slightly better when the block size is small. Since variational inference is typically much more efficient than MCMC, the former method is preferred for practical analysis, especially for high-dimensional cases. We will present run-time comparisons between these two approaches in the next section.

3.3. *Predictions out-of-sample.* Prediction power is a useful criterion for evaluating statistical models. When some data are missing, is the model sufficiently flexible to provide correct inferences and to predict the missing values with high accuracy? To answer this question, we randomly select $2/3$ of rows and $2/3$ of columns from the table, whose intersections are $4/9$ of the entries. We set half of them (i.e., $2/9$) as missing (to avoid eliminating an entire row or column), and run the model on the remaining $7/9$ entries. The first membership prediction accuracy is reported in Table 2. They are slightly lower than those estimated without missing values, but overall much better than the baseline probabilities $1/K_1$ and $1/K_2$. Furthermore, the prediction accuracy achieved by variational Bayes is comparable or better than those obtained by MCMC. This result reinforces our belief that variational Bayes is a good inference approach for the proposed blockmodel.

3.4. *Blockmodel matrix estimation.* Here we compare the variational Bayes and MCMC in terms of estimating the matrix B. The estimation error ε_B is defined as the 1-norm of the matrix $|B - \hat{B}|$, where $\hat{B}$ is the estimated matrix. The result for $K_1 = 2$, $K_2 = 3$ is shown in Table 3. Except for the case of $\alpha = \beta = 0.05$ and $N_1 = 10$, $N_2 = 15$, variational Bayes performs close to or better than MCMC. The true B in this simulation study is

$$\begin{pmatrix} -0.5009 & 0.0687 & 1.5887 \\ 0.4148 & -0.8086 & -1.3112 \end{pmatrix}.$$

3.5. *Sensitivity to initialization and priors specifications.* Here we analyzed the sensitivity of the inference to informative versus noninformative prior specifications, and to uniform versus random initialization of some constants in our model. The results show no significant sensitivity of the estimation error to these choices. This evidence supports our claim that inference is well behaved and that identifiability is not an issue for the model we proposed, in practice, in the data regimes we considered.

In Algorithm 1 (vEM) and the supplement (MCMC), we initialized a subset of parameters ($\pi, \eta, \nu, \varepsilon$ in vEM and D, E in MCMC) uniformly. To assess the sensitivity of inference to this initialization strategy, we tested alternative versions of these algorithms in which we initialized these parameters at random, on the

TABLE 2

Comparisons on row and column estimation accuracy between variational EM and MCMC, when 2/9 of the entries are missing. Standard errors are inside the parenthesis

		$K_1 = 2$ and $K_2 = 3$		$K_1 = 4$ and $K_2 = 6$		$K_1 = 6$ and $K_2 = 9$	
(N_1, N_2)	α/β	Row	Column	Row	Column	Row	Column
vEM							
(10, 15)	0.2	0.780 (0.148)	0.600 (0.094)	0.610 (0.110)	0.507 (0.118)	0.520 (0.063)	0.547 (0.103)
	0.05	0.900 (0.067)	0.853 (0.117)	0.730 (0.125)	0.547 (0.108)	0.700 (0.094)	0.613 (0.042)
(50, 75)	0.2	0.664 (0.067)	0.615 (0.134)	0.452 (0.081)	0.383 (0.039)	0.366 (0.034)	0.335 (0.037)
	0.05	0.930 (0.034)	0.843 (0.077)	0.570 (0.135)	0.564 (0.074)	0.504 (0.076)	0.444 (0.040)
(100, 150)	0.2	0.786 (0.091)	0.672 (0.128)	0.472 (0.124)	0.362 (0.055)	0.313 (0.049)	0.326 (0.059)
	0.05	0.751 (0.194)	0.749 (0.136)	0.656 (0.102)	0.503 (0.091)	0.397 (0.068)	0.373 (0.056)
MCMC							
(10, 15)	0.2	0.703 (0.100)	0.617 (0.083)	0.480 (0.091)	0.460 (0.085)	0.406 (0.012)	0.368 (0.054)
	0.05	0.770 (0.145)	0.726 (0.115)	0.540 (0.161)	0.446 (0.073)	0.454 (0.101)	0.456 (0.070)
(50, 75)	0.2	0.788 (0.145)	0.645 (0.104)	0.544 (0.064)	0.443 (0.032)	0.357 (0.062)	0.343 (0.039)
	0.05	0.809 (0.194)	0.647 (0.098)	0.606 (0.072)	0.567 (0.068)	0.473 (0.054)	0.479 (0.074)
(100, 150)	0.2	0.813 (0.103)	0.576 (0.102)	0.575 (0.061)	0.492 (0.042)	0.411 (0.048)	0.395 (0.028)
	0.05	0.867 (0.150)	0.834 (0.111)	0.639 (0.051)	0.524 (0.040)	0.514 (0.092)	0.497 (0.030)

TABLE 3
Comparisons on ε_B as the estimation error of B between variational Bayes and MCMC

(N_1, N_2)	(10,15)		(50,75)		(100,150)	
α/β	0.05	0.2	0.05	0.2	0.05	0.2
VB	0.152 (0.042)	0.022 (0.022)	0.048 (0.024)	0.061 (0.061)	0.053 (0.029)	0.002 (0.001)
MCMC	0.019 (0.006)	0.027 (0.047)	0.110 (0.058)	0.045 (0.066)	0.134 (0.065)	0.105 (0.058)

data set analyzed in Section 3.7. Briefly, in vEM, we initialized each $\vec{\phi}_{j\to k}$ and $\vec{\eta}_{j\leftarrow k}$ with random membership vectors, then initialized $\vec{\nu}_j$, $\vec{\xi}_k$ using equations (2.9) and (2.10). The blockmodel B is initialized as in Algorithm 1. In MCMC, we initialized each $\vec{D}_{j\to k}$, $\vec{E}_{j\leftarrow k}$ with a membership with a single positive entry assigned at random, we computed $\vec{D}_{j\to\cdot}$, $\vec{E}_{\cdot\leftarrow k}$, Y_{gh}, n_{gh} accordingly from these initial values of $\vec{D}$ and $\vec{E}$, then initialized $p(D_{j\to k,g}=1, E_{j\leftarrow k,h}=1)$ as detailed in the supplement [Airoldi, Wang and Lin (2013b)]. The results of this experiment are shown in Table 4.

Another input for Algorithm 1 and the MCMC inference algorithm is the Dirichlet parameters α and β. A priori, $\alpha, \beta < 1$ favor a single dominating membership component while $\alpha, \beta > 1$ favor diffuse membership. In the analysis of real data, we expect few dominating memberships, so we typically set $\alpha = \beta$ equal to either 0.2 or 0.05 and assess sensitivity of resulting estimated memberships and other parameters. However, the question arises as to whether an alternative strategy that features informative priors is more useful than using noninformative as we do. Using informative priors for the membership parameters might lead to improved inference, especially in the case of substantial nonidentifiability.

To evaluate this issue, we generated a data set with informative priors $\vec{\alpha} = (0.3, 0.7)'$ for the rows and $\vec{\beta} = (0.6, 0.3, 0.10)'$ for the columns. Then we fit the model with the vEM algorithm on this data set using both noninformative uniform priors ($\alpha = \beta = 0.05$) and informative priors with the vectors $\vec{\alpha}$, $\vec{\beta}$ set at the true values. The results are presented in Table 5 from which we see that the results are

TABLE 4
Comparisons on ε_B as the estimation error of B and the first highest membership accuracy between different initialization for variational Bayes and MCMC

	vEM			MCMC		
Init.	ε_B	Row	Column	ε_B	Row	Column
Random	0.200 (0.163)	0.916 (0.184)	0.907 (0.120)	0.171 (0.179)	0.872 (0.171)	0.818 (0.177)
Uniform	0.205 (0.173)	0.916 (0.117)	0.880 (0.102)	0.115 (0.175)	0.957 (0.047)	0.820 (0.157)

TABLE 5
Comparison of vEM fits using informative and noninformative priors, in terms of estimation error ε_B and accuracy in estimating the highest membership component

Noninformative priors			Informative priors		
ε_B	Row	Column	ε_B	Row	Column
0.385 (0.176)	0.868 (0.121)	0.827 (0.134)	0.203 (0.121)	0.788 (0.157)	0.870 (0.091)

comparable. This justifies the simple choice of noninformative prior in our algorithms.

3.6. *Run-time comparison*. As seen previously, variational Bayes performs as effectively as MCMC in parameter and membership estimation as well as held-out prediction accuracy. In the following, we present results on run-time comparison between these two approaches. Our goal is to quantify the magnitude of savings that variational Bayes can achieve while obtaining similar inferences to those obtained through MCMC.

For each experiment we run 10 times, and the average log run-time is recorded. The plots are shown in Figure 1. Three table sizes are considered in this simulation: 10×15, 50×75 and 100×150. From this figure, the run-time for MCMC is consistently several times larger more than that of variational Bayes. For example, when block sizes equal $(6, 9)$, and Dirichlet parameters equal 0.05, one experiment takes about 30 minutes to run for variational Bayes, and it takes roughly 6 hours for MCMC. This trend continues when table size increases, and the saving on computational cost can be much more. These results suggest that variational Bayes should be preferred for analyzing large tables. Recently developed inference strategies based on spectral clustering [Rohe and Yu (2012)] and binary factor graphs [Azari and Airoldi (2012)] should also be considered.

3.7. *Quantifying the effects of censoring*. One of the issues in existing studies of coordinated cellular responses is the preprocessing of the original measurements. This kind of censoring reduces data utility and decreases estimation accuracy. The goal of this study is to quantify the effects of censoring by thresholding on the estimation of the blockmodel.

The data Y are generated from $Y(j,k) \sim \text{Normal}(\vec{\pi}_j' B \vec{p}_k, \sigma^2)$. The domain of $Y(j,k)$ is $(-\infty, +\infty)$. We perform Inverse Fisher Transformation (IFT) that maps $Y(j,k)$ to $\rho(j,k)$ so that its range is $[-1, 1]$. The censored data are defined as $S(j,k) = \mathbf{1}(|\rho(j,k)| \geq \tau)$, where τ can be median, mean or 0.5. Clearly, $S(j,k) \in \{0, 1\}$.

The Normal blockmodel is applied to the original data $Y(j,k)$ and the Bernoulli blockmodel described in Section 2.2 is applied to the censored data $S(j,k)$.

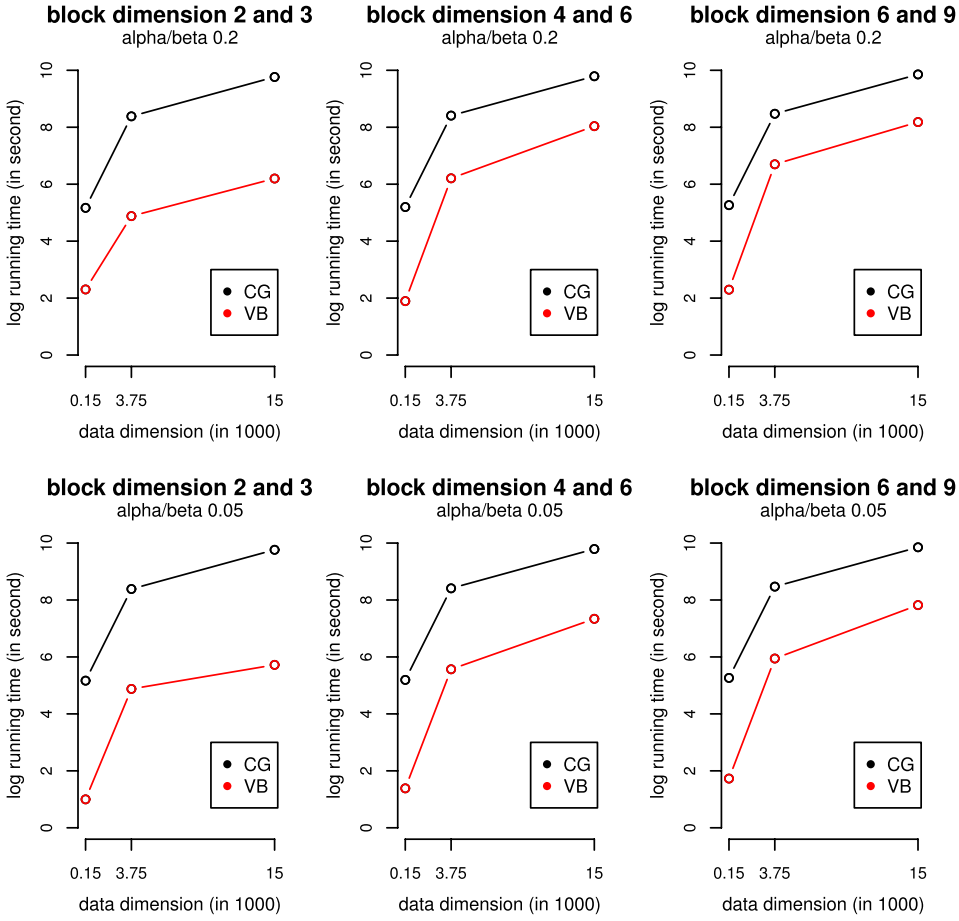

FIG. 1. *Log run-time for simulated data. Red lines represent variational Bayes and black lines represent MCMC via collapsed Gibbs. The x-axis is the number of elements in a table. For instance,* 0.15 (*thousand*) *represents a* 10 *by* 15 *table with* 150 *elements.*

To make the comparison in the same scale, we define $\hat{\rho}(j,k)$ as the IFT of $\vec{\phi}'_{j\to k}\hat{B}\vec{\eta}_{j\leftarrow k}$, where $\vec{\phi}_{j\to k}$, $\hat{B}$ and $\vec{\eta}_{j\leftarrow k}$ are estimated from the Normal block-model. The estimation error is defined as

$$\varepsilon = \frac{\sum_{(j,k)} |\rho(j,k) - \hat{\rho}(j,k)|}{N_1 \times N_2}.$$

The estimation error for the censored experiment is computed in the same fashion, with $\hat{\rho}(j,k) = \vec{\phi}'_{j\to k}\hat{B}\vec{\eta}_{j\leftarrow k}$, where $\vec{\phi}_{j\to k}$, $\hat{B}$ and $\vec{\eta}_{j\leftarrow k}$ are estimated from the Bernoulli blockmodel, and $\rho(j,k)$ replaced by $|\rho(j,k)|$.

We compare our model with a bi-clustering method popular in computational biology [Cheng and Church (2000)], fit to both the raw and censored correlations.

TABLE 6
Comparison of estimation error on censored and noncensored data. Standard errors are inside the parenthesis

Data	Method	Bic.	Error ε	Recall	Precision
Raw ρ_{ij}	2-way Normal	6	0.054 (0.010)	0.841 (0.169)	0.881 (0.116)
	Hier. clustering	6	0.221 (–)	0.967 (–)	0.970 (–)
	Cheng & Church	2	–	0.367 (–)	0.679 (–)
$\lvert\rho_{ij}\rvert > \rho_{(0.5)}$	2-way Bernoulli	6	0.175 (0.006)	0.518 (0.017)	0.722 (0.048)
	Hier. clustering	6	0.125 (–)	0.700 (–)	0.850 (–)
	Cheng & Church	2	–	0.232 (–)	0.640 (–)
	Cross-associations	4	–	0.667 (–)	0.762 (–)
$\lvert\rho_{ij}\rvert > \bar{\rho}$	2-way Bernoulli	6	0.182 (0.003)	0.528 (0.014)	0.773 (0.056)
	Hier. clustering	6	0.187 (–)	0.500 (–)	0.841 (–)
	Cheng & Church	3	–	0.237 (–)	0.640 (–)
	Cross-associations	6	–	0.667 (–)	0.841 (–)
$\lvert\rho_{ij}\rvert > 0.5$	2-way Bernoulli	6	0.158 (0.002)	0.528 (0.022)	0.835 (0.030)
	Hier. clustering	6	0.189 (–)	0.500 (–)	0.841 (–)
	Cheng & Church	3	–	0.239 (–)	0.640 (–)
	Cross-associations	8	–	0.613 (–)	0.667 (–)

We match each estimated bicluster to a true block and compute recall and precision in estimating absolute correlations above a threshold. Results are presented in Table 6, where the results obtained with BCCC are optimized over a range of input parameter values. For completeness, we also add results obtained with hierarchical clustering to rows and columns independently, and with cross-association [Chakrabarti et al. (2004)].

The effects of censoring are clearly seen from Table 6. The estimation error increases more than threefold when using the censored data with the Bernoulli block model. The effect of thresholding parameter τ is not very significant.

4. Analyzing transcriptional and metabolic coordination in response to starvation. Functions in a cell are executed by cascades of molecular events. Intuitively, proteins are the messengers, while metabolites and other small molecules are the messages. Measuring protein activity over time, directly, is difficult and expensive. An indication of the abundance of most proteins, however, can be inferred from the amount of the messenger RNA transcripts. These transcripts are copies of genes and lead to the translation of proteins. This is especially true in yeast where alternatives to the transcription-translation hypothesis, such as alternative splicing, are not frequent. Metabolite concentrations add an essential perspective to the study of cascades of molecular events.

We conducted an integrated analysis of two data collections recently published: temporal profiles of metabolite concentrations [Brauer et al. (2006)] and tempo-

ral profiles of gene expression [Bradley et al. (2009)], both measured in *Saccharomyces cerevisiae* with matching sampling schemes.

An integrated analysis of the coordination between gene expression and metabolite concentrations may lead to the identification of sets of genes (i.e., the corresponding proteins) and metabolites that are functionally related, which will provide additional insights into regulatory mechanisms at multiple levels and open avenues of inquiry. The identification and quantification of such coordinated regulatory behavior is the goal of our analysis.

The methodology in Section 2 allows us to identify genes and metabolites that show correlated responses to metabolic stress, namely, starvation. To evaluate the biological significance of the results, we quantify to what extent correlated responses are associated with metabolic-related functions and to what extent estimated models can be used to identify functionally related genes and metabolites out-of-sample.

4.1. *Data and experimental design.* The expression data consist of messenger RNA transcript levels measured using Agilent microarrays on cultures of *S. cerevisiae* before and after carbon starvation (glucose removal), and before and after nitrogen starvation (ammonium removal). Collection times were 0 minutes (before starvation) and 10, 30, 60, 120, 240 and 480 minutes after starvation. For more details about the data and the experimental protocol see Bradley et al. (2009). The metabolite concentrations data were obtained using liquid chromatography-mass spectrometry before and after carbon starvation (glucose removal), and before and after nitrogen starvation (ammonium removal). Collection times were 0 minutes (before starvation) and 10, 30, 60, 120, 240 and 480 minutes after starvation. For more details about the data and the experimental protocol see Brauer et al. (2006).

The concentration of each metabolite and the transcript level of each gene at time point t are expressed as $\log_2$ ratios versus the corresponding measurements at the zero time point. Thus, for each gene j we have a sequence G_{jt}, $t = 1, \ldots, 6$, and for each metabolite k we have a sequence M_{kt}, $t = 1, \ldots, 6$, representing for the 6 time points observation after time 0. Complete temporal profiles are available for 5039 genes and for 61 metabolites; 783 genes and 7 metabolites with missing data were not considered.

Using the temporal profile, we can calculate the sample correlation coefficient of each gene and metabolite pair (j,k): $\rho(j,k) = \frac{\sum_{t=1}^{T}(G_{jt}-\bar{G}_j)(M_{kt}-\bar{M}_k)}{(T-1)S_G S_M}$, where $\bar{G}_j = \sum_t G_{jt}/T$ and $\bar{M}_k = \sum_t M_{kt}/T$ are the sample mean, and S_G and S_M are the sample standard deviation. We then transform these correlations using the Fisher transformation $Z(j,k) = \frac{1}{2}\log\frac{1+\rho(j,k)}{1-\rho(j,k)}$. With the true correlation between genes and metabolites denoted as ρ_0, we have $Z(j,k)$ following asymptotically Normal distribution with mean $\mu_z = \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}$ and standard error $1/\sqrt{T-3}$ [Fisher (1915, 1921)]. Under the hypothesis that there is no correlation between

genes and metabolites, we will expect $\rho_0 = 0$ and $\mu_z = \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0} = 0$. These Fisher transformed quantities provide the input to our model.

We do not expect the multi-way blockmodel assumption to hold for all the 5039 genes. Instead, we provide separate joint analyses on subsets of genes and all 61 metabolites, for a number of gene lists of interest, which we expect to be involved in the cellular response to starvation. We consider gene lists that were obtained in studies exploring the environmental stress response (ESR), cellular proliferation, metabolism and the cell cycle [Gasch et al. (2000), Tu et al. (2005), Brauer et al. (2008), Airoldi et al. (2009, 2013a), Slavov et al. (2013)].

For all the experiments we rely on variational Bayes implementation of our model due to its advantage in convergence speed, which is crucial when dealing with correlation tables involving hundreds of genes. We adopt the setting as described in Section 3.6 for VB, with specific changes described as they become relevant. In the remainder of this section, with the exception of Section 4.5, we consistently set the number of metabolite blocks $K_2 = 4$, since there are four metabolite classes, and we use informative priors for the memberships of each metabolite, depending on which class they are known to belong to. Specifically, each of the 61 metabolites belongs to one of the four classes: TCA, AA, GLY, BSI. If a metabolite is in class TCA, say, Aconitate, its $\vec{\xi}$ vector will be initialized as $\vec{\xi} = [100\ 1\ 1\ 1]$, normalized to unit norm. By assuming a dominating component on the true index in the initial membership, the metabolites will mostly remain stably associated to their classes during VB inference.

For the optimal number of gene blocks, we select K_1 by minimizing the Bayesian information criterion (BIC). The general BIC formula is $-2\log L + k \times \log(n)$, in which k is the number of parameters, n is the number of observations, and L is the likelihood. For our model, the approximated BIC is

$$-2\log L + |B, \vec{\pi}, \vec{p}| \times \log|Y|,$$

where $|B, \vec{\pi}, \vec{p}|$ is the number of parameters, which is approximately equal to $K_1 \times K_2$, and $|Y|$ is the number of entries in the table, that is, $|Y| = N_1 \times N_2$.

In Section 4 we present some of the results with the goal of showcasing how the data analysis, via the multi-way block model, supports the biological research.

4.2. *Multifaceted functional evaluation of coordinated responses.* Here we evaluate to what extent the proposed model is useful in revealing the genes' multifaceted functional roles. We rely on the functional enrichment analysis using the Gene Ontology to evaluate the functional content of clusters of genes [Ashburner et al. (2000), Boyle et al. (2004)].

One aspect of our model that distinguishes it from clustering and bi-clustering methods is the mixed membership assumption. That is, in our model, each gene can participate in multiple functions, as modeled via the gene-specific latent membership vectors $\vec{\pi}$. In practice, the membership assumption lets us identify multiple levels of functional enrichment.

To illustrate this point, we consider 521 genes that were found to be strongly associated with metabolic activities, that is, up-regulated in response to increasing growth rate, in previous studies [Brauer et al. (2008), Airoldi et al. (2009)]. We use the largest estimated memberships for each gene π_{gi}, $i = 1, \ldots, K_1$, to assign genes $g = 1, \ldots, 521$ to metabolite classes $j = 1, \ldots, 4$. Then we perform functional analysis on the resulting sets of genes associated with each metabolite class.

More formally, we proceed as follows. First, the largest estimated membership is used to assign gene g to gene block i, according to $\hat{i}_g = \arg\max_{i=1,\ldots,K_1} \pi_{gi}$. Then the largest estimated gene-block to metabolite-class association $|B_{ij}|$ is then used to assign gene g with a metabolite class, according to

$$\hat{j}_g = \arg\max_{j=1,\ldots,4} |B_{\hat{i}_g, j}|.$$

The collection of estimated gene-to-metabolite class associations, $\{\hat{j}_g,\ g = 1, \ldots, 521\}$, is used to partition genes into four sets, for example, $AA = \{g \text{ s.t. } \hat{j}_g = 1\}$. We perform functional enrichment analysis for each of these four sets. In addition, the mixed membership nature of the proposed multi-way block-model allows us to analyze second-order functional enrichment. We repeat the procedure above but we estimate $\hat{i}_g$ using the second-largest membership in $\vec{\pi}_g$.

The functional analysis results obtained for both first- and second-largest memberships are reported in Table 7. Interestingly, subsets of genes associated with

TABLE 7

Example functional evaluation. Gene Ontology terms associated with first- and second-largest membership scores for the Nitrogen starvation experiment

Memb.	**Class**	**Ontology**	**Term description**	***p*-value**
First	AA	Component	DNA-directed RNA polymerase I complex	9.8E–6
First	AA	Component	Preribosome, small subunit precursor	0.00324
First	AA	Function	Translation factor activity, nucleic acid binding	5.11E–14
First	AA	Function	Translation initiation factor activity	1.64E–10
First	AA	Function	DNA-directed RNA polymerase activity	1.45E–6
First	BSI	Component	Preribosome, large subunit precursor	0.00013
First	BSI	Function	GTP binding	0.03314
First	BSI	Function	Guanyl ribonucleotide binding	0.03314
Second	AA	Component	DNA-directed RNA polymerase III complex	6.76E–10
Second	AA	Component	DNA-directed RNA polymerase II core complex	0.00322
Second	AA	Function	RNA polymerase activity	1.27E–6
Second	AA	Function	ATP-dependent RNA helicase activity	3.43E–6
Second	BSI	Component	Ribonucleoprotein complex	6.47E–12
Second	BSI	Component	90S preribosome	0.00124
Second	BSI	Function	Aminoacyl-tRNA ligase activity	0.0000817
Second	BSI	Function	N-methyltransferase activity	0.0455

the same metabolite class, for instance, AA, are functionally enriched for multiple functions, to different degrees. For instance, genes use AA metabolites when performing translational activities in the nucleus primarily, however, they use AA metabolites when performing polymerase-related activities on the polymerase II and II complexes to a lesser extent. Similarly, genes use BSI metabolites for binding activities in the preribosome primarily, and for ligase and transferase activities in the preribosome and the ribonucleoprotein complex to a minor extent. The magnitude of the components of the relevant mixed membership vectors provides more information on the degree of involvement the various gene blocks in these many activities. This type of multifaceted functional analysis is possible thanks to the mixed membership assumption encoded in the multi-way blockmodel.

These results highlight the role of the mixed membership assumption in supporting a detailed multifaceted functional analysis, which is not possible with traditional methods.

4.3. *Predicting functional annotations out-of-sample*. Here we assess the goodness of fit of the proposed method on real data, in terms of predictions out-of-sample. We present results of an experiment in which we predict held-out functional annotations π_{gi}. This analysis leverages use of informative priors on a subset of known functional annotations.

We consider 57 genes that were found in previous studies to be strongly associated with cellular growth [Airoldi et al. (2009), Brauer et al. (2008)], 760 genes that were found to be involved in the environmental response to stress [Gasch et al. (2000)], and 19 genes that were found to be involved in metabolic cycling [Tu et al. (2005)].

Good out-of-sample prediction performance will enable biologists to use this method to guide which functions they should be testing at the bench, speeding up the exploration of the functional landscape through statistical analysis of gene–metabolite associations.

To establish the ground truth for this experiment, we collected functional annotations for each gene in the same four lists as in Section 4.2 which will be held-out and predicted using the multi-way blockmodel. Table 8 reports summary statistics

TABLE 8

Statistics for the lists of genes. Column three reports the number of genes with one, some and no functional annotations. K_1 is the number of gene blocks in the fitted blockmodel

	No. of genes		**No. of functional annotations**						
Gene list	**Total**	**One/some/none**	**Min**	**25%**	**50%**	**75%**	**Max**	**Mean**	K_1
Growth rate	57	5/19/38	1	1.25	4	7	7	4.26	12
ESR induced	240	0/215/25	2	5	7	12	31	9.31	76
ESR repressed	520	1/503/17	1	10	19	22	31	16.93	78
Metabolic cycle	19	4/14/5	1	1	6.5	10	20	7.29	25

of the functional annotations in each list of genes, obtained using the Gene Ontology term finder (SGD). Column two reports the total number of genes in each list. Column three reports the number of genes with one, some and no functional annotations. Columns 4–9 report the quantiles from the distribution of the number of functional annotations for the genes in each list. Column 10 reports the value of K_1 we selected for fitting the blockmodel.

To perform the second experiment, we held out the annotations for 50% of the genes with multiple functional annotations, and we also held out the annotation for 50% of the genes with a single functional annotation. When fitting the multi-way blockmodel, in addition to using informative priors for the memberships of each metabolite depending on which class they are known to belong, as detailed in Section 4.1, we used informative priors for the functional annotations we *did not* hold out. For the held-out annotations, we used noninformative values for the hyperparameters instead. For instance, suppose that the known vector of functional annotations for gene g is $\vec{a}_g = [1\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1]$, and that $a_g(1)$ and $a_g(4)$ were to be held out in a particular replication, so that we have $\vec{a}_g = [\text{NA}\ 0\ 0\ \text{NA}\ 1\ 0\ 0\ 0\ 0\ 1]$. The prior for the functional annotation for that gene would be set at $\vec{\xi}_g = [1\ 1\ 1\ 1\ 100\ 1\ 1\ 1\ 1\ 100]$, normalized to unit norm. The rationale for this choice is to fit a multi-way blockmodel with known biological structure for those genes and metabolites that are used for parameter estimation, but agnostic about the biology we want to predict out-of-sample. We claimed success in each prediction if the imputed annotations, $\hat{\xi}_g(k) = 1$, corresponded to real held-out annotations, and if the imputed absences of annotations, $\hat{\xi}_g(i) = 0$, corresponded to absences of real held-out annotations. We repeated this procedure 10 times, for each of the four lists of genes.

Table 9 reports the accuracy results, detailed by genes with single and multiple annotations, and evaluated separately for annotations (i.e., the 1s) and lack of annotations (i.e., the 0s). The baseline accuracy for predicting single annotations, using random guesses for each gene independently, ranges between $1/19 \approx 5\%$ for

TABLE 9

Out-of-sample predictions of functional annotations for the Nitrogen starvation experiment. Accuracy in recovering (single/multiple) annotations for four lists of genes

	Single annotations				**Multiple annotations**			
	Observed		**Missing**		**Observed**		**Missing**	
Gene list	**0s**	**1s**	**0s**	**1s**	**0s**	**1s**	**0s**	**1s**
Growth rate	0.94	0.36	0.94	0.36	0.83	0.75	0.84	0.74
ESR induced	–	–	–	–	0.92	0.39	0.92	0.46
ESR repressed	–	–	0.99	0.00	0.84	0.43	0.85	0.50
Metabolic cycle	0.97	0.25	0.96	0.10	0.78	0.65	0.80	0.63

the metabolic cycle genes to $1/520 \approx 0.2\%$ for the ERS genes. The baseline accuracy is slightly higher for predicting multiple annotations, since predicting single annotations is a harder problem.

For completeness, we also report the accuracy in predicting annotations that were known during model fitting to get a sense of the goodness of fit from a substantive, biological perspective. In fact, if the model assumptions are accurate, we would expect accurate predictions for the known annotations. If the model is not accurate, or if the model provides too much shrinkage, we would expect lower accuracy on known annotations.

Overall, the blockmodel assumptions are substantiated by the results in Table 9. The model is useful for encoding biological information about single and multiple functional annotations. The out-of-sample prediction accuracy of the multi-way blockmodel is solid and consistently much higher than the baseline. These results complement and confirm the out-of-sample prediction results we obtained in Section 3.3.

4.4. *Coordinated and differential regulatory response to Nitrogen and Carbon starvation*. Here we provide an illustration of how the multi-way blockmodel can be used to perform quantitative and qualitative analysis of coordinated regulation in response to Nitrogen starvation and differential regulation in response to Carbon starvation. We perform this analysis for the same four lists of genes we considered in Section 4.2.

The quantitative analysis of *coordinated regulation* is based on the number of genes which are estimated to be associated with the various metabolite classes, in both the Nitrogen and the Carbon starvation experiments.

We used the same procedure described in Section 4.2 to estimate the metabolite classes associated with each gene, using the estimated largest and second-largest (gene-block) memberships. Table 10 reports the number of genes that were found to be associated with a primary metabolite class (largest membership) and with

TABLE 10

Quantitative evaluation of coordinated regulatory responses. Number of genes associated with the same metabolite class in both the Nitrogen and Carbon starvation experiments. The association is estimated using both largest and second-largest membership scores

	Largest membership				**Second-largest membership**			
Gene list	**AA**	**BSI**	**GLY**	**TCA**	**AA**	**BSI**	**GLY**	**TCA**
Growth rate	11	6	10	2	10	5	3	1
ESR induced	55	13	4	0	48	21	4	0
ESR repressed	128	22	1	17	109	52	0	27
Metabolic cycle	2	3	1	0	2	3	1	0

a secondary metabolite class (second-largest membership) for each of the four lists of genes we consider.

About a fourth of the genes are found to be associated with a primary metabolite class. Despite the similarity in the patterns of primary and secondary associations, the gene sets involved in them are different. These results imply that another fourth of the genes are found to be associated to a secondary metabolite class. Overall, the blockmodel suggests a substantial amount of overlap between the coordinated regulatory response to Nitrogen and Carbon starvation. A similar quantitative analysis could be conducted for highlighting Nitrogen- and Carbon-specific coordinated regulatory responses.

The qualitative analysis of *differential regulation* is based on the functional enrichment analysis of those genes associated with a given metabolite class in the Nitrogen experiment, but associated with a different metabolic class in the Carbon experiment. For this analysis, we used the procedure above to estimate the metabolite classes associated with each gene, using the estimated largest memberships only, for the list of genes that were found to be ESR induced. The results of the functional analysis obtained for the largest memberships, using the Gene Ontology term finder, are reported in Table 11.

TABLE 11

Functional evaluation of gene–metabolite associations that are differentially regulated in Nitrogen and Carbon. Gene Ontology terms for gene–metabolite associations unique to the Nitrogen starvation experiment. Association is computed using the largest memberships

Class	Ontology	Term description	*p*-value
AA	Function	Alcohol dehydrogenase (NADP+) activity	0.01775
AA	Function	Aldo-keto reductase (NADP) activity	0.01775
AA	Process	Vacuolar protein catabolic process	0.00026
AA	Process	Catabolic process	0.00808
BSI	Function	Peroxidase activity	0.0000568
BSI	Function	Antioxidant activity	0.0004
BSI	Function	Carbohydrate kinase activity	0.00083
BSI	Function	Glutathione peroxidase activity	0.0214
BSI	Process	Carbohydrate catabolic process	2.98E–7
BSI	Process	Cellular response to oxidative stress	0.0000131
BSI	Process	Trehalose metabolic process	0.0000203
BSI	Process	Alcohol catabolic process	0.0000231
BSI	Process	Glycoside metabolic process	0.000061
GLY	Function	Oxidoreductase activity	0.0000116
GLY	Process	Oxidation–reduction process	0.00097
GLY	Process	Cellular carbohydrate metabolic process	0.0011
GLY	Process	Carbohydrate metabolic process	0.00143
GLY	Process	Cellular aldehyde metabolic process	0.00351

These results highlight how the same set of genes (a proxy for proteins) may be using metabolites differently to execute a response to the Carbon. For instance, metabolites in the BSI class are used to process glycoside, alcohol and trehalose, as part of antioxidant activities. Metabolites in the GLY class are used to execute oxidoreductase activities and metabolize aldehyde and carbohydrates. The magnitude of the components of the relevant mixed membership vectors provides more information on the degree of involvement the various gene blocks in these many activities.

A similar qualitative analysis could be carried out to explore the functional landscape, that is, shared by the Nitrogen and Carbon coordinated regulatory responses to starvation.

4.5. *Comparative analysis of raw and preprocessed data.* Here we compare a blockmodel analysis of coordinated regulation with an analysis using cross-association Chakrabarti et al. (2004), quantitatively, in terms of number of gene–metabolite class associations found. We consider the four lists of genes above for this analysis.

Cross-association takes a binary table as input. We built such a genes-by-metabolites binary matrix Y by thresholding the corresponding matrix of correlations. We assign $Y(j,k) = 1$ whenever $\rho(j,k)$ is above the 75th percentile or below the 25th percentile of the empirical correlation distribution.

Cross-association provides a two-way blockmodel as output, in which K_1 and K_2 are estimated using a metric based on information gain. To make a valid comparison, we fit the stochastic multi-way blockmodel with the same number of gene and metabolite blocks.

An additional complication in this analysis is that the number of metabolite blocks can be different from four, for both cross-association and the stochastic blockmodel. We use noninformative priors on the metabolites memberships in the stochastic blockmodel. In addition, we developed a greedy matching procedure to associate metabolite blocks to metabolite classes, after inference. We proceeded as follows. Each metabolite was associated with a block using its largest (metabolite-block) membership. Each metabolite is associated with a known metabolite class. We assigned a metabolite class label to each metabolite block according to a simple majority rule.

We used the same procedure described in Section 4.2 to estimate the metabolite classes associated with each gene, using the estimated largest and second-largest (gene-block) memberships.

Table 12 reports the number of Gene Ontology terms that were found to be associated with a primary metabolite class (largest membership) in the first four rows, and with a secondary metabolite class (second-largest membership) in the next four rows, for each of the four lists of Gene Ontology terms we consider. The last four rows report the number of genes that were found to be associated

TABLE 12

Quantitative evaluation of Gene Ontology terms associated with gene–metabolite class found. Shown in the tables are results for multi-way blockmodel's largest (1*st*) *and second-largest* (2*nd*) *memberships as well as cross-associations* (*CA*)

		AA			BSI			TCA			
Memb.	**Gene list**	**BP**	**CC**	**MF**	**BP**	**CC**	**MF**	**BP**	**CC**	**MF**	**Total**
1st	Growth rate	1	–	–	1	6	5	–	–	–	13
1st	ESR induced	33	7	11	2	1	2	16	–	5	77
1st	ESR repressed	–	45	26	32	23	5	–	–	–	131
1st	Metabolic cycle	13	6	6	–	–	–	–	–	–	25
2nd	Growth rate	4	6	3	–	–	–	–	–	–	13
2nd	ESR induced	–	1	6	–	16	14	–	–	1	38
2nd	ESR repressed	–	47	21	–	34	11	–	–	–	113
2nd	Metabolic cycle	–	–	–	12	6	8	–	–	–	26
CA	Growth rate	–	6	2	2	–	2	–	–	–	12
CA	ESR induced	–	–	–	–	21	19	–	–	–	40
CA	ESR repressed	–	47	20	–	30	20	–	–	–	117
CA	Metabolic cycle	–	–	–	12	6	7	–	–	–	25

with a metabolite class using cross-association. The multi-way stochastic blockmodel finds more primary associations than cross-association, 246 versus 194. In addition, if we consider the secondary associations, the blockmodel analysis uncovers 190 more associations. In fact, subsets of genes associated with the same primary and secondary metabolite class, for instance, AA, are not overlapping by construction.

Overall, cross-association is not well suited for any analysis of biological correlations because of a number of shortcomings, including its reliance on binary input and its lack of flexibility for incorporating prior biological information, for example, the number of metabolite blocks. Our results show that the multi-way stochastic blockmodel outperforms cross-associations quantitatively, even when we do not make use of biological prior knowledge.

5. Concluding remarks. In order to analyze the temporal coordination between gene expression and metabolite concentrations in yeast cells, in response to starvation, we developed a family of multi-way stochastic blockmodels. These models extend the mixed membership stochastic blockmodel [Airoldi et al. (2008)] to the case of two sets of measurements and to the case of Gaussian and binary responses. We developed and compared various inference schemes for multi-way blockmodels, including Monte Carlo Markov chains and variational Bayes.

We further explored the impact of *thresholding* and *binning* on the analysis. These censoring mechanisms are often used as preprocessing steps. The transformed data are then amenable to the analysis of coordination using off-the-shelf

methods, including Bayesian networks and popular blocking algorithms from the data mining literature [Bradley et al. (2009), Chakrabarti et al. (2004)]. The sensitivity analysis suggests that the impact of preprocessing steps that involve censoring is substantial, both from a quantitative perspective and in terms of its impact on biological discovery, in our case study.

Acknowledgments. The authors thank David Madigan for suggesting extensions of the mixed membership blockmodel, in the context of the analysis of adverse events. EMA is an Alfred P. Sloan Research Fellow.

SUPPLEMENTARY MATERIAL

Supplement to "Multi-way blockmodels for analyzing coordinated high-dimensional responses" (DOI: 10.1214/13-AOAS643SUPP; .pdf). We provide additional supporting plots that show both good and poor performance of the Hill estimator for the index of regular variation in a variety of examples.

REFERENCES

AIROLDI, E. M. (2007). Getting started in probabilistic graphical models. *PLoS Computational Biology* **3** e252.

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9** 1981–2014.

AIROLDI, E. M., HUTTENHOWER, C., GRESHAM, D., LU, C., CAUDY, A., DUNHAM, M., BROACH, J. R., BOTSTEIN, D. and TROYANSKAYA, O. G. (2009). Predicting cellular growth from gene expression signatures. *PLoS Computational Biology* **5** e1000257.

AIROLDI, E. M., HASHIMOTO, T. B., BRANDT, N., BAHMANI, T., ATHANASIADOU, N. and GRESHAM, D. J. (2013a). Coordinated dynamics of cell growth and transcription. Preprint.

AIROLDI, E. M., WANG, X. and LIN, X. (2013b). Supplement to "Multi-way blockmodels for analyzing coordinated high-dimensional responses." DOI:10.1214/13-AOAS643SUPP.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBINAND, G. M. and SHERLOCK, G. (2000). Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nature Genetics* **25** 25–29.

AZARI, H. and AIROLDI, E. M. (2012). Graphlet decomposition of a weighted network. *Journal of Machine Learning Research*, *W&CP* (*AI&Stat*) **22** 54–63.

BLEI, D. M. and JORDAN, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143 (electronic). MR2227367

BLOCKER, A. W. and MENG, X. L. (2013). The perils of data pre-processing. *Bernoulli* **19** 1176–1211.

BOYLE, E. I., WENG, S., GOLLUB, J., JIN, H., BOTSTEIN, D., CHERRY, J. M. and SHERLOCK, G. (2004). GO::TermFinder—open source software for accessing gene ontology terms associated with a list of genes. *Bioinformatics* **20** 3710–3715.

BRADLEY, P. H., BRAUER, M. J., RABINOWITZ, J. D. and TROYANSKAYA, O. G. (2009). Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. *PLoS Computational Biology* **5** e1000270.

BRAUER, M. J., HUTTENHOWER, C., AIROLDI, E. M., ROSENSTEIN, R., MATESE, J. C., GRESHAM, D., BOER, V. M., TROYANSKAYA, O. G. and BOTSTEIN, D. (2008). Coordination of growth rate, cell cycle, stress response and metabolic activity in yeast. *Molecular Biology of the Cell* **19** 352–367.

BRAUER, M. J., YUAN, J., BENNETT, B. D., LU, W., KIMBALL, E., BOTSTEIN, D. and RABINOWITZ, J. D. (2006). Conservation of the metabolomic response to starvation across two divergent microbes. *Proc. Natl. Acad. Sci. USA* **103** 19302–19307.

BRAUN, M. and MCAULIFFE, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.* **105** 324–335. MR2757203

CHAKRABARTI, D., PAPADIMITRIOU, S., MODHA, D. and FALOUTSOS, C. (2004). Fully automatic cross-associations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **10** 79–88.

CHENG, Y. and CHURCH, G. M. (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* **8** 93–103.

CHERRY, J. M., BALL, C., WENG, S., JUVIK, G., SCHMIDT, R., ADLER, B., DUNN, C., DWIGHT, S., RILES, L., MORTIMER, R. K. and BOTSTEIN, D. (1997). Genetic and physical maps of Saccharomyces cerevisiae. *Nature* **387** 67–73.

COPE, L., ZHONG, X., GARRETT, E. and PARMIGIANI, G. (2004). MergeMaid: R tools for merging and cross-study validation of gene expression data. *Stat. Appl. Genet. Mol. Biol.* **3** a29. MR2101476

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* **10** 507–521.

FISHER, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

FRANKS, A. M., CSÁRDI, G., DRUMMOND, D. A. and AIROLDI, E. M. (2012). Estimating a structured covariance matrix from multi-lab measurements in high-throughput biology. Preprint.

GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D. and BROWN, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **11** 4241–4257.

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2009). Statistical models of networks. *Foundations and Trends in Machine Learning* **2** 1–117.

JORDAN, M., GHAHRAMANI, Z., JAAKKOLA, T. and SAUL, L. (1999). Introduction to variational methods for graphical models. *Machine Learning* **37** 183–233.

JOUTARD, C., AIROLDI, E. M., FIENBERG, S. E. and LOVE, T. M. (2008). Discovery of latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. In *Data Mining Patterns, New Methods and Applications*. IGI Global, Hershey, PA.

KANEHISA, M. and GOTO, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** 27–30.

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. MR1294740

LU, R., MARKOWETZ, F., UNWIN, R. D., LEEK, J. T., AIROLDI, E. M., MACARTHUR, B. D., LACHMANN, A., ROZOV, R., MA'AYAN, A., BOYER, L. A., TROYANSKAYA, O. G., WHETTON, A. D. and LEMISCHKA, I. R. (2009). Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature* **462** 358–362.

MARKOWETZ, F., AIROLDI, E. M., LEMISCHKA, I. R. and TROYANSKAYA, O. G. (2009). Mapping dynamic histone acetylation patterns to gene expression in nanog-depleted murine embryonic stem cells. Unpublished manuscript.

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255

ROHE, K. and YU, B. (2012). Co-clustering for directed graphs: The stochastic co-blockmodel and a spectral algorithm. Available at arXiv:1204.2296.

SGD project. Saccharomyces genome database. Available at http://www.yeastgenome.org/.

SLAVOV, N., AIROLDI, E. M., VAN OUDENAARDEN, A. and BOTSTEIN, D. (2013). A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes. *Molecular Biology of the Cell* **23** 1986–1997.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester. MR0838090

TROYANSKAYA, O. G., DOLINSKI, K., OWEN, A. B., ALTMAN, R. B. and BOTSTEIN, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). *Proc. Natl. Acad. Sci. USA* **100** 10623–10628.

TU, B. P., KUDLICKI, A., ROWICKA, M. and MCKNIGHT, S. L. (2005). Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science* **310** 1152–1158.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** 290–295. MR0652727

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178–205. MR0773161

E. M. AIROLDI
HARVARD UNIVERSITY
1 OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: airoldi@fas.harvard.edu

X. WANG
UNIVERSITY OF CINCINNATI
2815 COMMONS WAY
CINCINNATI, OHIO 45221
USA

X. LIN
RUTGERS UNIVERSITY
100 ROCK AVENUE
PISCATAWAY, NEW JERSEY 08854
USA

Submitted to the Annals of Applied Statistics
arXiv: stat.ME./0907.0000

SUPPLEMENT TO "MULTI-WAY BLOCKMODELS FOR ANALYZING COORDINATED HIGH-DIMENSIONAL RESPONSES"

By Edoardo M Airoldi[*], Xiaopei Wang[†] and Xiaodong Lin[‡]

Harvard University[*], University of Cincinnati[†] and Rutgers University[‡]

APPENDIX A: DETAILS OF THE VARIATIONAL INFERENCE

To carry out parameter estimation in the proposed multi-way blockmodel, we developed a variational EM algorithm in a similar fashion to (Airoldi, 2007; Jordan et al., 1999) with inference detailed as follows.

A.1. Variational objective. Following notations in Section **??**, assume that we have observations $Y = Y(j,k)$, unknown parameters $\Theta = \{\alpha, \beta, B, \sigma^2\}$, and latent variables $X = \{\vec{\pi}_j, \vec{p_k}, \vec{D}_{j\to k}, \vec{E}_{j\leftarrow k}\}$, where $1 \le j \le N_1$ and $1 \le k \le N_2$. The latent variables X are assumed to follow distribution $q(X)$. By Jensen's inequality, we obtain the following lower bound on the observed data log-likelihood:

$$\log p(Y|\Theta) \ge E_q[\log p(Y, X|\Theta)] - E_q[\log q(X)]. \tag{A.1}$$

For the complete data likelihood $p(Y, X|\Theta)$ we have

$$\begin{aligned} p(Y, X|\alpha, \beta, B, \sigma^2) &= \prod_j p_1(\vec{\pi}_j|\alpha) \prod_k p_1(\vec{p_k}|\beta) \\ &\quad \prod_{j,k} p_0(Y(j,k)|\vec{D}_{j\to k}, \vec{E}_{j\leftarrow k}, B, \sigma^2) p_2(\vec{D}_{j\to k}|\vec{\pi}_j) p_2(\vec{E}_{j\leftarrow k}|\vec{p_k}), \end{aligned} \tag{A.2}$$

where p_0 is a Normal distribution with mean $\mu = \vec{D}'_{j\to k} B \vec{E}_{j\leftarrow k}$ and variance σ^2, p_1 is a Dirichlet distribution and p_2 is a Multinomial distribution with $n = 1$. The latent variable distribution $q(X)$ is approximated by the mean-field fully-factorized family as

$$q(\vec{\pi}, \vec{p}, \vec{D}, \vec{E}|\vec{\nu}, \vec{\xi}, \vec{\phi}, \vec{\eta}) = \prod_j q_1(\vec{\pi}_j|\vec{\nu}_j) \prod_k q_1(\vec{p_k}|\vec{\xi_k}) \prod_{j,k} q_2(\vec{D}_{j\to k}|\vec{\phi}_{j\to k}) q_2(\vec{E}_{j\leftarrow k}|\vec{\eta}_{j\leftarrow k}), \tag{A.3}$$

where q_1 is a Dirichlet distribution and q_2 is a Multinomial distribution with $n = 1$, and $\vec{\nu}, \vec{\xi}, \vec{\phi}, \vec{\eta}$ are the free parameters in the factorized approximation.

Substituting (A.3) and (A.2) into (A.1), we have the approximated lower bound for the log likelihood $L_\Delta(q,\Theta)$ that we aim to maximize.

$$\begin{aligned} L_\Delta(q,\Theta) &= E_q[\log \prod_{j,k} p_0(Y(j,k)|\vec{D}_{j\to k},\vec{E}_{j\leftarrow k},B,\sigma^2)] \\ &+ E_q[\log \prod_{j,k} p_2(\vec{D}_{j\to k}|\vec{\pi}_j)] + E_q[\log \prod_{j,k} p_2(\vec{E}_{j\leftarrow k}|\vec{p}_k)] \\ &+ E_q[\log \prod_{j} p_1(\vec{\pi}_j|\alpha)] + E_q[\log \prod_{k} p_1(\vec{p}_k|\beta)] \\ &- E_q[\log \prod_{j} q_1(\vec{\pi}_j|\vec{\nu}_j)] - E_q[\log \prod_{k} q_1(\vec{p}_k|\vec{\xi}_k)] \\ &- E_q[\log \prod_{j,k} q_2(\vec{D}_{j\to k}|\vec{\phi}_{j\to k})] - E_q[\log \prod_{j,k} q_2(\vec{E}_{j\leftarrow k}|\vec{\eta}_{j\leftarrow k})]. \end{aligned} \tag{A.4}$$

For the first term in (A.4), we have

$$\begin{aligned} & E_q[\log \prod_{j,k} p_0(Y(j,k)|\vec{D}_{j\to k},\vec{E}_{j\leftarrow k},B,\sigma^2)] \\ &= \sum_{j,k} E_q[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{(Y^2(j,k)+\vec{D}'_{j\to k}B^2\,\vec{E}_{j\leftarrow k} - 2Y(j,k)\vec{D}'_{j\to k}B\,\vec{E}_{j\leftarrow k})}{2\sigma^2}] \\ &= -\frac{1}{2}N_1N_2\log 2\pi + \sum_{j,k,g,h}\phi_{j\to k,g}\eta_{j\leftarrow k,h}f(Y(j,k),B(g,h),\sigma^2), \end{aligned}$$

where

$$f(Y(j,k),B(g,h),\sigma^2) = -\frac{\log\sigma^2}{2} - \frac{Y^2(j,k)}{2\sigma^2} - \frac{B^2(g,h)}{2\sigma^2} + \frac{Y(j,k)B(g,h)}{\sigma^2}$$

$$\frac{\partial f}{\partial B(g,h)} = -\frac{B(g,h)}{\sigma^2} + \frac{Y(j,k)}{\sigma^2}$$

$$e^{f(Y(j,k),B(g,h),\sigma^2)} = (\sigma^2\cdot e^{\frac{(Y(j,k)-B(g,h))^2}{\sigma^2}})^{-\frac{1}{2}}.$$

Since

$$E[\log(X_i)] = \psi(\alpha_i) - \psi(\sum_i \alpha_i), \text{for } \vec{X}\sim \text{Dirichlet}(\vec{\alpha}), \text{where } \psi(x) = \frac{d}{dx}\log\Gamma(x),$$

we have

$$
\begin{aligned}
& L_\Delta(q,\Theta) \\
&= -\frac{1}{2}N_1N_2\log 2\pi + \sum_{j,k,g,h}\phi_{j\to k,g}\eta_{j\leftarrow k,h}f(Y(j,k),B(g,h),\sigma^2) \\
&+ \sum_{j,k,g}\phi_{j\to k,g}(\psi(\nu_{j,g})-\psi(\sum_g \nu_{j,g})) + \sum_{j,k,h}\eta_{j\leftarrow k,h}(\psi(\xi_{k,h})-\psi(\sum_h \xi_{k,h})) \\
&+ N_1\log\Gamma(K_1\alpha) - N_1K_1\log\Gamma(\alpha) + \sum_{j,g}(\alpha-1)(\psi(\nu_{j,g})-\psi(\sum_g \nu_{j,g})) \\
&+ N_2\log\Gamma(K_2\beta) - N_2K_2\log\Gamma(\beta) + \sum_{k,h}(\beta-1)(\psi(\xi_{k,h})-\psi(\sum_h \xi_{k,h})) \\
&- \sum_j \log\Gamma(\sum_g \nu_{j,g}) + \sum_{j,g}\log\Gamma(\nu_{j,g}) - \sum_{j,g}(\nu_{j,g}-1)(\psi(\nu_{j,g})-\psi(\sum_g \nu_{j,g})) \\
&- \sum_k \log\Gamma(\sum_h \xi_{k,h}) + \sum_{k,h}\log\Gamma(\xi_{k,h}) - \sum_{k,h}(\xi_{k,h}-1)(\psi(\xi_{k,h})-\psi(\sum_h \xi_{k,h})) \\
&- \sum_{j,k,g}\phi_{j\to k,g}\log\phi_{j\to k,g} - \sum_{j,k,h}\eta_{j\leftarrow k,h}\log\eta_{j\leftarrow k,h}.
\end{aligned}
$$

A.2. Variational E-step. Isolating terms containing $\phi_{j\to k,g}$, we get $L_{\phi_{j\to k,g}}$. Differentiate it with respect to $\phi_{j\to k,g}$, we have

$$
\begin{aligned}
\frac{\partial L_{\phi_{j\to k,g}}}{\partial \phi_{j\to k,g}} &= \sum_h \eta_{j\leftarrow k,h}f(Y(j,k),B(g,h),\sigma^2) + \psi(\nu_{j,g}) - \psi(\sum_g \nu_{j,g}) - \log\phi_{j\to k,g} - 1 \\
&= 0.
\end{aligned}
$$

Thus,

$$
\phi_{j\to k,g} \propto e^{\psi(\nu_{j,g})-\psi(\sum_g \nu_{j,g})}\prod_h (\sigma^2\cdot e^{\frac{(Y(j,k)-B(g,h))^2}{\sigma^2}})^{-\frac{1}{2}\eta_{j\leftarrow k,h}}. \tag{A.5}
$$

Isolating terms containing $\eta_{j\leftarrow k,h}$, we get $L_{\eta_{j\leftarrow k,h}}$. Differentiate it with respect to $\eta_{j\leftarrow k,h}$, we have

$$
\begin{aligned}
\frac{\partial L_{\eta_{j\leftarrow k,h}}}{\partial \eta_{j\leftarrow k,h}} &= \sum_g \phi_{j\to k,g}f(Y(j,k),B(g,h),\sigma^2) + \psi(\xi_{k,h}) - \psi(\sum_h \xi_{k,h}) - \log\eta_{j\leftarrow k,h} - 1 \\
&= 0.
\end{aligned}
$$

Thus,

$$
\eta_{j\leftarrow k,h} \propto e^{\psi(\xi_{k,h})-\psi(\sum_h \xi_{k,h})}\prod_g (\sigma^2\cdot e^{\frac{(Y(j,k)-B(g,h))^2}{\sigma^2}})^{-\frac{1}{2}\phi_{j\to k,g}}. \tag{A.6}
$$

Isolating terms containing $\nu_{j,g}$, we get $L_{\nu_{j,g}}$. Differentiate it with respect to $\nu_{j,g}$, we have

$$\begin{aligned}\frac{\partial L_{\nu_{j,g}}}{\partial \nu_{j,g}} &= (\sum_k \phi_{j\to k,g} + \alpha - \nu_{j,g})\psi'(\nu_{j,g}) - \sum_g (\sum_k \phi_{j\to k,g} + \alpha - \nu_{j,g})\psi'(\sum_g \nu_{j,g}) \\ &= 0.\end{aligned}$$

Thus,

$$\nu_{j,g} = \sum_k \phi_{j\to k,g} + \alpha. \tag{A.7}$$

Isolating terms containing $\xi_{k,h}$, we get $L_{\xi_{k,h}}$. Differentiate it with respect to $\xi_{k,h}$, we have

$$\begin{aligned}\frac{\partial L_{\xi_{k,h}}}{\partial \xi_{k,h}} &= (\sum_j \eta_{j\leftarrow k,h} + \beta - \xi_{k,h})\psi'(\xi_{k,h}) - \sum_h (\sum_j \eta_{j\leftarrow k,h} + \beta - \xi_{k,h})\psi'(\sum_h \xi_{k,h}) \\ &= 0.\end{aligned}$$

Thus,

$$\xi_{k,h} = \sum_j \eta_{j\leftarrow k,h} + \beta. \tag{A.8}$$

A.3. Variational M-step. Isolating terms containing B, we get L_B. Differentiate it with respect to $B(g,h)$, we have

$$\begin{aligned}\frac{\partial L_B}{\partial B(g,h)} &= \sum_{j,k} \phi_{j\to k,g}\eta_{j\leftarrow k,h}\frac{\partial f}{\partial B(g,h)} \\ &= \sum_{j,k} \phi_{j\to k,g}\eta_{j\leftarrow k,h}(-\frac{B(g,h)}{\sigma^2} + \frac{Y(j,k)}{\sigma^2}) \\ &= 0.\end{aligned}$$

Thus,

$$B(g,h) = \frac{\sum_{j,k} \phi_{j\to k,g}\eta_{j\leftarrow k,h}Y(j,k)}{\sum_{j,k} \phi_{j\to k,g}\eta_{j\leftarrow k,h}}. \tag{A.9}$$

APPENDIX B: DETAILS OF THE MCMC INFERENCE

In the following we present an alternative inference approach using the collapsed Gibbs sampler (Liu, 1994). Considering B as a latent parameter, we have the set of latent parameters $X = \{\vec{\pi}_j, \vec{p_k}, B\}$, and the set

of latent variables $Z = \{\vec{D}_{j\to k}, \vec{E}_{j\leftarrow k}\}$. The set of hyper-parameters are $\Theta = \{\alpha, \beta, \sigma^2, \mu_B(g,h), \sigma_B^2(g,h)\}$. After integrating out the latent parameters $\vec{\pi}$, $\vec{p}$ and B from the complete data likelihood $p(Y, \vec{\pi}, \vec{p}, \vec{D}, \vec{E}, B|\Theta)$, we obtain the marginal distributions $p(Y, \vec{D}, \vec{E})$ and $p(Y, D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}})$. $D_{-\{j\to k\}}$ denotes excluding the membership of $j \to k$ from the set of latent memberships D, and similarly for $E_{-\{j\leftarrow k\}}$. $\neg\{j,k\}$ denotes excluding the entry (j,k). Subsequently, the posterior distribution of the latent variables $\vec{D}_{j\to k}$ and $\vec{E}_{j\leftarrow k}$ can be computed using B.1, and $\vec{\pi}$, $\vec{p}$ and B can be updated via B.2-B.4. Computational procedures are illustrated in Algorithm 1 with details following. Therein, g_0 and h_0 denote the group memberships before the update, and g_1 and h_1 represent the group memberships after the update.

MCMC $\left(Y(j,k)_{j=1,k=1}^{N_1,N_2}, \alpha, \beta, \sigma^2, \mu_B(g,h), \sigma_B^2(g,h) \right)$

initialize $\vec{D}_{j\to k} := -1$ for all j and k
initialize $\vec{E}_{j\leftarrow k} := -1$ for all j and k
initialize $\vec{D}_{j\to\cdot} := 0$ for all j
initialize $\vec{E}_{\cdot\leftarrow k} := 0$ for all k
initialize $Y_{gh} := 0$ for all g and h
initialize $n_{gh} := 0$ for all g and h
initialize $p(D_{j\to k,g} = 1, E_{j\leftarrow k,h} = 1) := 1/(K_1 * K_2)$ for all j and k, and all g and h
repeat
  for $j = 1$ *to* N_1 **do**
    for $k = 1$ *to* N_2 **do**
      if *the current membership is* (g_0, h_0)*:* $D_{j\to k,g_0} = 1, E_{j\leftarrow k,h_0} = 1$ **then**
        subtract $Y(j,k)$ from $Y_{g_0h_0}$ and 1 from $n_{g_0h_0}$
        subtract 1 from $D_{j\to\cdot,g_0}$ and 1 from $E_{\cdot\leftarrow k,h_0}$
      update $p(D_{j\to k,g} = 1, E_{j\leftarrow k,h} = 1|D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y)$ for all g and h using Eq.(B.1) and normalize to sum to 1
      draw sample for $\vec{D}_{j\to k}$ $\vec{E}_{j\leftarrow k}$ simultaneously from the probability
      if *the new membership is* (g_1, h_1)*:* $D_{j\to k,g_1} = 1, E_{j\leftarrow k,h_1} = 1$ **then**
        add $Y(j,k)$ to $Y_{g_1h_1}$ and 1 to $n_{g_1h_1}$
        add 1 to $D_{j\to\cdot,g_1}$ and 1 to $E_{\cdot\leftarrow k,h_1}$
until N iterations after burn-in;
return $(\vec{D}, \vec{E}, \vec{D}_{j\to\cdot}, \vec{E}_{\cdot\leftarrow k}, Y_{gh}, n_{gh})$
estimate $(\vec{\pi}, \vec{p}, B)$ by Eq.(B.2)(B.3)(B.4)

Algorithm 1: The MCMC algorithm.

$$
\begin{aligned}
(\text{B.1}) \qquad & p(D_{j\to k,g}=1, E_{j\leftarrow k,h}=1 | D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y) \\
\propto \quad & \sqrt{\frac{n_{gh}^{\neg\{j,k\}}\sigma_B^2(g,h)+\sigma^2}{(n_{gh}^{\neg\{j,k\}}+1)\sigma_B^2(g,h)+\sigma^2}}(\alpha+D_{j\to\cdot,g}^{\neg\{j,k\}})(\beta+E_{\cdot\leftarrow k,h}^{\neg\{j,k\}}) \\
\times \quad & \exp\{\frac{(\frac{Y_{gh}^{\neg\{j,k\}}+Y(j,k)}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)})^2}{2(\frac{(n_{gh}^{\neg\{j,k\}}+1)}{\sigma^2}+\frac{1}{\sigma_B^2(g,h)})} - \frac{(\frac{Y_{gh}^{\neg\{j,k\}}}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)})^2}{2(\frac{n_{gh}^{\neg\{j,k\}}}{\sigma^2}+\frac{1}{\sigma_B^2(g,h)})}\}.
\end{aligned}
$$

$$
\begin{aligned}
(\text{B.2}) \qquad \pi_{j,g} &= \frac{\alpha+D_{j\to\cdot,g}}{\sum_g(\alpha+D_{j\to\cdot,g})} \\
&= \frac{\alpha+D_{j\to\cdot,g}}{K_1\alpha+N_2}.
\end{aligned}
$$

$$
\begin{aligned}
(\text{B.3}) \qquad p_{k,h} &= \frac{\beta+E_{\cdot\leftarrow k,h}}{\sum_h(\beta+E_{\cdot\leftarrow k,h})} \\
&= \frac{\beta+E_{\cdot\leftarrow k,h}}{K_2\beta+N_1}.
\end{aligned}
$$

$$
(\text{B.4}) \qquad B(g,h) = \frac{\frac{Y_{gh}}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)}}{\frac{n_{gh}}{\sigma^2}+\frac{1}{\sigma_B^2(g,h)}}.
$$

B.1. Complete data likelihood.

$$
\begin{aligned}
& p(Y,\vec{\pi},\vec{p},\vec{D},\vec{E},B|\alpha,\beta,\mu_B,\sigma_B^2,\sigma^2) \\
(\text{B.5}) \qquad = \quad & \prod_{j,k} p_0(Y(j,k)|\vec{D}_{j\to k},\vec{E}_{j\leftarrow k},B,\sigma^2) \\
(\text{B.6}) \qquad \times \quad & \prod_j p_1(\vec{\pi}_j|\alpha)\prod_k p_1(\vec{p}_k|\beta) \\
(\text{B.7}) \qquad \times \quad & \prod_{j,k} p_2(\vec{D}_{j\to k}|\vec{\pi}_j)p_2(\vec{E}_{j\leftarrow k}|\vec{p}_k) \\
(\text{B.8}) \qquad \times \quad & \prod_{g,h} p_3(B(g,h)|\mu_B(g,h),\sigma_B^2(g,h)),
\end{aligned}
$$

where p_0 is a Normal distribution with mean $\mu = \vec{D}'_{j\to k} B \vec{E}_{j\leftarrow k}$ and variance σ^2, p_1 is a Dirichlet distribution, p_2 is a Multinomial distribution with $n = 1$, and p_3 is a Normal distribution.

$$(B.5) \quad = \quad (\frac{1}{\sqrt{2\pi}\sigma})^{N_1 N_2} \exp\{-\frac{\sum_{j,k} Y^2(j,k)}{2\sigma^2}\} \prod_{g,h} \exp\{-\frac{B^2(g,h) n_{gh}}{2\sigma^2}\} \exp\{\frac{B(g,h) Y_{gh}}{\sigma^2}\},$$

where

$$n_{gh} \stackrel{def}{=} \sum_{j,k} \mathbf{1}(D_{j\to k,g} = 1 \text{ and } E_{j\leftarrow k,h} = 1)$$

and

$$Y_{gh} \stackrel{def}{=} \sum_{j,k} Y(j,k) \mathbf{1}(D_{j\to k,g} = 1 \text{ and } E_{j\leftarrow k,h} = 1).$$

$$(B.6) = \prod_j \left(\frac{\Gamma(K_1\alpha)}{(\Gamma(\alpha))^{K_1}} \prod_g \pi_{j,g}^{\alpha-1} \right) \prod_k \left(\frac{\Gamma(K_2\beta)}{(\Gamma(\beta))^{K_2}} \prod_h p_{k,h}^{\beta-1} \right).$$

$$(B.7) \quad = \quad \prod_{j,g} \pi_{j,g}^{D_{j\to\cdot,g}} \times \prod_{k,h} p_{k,h}^{E_{\cdot\leftarrow k,h}},$$

where

$$D_{j\to\cdot,g} \stackrel{def}{=} \sum_k \mathbf{1}(D_{j\to k,g} = 1)$$

and

$$E_{\cdot\leftarrow k,h} \stackrel{def}{=} \sum_j \mathbf{1}(E_{j\leftarrow k,h} = 1).$$

$$(B.8) \quad = \quad \prod_{g,h} \frac{1}{\sqrt{2\pi}\sigma_B(g,h)} \exp\{-\frac{B^2(g,h)}{2\sigma_B^2(g,h)}\} \exp\{-\frac{\mu_B^2(g,h)}{2\sigma_B^2(g,h)}\} \exp\{\frac{B(g,h)\mu_B(g,h)}{\sigma_B^2(g,h)}\}.$$

B.2. Marginal distribution. For the marginal distribution $p(\vec{D}, \vec{E})$, we have

$$\begin{aligned} p(\vec{D}, \vec{E}) &= p(\vec{D}) p(\vec{E}) \\ &= \int p(\vec{D}|\vec{\pi}) p(\vec{\pi}) d\vec{\pi} \int p(\vec{E}|\vec{p}) p(\vec{p}) d\vec{p}. \end{aligned}$$

From (B.6) and (B.7) we have

$$(B.6) \times (B.7) = \frac{(\Gamma(K_1\alpha))^{N_1}}{(\Gamma(\alpha))^{K_1N_1}} \prod_{j,g} \pi_{j,g}^{\alpha+D_{j\to\cdot,g}-1} \tag{B.9}$$

$$\times \frac{(\Gamma(K_2\beta))^{N_2}}{(\Gamma(\beta))^{K_2N_2}} \prod_{k,h} p_{k,h}^{\beta+E_{\cdot\leftarrow k,h}-1}. \tag{B.10}$$

Integrate out $\vec{\pi}$ from (B.9) and $\vec{p}$ from (B.10), we have

$$\begin{aligned} p(\vec{D},\vec{E}) &= \int \frac{(\Gamma(K_1\alpha))^{N_1}}{(\Gamma(\alpha))^{K_1N_1}} \prod_{j,g} \pi_{j,g}^{\alpha+D_{j\to\cdot,g}-1} \frac{(\Gamma(K_2\beta))^{N_2}}{(\Gamma(\beta))^{K_2N_2}} \prod_{k,h} p_{k,h}^{\beta+E_{\cdot\leftarrow k,h}-1} d\vec{\pi} d\vec{p} \\ &= \frac{(\Gamma(K_1\alpha))^{N_1}(\Gamma(K_2\beta))^{N_2} \prod_{j,g}\Gamma(\alpha+D_{j\to\cdot,g}) \prod_{k,h}\Gamma(\beta+E_{\cdot\leftarrow k,h})}{(\Gamma(\alpha))^{K_1N_1}(\Gamma(\beta))^{K_2N_2}(\Gamma(K_1\alpha+N_2))^{N_1}(\Gamma(K_2\beta+N_1))^{N_2}}. \end{aligned}$$

For distribution $p(Y|\vec{D},\vec{E})$, we have

$$p(Y|\vec{D},\vec{E}) = \int p(Y|\vec{D},\vec{E},B)p(B)dB.$$

From (B.5) and (B.8) we have

$$\begin{aligned} &(B.5) \times (B.8) \\ &= (\frac{1}{\sqrt{2\pi}\sigma})^{N_1N_2} \exp\{-\frac{\sum_{j,k} Y^2(j,k)}{2\sigma^2}\} \\ &\times \prod_{g,h} \frac{\exp\{-\frac{\mu_B^2(g,h)}{2\sigma_B^2(g,h)}\}}{\sqrt{2\pi}\sigma_B(g,h)} \exp\{-B^2(g,h)(\frac{n_{gh}}{2\sigma^2}+\frac{1}{2\sigma_B^2(g,h)})\} \exp\{B(g,h)(\frac{Y_{gh}}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)})\}. \end{aligned}$$

Integrate out $B(g,h)$ for all g,h, we have

$$\begin{aligned} p(Y|\vec{D},\vec{E}) &= (\frac{1}{\sqrt{2\pi}\sigma})^{N_1N_2} \exp\{-\frac{\sum_{j,k} Y^2(j,k)}{2\sigma^2}\} \prod_{g,h} \sqrt{\frac{\sigma^2}{n_{gh}\sigma_B^2(g,h)+\sigma^2}} \\ &\times \exp\{-\frac{1}{2}[\frac{\mu_B^2(g,h)}{\sigma_B^2(g,h)} - \frac{(\frac{Y_{gh}}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)})^2}{(\frac{n_{gh}}{\sigma^2}+\frac{1}{\sigma_B^2(g,h)})}]\}. \end{aligned}$$

For marginal distribution $p(Y, \vec{D}, \vec{E})$, we have

$$\begin{aligned} p(Y, \vec{D}, \vec{E}) &= (\frac{1}{\sqrt{2\pi}\sigma})^{N_1 N_2} \exp\{-\frac{\sum_{j,k} Y^2(j,k)}{2\sigma^2}\} \\ &\times \prod_{g,h} \sqrt{\frac{\sigma^2}{n_{gh}\sigma_B^2(g,h)+\sigma^2}} \exp\{-\frac{1}{2}[\frac{\mu_B^2(g,h)}{\sigma_B^2(g,h)} - \frac{(\frac{Y_{gh}}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)})^2}{(\frac{n_{gh}}{\sigma^2}+\frac{1}{\sigma_B^2(g,h)})}]\} \\ &\times \frac{(\Gamma(K_1\alpha))^{N_1}(\Gamma(K_2\beta))^{N_2}\prod_{j,g}\Gamma(\alpha+D_{j\to\cdot,g})\prod_{k,h}\Gamma(\beta+E_{\cdot\leftarrow k,h})}{(\Gamma(\alpha))^{K_1N_1}(\Gamma(\beta))^{K_2N_2}(\Gamma(K_1\alpha+N_2))^{N_1}(\Gamma(K_2\beta+N_1))^{N_2}}. \end{aligned}$$

B.3. Full conditionals and collapsed Gibbs sampling. In the following, the quantities Y_{gh}, n_{gh}, $\vec{D}_{j\to k}$ and $\vec{E}_{j\leftarrow k}$ are the current estimates before the new iteration. In addition, we assume that in the current iteration, $D_{j\to k,g_0} = 1$, $E_{j\leftarrow k,h_0} = 1$, and in the new iteration $D_{j\to k,g_1} = 1$, $E_{j\leftarrow k,h_1} = 1$. The notions $\neg\{j,k\}$, $-\{j\to k\}$ and $-\{j\leftarrow k\}$ represent excluding the entry (j,k), the latent memberships $D_{j\to k,g_0} = 1$, and $E_{j\leftarrow k,h_0} = 1$, respectively.

Suppose $g_1 \neq g_0$ and $h_1 \neq h_0$, we have

$$\begin{aligned} &p(D_{j\to k,g_1}=1, E_{j\leftarrow k,h_1}=1 | D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y) \\ &\propto \frac{p(D_{j\to k,g_1}=1, E_{j\leftarrow k,h_1}=1, D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y)}{p(D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y)}, \end{aligned}$$

where

$$\begin{aligned} &p(D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y) \\ &\propto \sqrt{\frac{\sigma^2}{(n_{g_0h_0}-1)\sigma_B^2(g_0,h_0)+\sigma^2}} \exp\{\frac{(\frac{Y_{g_0h_0}-Y(j,k)}{\sigma^2}+\frac{\mu_B(g_0,h_0)}{\sigma_B^2(g_0,h_0)})^2}{2(\frac{(n_{g_0h_0}-1)}{\sigma^2}+\frac{1}{\sigma_B^2(g_0,h_0)})}\} \\ &\times \Gamma(\alpha+D_{j\to\cdot,g_0}-1)\Gamma(\beta+E_{\cdot\leftarrow k,h_0}-1) \\ &\times \sqrt{\frac{\sigma^2}{n_{g_1h_1}\sigma_B^2(g_1,h_1)+\sigma^2}} \exp\{\frac{(\frac{Y_{g_1h_1}}{\sigma^2}+\frac{\mu_B(g_1,h_1)}{\sigma_B^2(g_1,h_1)})^2}{2(\frac{n_{g_1h_1}}{\sigma^2}+\frac{1}{\sigma_B^2(g_1,h_1)})}\} \\ &\times \Gamma(\alpha+D_{j\to\cdot,g_1})\Gamma(\beta+E_{\cdot\leftarrow k,h_1}), \end{aligned}$$

and

$$\begin{aligned}
&p(D_{j\to k,g_1}=1, E_{j\leftarrow k,h_1}=1, D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y)\\
&\propto \sqrt{\frac{\sigma^2}{(n_{g_0h_0}-1)\sigma_B^2(g_0,h_0)+\sigma^2}}\exp\{\frac{(\frac{Y_{g_0h_0}-Y(j,k)}{\sigma^2}+\frac{\mu_B(g_0,h_0)}{\sigma_B^2(g_0,h_0)})^2}{2(\frac{(n_{g_0h_0}-1)}{\sigma^2}+\frac{1}{\sigma_B^2(g_0,h_0)})}\}\\
&\times \Gamma(\alpha+D_{j\to\cdot,g_0}-1)\Gamma(\beta+E_{\cdot\leftarrow k,h_0}-1)\\
&\times \sqrt{\frac{\sigma^2}{(n_{g_1h_1}+1)\sigma_B^2(g_1,h_1)+\sigma^2}}\exp\{\frac{(\frac{Y_{g_1h_1}+Y(j,k)}{\sigma^2}+\frac{\mu_B(g_1,h_1)}{\sigma_B^2(g_1,h_1)})^2}{2(\frac{(n_{g_1h_1}+1)}{\sigma^2}+\frac{1}{\sigma_B^2(g_1,h_1)})}\}\\
&\times \Gamma(\alpha+D_{j\to\cdot,g_1}+1)\Gamma(\beta+E_{\cdot\leftarrow k,h_1}+1).
\end{aligned}$$

Therefore

$$\begin{aligned}
&p(D_{j\to k,g_1}=1, E_{j\leftarrow k,h_1}=1|D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y)\\
&\propto \sqrt{\frac{n_{g_1h_1}\sigma_B^2(g_1,h_1)+\sigma^2}{(n_{g_1h_1}+1)\sigma_B^2(g_1,h_1)+\sigma^2}}(\alpha+D_{j\to\cdot,g_1})(\beta+E_{\cdot\leftarrow k,h_1})\\
&\times \exp\{\frac{(\frac{Y_{g_1h_1}+Y(j,k)}{\sigma^2}+\frac{\mu_B(g_1,h_1)}{\sigma_B^2(g_1,h_1)})^2}{2(\frac{(n_{g_1h_1}+1)}{\sigma^2}+\frac{1}{\sigma_B^2(g_1,h_1)})}-\frac{(\frac{Y_{g_1h_1}}{\sigma^2}+\frac{\mu_B(g_1,h_1)}{\sigma_B^2(g_1,h_1)})^2}{2(\frac{n_{g_1h_1}}{\sigma^2}+\frac{1}{\sigma_B^2(g_1,h_1)})}\}.
\end{aligned}$$

Similar results can be derived for $\{g_1=g_0, h_1\neq h_0\}$, $\{g_1\neq g_0, h_1=h_0\}$, and $\{g_1=g_0, h_1=h_0\}$. In summary, we have the conditional distribution for sampling the pair (g,h) for (j,k) in a new iteration:

$$\begin{aligned}
&p(D_{j\to k,g}=1, E_{j\leftarrow k,h}=1|D_{-\{j\to k\}}, E_{-\{j\leftarrow k\}}, Y)\\
&\propto \sqrt{\frac{n_{gh}^{\neg\{j,k\}}\sigma_B^2(g,h)+\sigma^2}{(n_{gh}^{\neg\{j,k\}}+1)\sigma_B^2(g,h)+\sigma^2}}(\alpha+D_{j\to\cdot,g}^{\neg\{j,k\}})(\beta+E_{\cdot\leftarrow k,h}^{\neg\{j,k\}})\\
&\times \exp\{\frac{(\frac{Y_{gh}^{\neg\{j,k\}}+Y(j,k)}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)})^2}{2(\frac{(n_{gh}^{\neg\{j,k\}}+1)}{\sigma^2}+\frac{1}{\sigma_B^2(g,h)})}-\frac{(\frac{Y_{gh}^{\neg\{j,k\}}}{\sigma^2}+\frac{\mu_B(g,h)}{\sigma_B^2(g,h)})^2}{2(\frac{n_{gh}^{\neg\{j,k\}}}{\sigma^2}+\frac{1}{\sigma_B^2(g,h)})}\}.
\end{aligned}$$

B.4. Parameter estimation. After the collapsed Gibbs sampling, we can estimate $\{\vec{\pi}_j, \vec{p_k}, B\}$ using $\{\vec{D}_{j\to k}, \vec{E}_{j\leftarrow k}\}$. First

$$\vec{\pi} = \arg\max_{\vec{\pi}}\{\log(p(\vec{D}|\vec{\pi})p(\vec{\pi}))-\sum_j\lambda_j(\sum_g\pi_{j,g}-1)\},$$

where

$$p(\vec{D}|\vec{\pi})p(\vec{\pi}) = \frac{(\Gamma(K_1\alpha))^{N_1}}{(\Gamma(\alpha))^{K_1N_1}} \prod_{j,g} \pi_{j,g}^{\alpha+D_{j\to\cdot,g}-1}.$$

To enable nonnegative estimates, we have

$$\begin{aligned} \pi_{j,g} &= \frac{\alpha + D_{j\to\cdot,g}}{\sum_g(\alpha + D_{j\to\cdot,g})} \\ &= \frac{\alpha + D_{j\to\cdot,g}}{K_1\alpha + N_2}. \end{aligned}$$

For $\vec{p}$ we have

$$\vec{p} = \arg\max_{\vec{p}} \{\log(p(\vec{E}|\vec{p})p(\vec{p})) - \sum_k \lambda_k(\sum_h p_{k,h} - 1)\},$$

where

$$p(\vec{E}|\vec{p})p(\vec{p}) = \frac{(\Gamma(K_2\beta))^{N_2}}{(\Gamma(\beta))^{K_2N_2}} \prod_{k,h} p_{k,h}^{\beta+E_{\cdot\leftarrow k,h}-1}.$$

To enable nonnegative estimates, we have

$$\begin{aligned} p_{k,h} &= \frac{\beta + E_{\cdot\leftarrow k,h}}{\sum_h(\beta + E_{\cdot\leftarrow k,h})} \\ &= \frac{\beta + E_{\cdot\leftarrow k,h}}{K_2\beta + N_1}. \end{aligned}$$

For the estimate on B we have

$$\begin{aligned} B &= \arg\max_B \log(p(Y|\vec{D},\vec{E},B)p(B)) \\ &= \arg\max_B \log\{\prod_{g,h} \exp[-B^2(g,h)(\frac{n_{gh}}{2\sigma^2} + \frac{1}{2\sigma_B^2(g,h)}) + B(g,h)(\frac{Y_{gh}}{\sigma^2} + \frac{\mu_B(g,h)}{\sigma_B^2(g,h)})]\}. \end{aligned}$$

Therefore,

$$B(g,h) = \frac{\frac{Y_{gh}}{\sigma^2} + \frac{\mu_B(g,h)}{\sigma_B^2(g,h)}}{\frac{n_{gh}}{\sigma^2} + \frac{1}{\sigma_B^2(g,h)}}.$$

APPENDIX C: COMPARATIVE ANALYSIS OF CONVERGENCE

Variational EM (vEM) and its stochastic counterparts have become an popular approach in statistics and machine learning for models where exact posterior inference is challenging (Jordan et al., 1999; Le Cun, 2004; Bottou, 2010; Toulis et al., 2013). The relative merits between vEM and MCMC have

been explored in detail (e.g, see Braun and McAuliffe, 2010). In the main text (Airoldi et al., 2013), we aimed at exploring the trade-off between estimation accuracy and computational burden that vEM helps manage in the context of multi-way blockmodels. Here we offer details about the empirical convergence analysis outlined in the main text.

The vEM inference procedure solves an optimization problem, no sampling is involved. Inference by means of vEM requires key choices about (1) the error tolerance for both the approximate E step and the M step (we picked 10^{-7} for the M Step and 10 iterations for the approximate E step), (2) how to design multiple initializations, and (3) how many to use (we tested 10 uniform and 10 random initializations, in this revision). Note that the vEM procedure involves two maximizations: one for the M step, as in regular EM, and one for the E step, within each iteration of the M step, which serves to tighten the variational lower bound for the likelihood. While convergence of the entire vEM procedure requires at most 1,000 M step iterations, for our chosen error tolerance, only few iterations of the E step updates are needed to optimize the variational lower bound for the likelihood, within each M step iteration. In detail, the Equations in the main text state the approximate E step updates for the set of variational free parameters (ϕ, η, ν, ξ), within each M Step iteration. We found that 10 iterations are enough to obtain a reasonable approximation for the variational lower bound. The principle at work here is similar to that underlying the 1-step move required in SIRM particle filters, and the 1-step Newton improvement of unbiased estimators in Lehmann and Casella (2003).

MCMC is a sampling approach. Inference by means of MCMC requires key choices about: (1) convergence criteria (we used the Gelman-Rubin and Raftery-Lewis for the median, leading to about 6,000 iterations), (2) burn-in (1,000 iteration), (3) thinning to reduce autocorrelation (we recorded one sample every 10 iterations), and (4) multiple chains (we used 10).

One common issue with MCMC is the correlation among subsequent samples. Figure 1 illustrates the properties of MCMC inference for one component of $\hat{B}$, obtained in one of our experiments via collapsed Gibbs sampler. From left to right, the panels show the autocorrelation function (ACF) before and after thinning, and the trace of the last 1000 iterations. These results suggest that the autocorrelations are reduced significantly from thinned samples after burn-in. Figure 2 shows the log-likelihood in the same experiment, both for variational EM (left panel) and for MCMC (right panel).

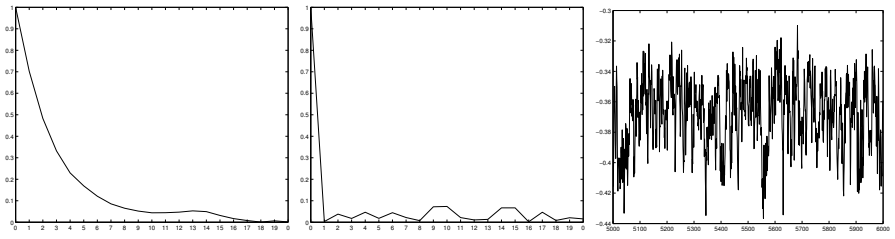

FIG 1. *Illustration of the properties of MCMC inference for one component of* $\hat{B}$. *From left to right: absolute ACF for the entire sample chain, absolute ACF for thinned samples after burn-in, trace for the last 1000 iterations.*

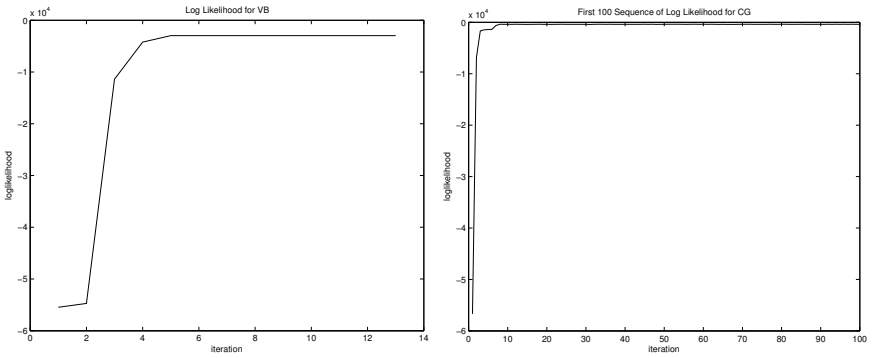

FIG 2. *Log-likelihood of variational Bayes (all iterations before convergence) and MCMC (the first 100 iterations).*

REFERENCES

E. M. Airoldi. Getting started in probabilistic graphical models. *PLoS Computational Biology*, 3(12):e252, 2007.

E. M. Airoldi, X. Wang, and X. Lin. Multi-way blockmodels for analyzing coordinated high-dimensional responses. *Annals of Applied Statistics*, 2013.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Compstat*, volume 2010, pages 177–186, 2010.

M. Braun and J. D. McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

L.B.Y. Le Cun. Large scale online learning. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, volume 16, page 217. MIT Press, 2004.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2nd edition, 2003.

J. S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427): 958–966, 1994.

P. Toulis, J. Rennie, and E. M. Airoldi. Fitting large scale GLMs with implicit updates.

Manuscript, February 2013.