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 $1 \leq j \leq N_1$ and $1 \leq k \leq N_2$. 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:

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

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

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

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.

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

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

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

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,

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

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,

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

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.

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

$$\pi_{j,g} = \frac{\alpha + D_{j\to\cdot,g}}{\sum_g(\alpha + D_{j\to\cdot,g})} = \frac{\alpha + D_{j\to\cdot,g}}{K_1\alpha + N_2}. \tag{B.2}$$

$$p_{k,h} = \frac{\beta + E_{\cdot\leftarrow k,h}}{\sum_h(\beta + E_{\cdot\leftarrow k,h})} = \frac{\beta + E_{\cdot\leftarrow k,h}}{K_2\beta + N_1}. \tag{B.3}$$

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

B.1. Complete data likelihood.

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

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

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

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

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

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

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

$$\begin{aligned} &(B.5) \times (B.8) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{N_1 N_2} \exp\{-\frac{\sum_{j,k} Y^2(j,k)}{2\sigma^2}\} \\ &\times \prod_{g,h} \frac{\exp\{-\frac{\mu_B^2(g,h)}{2\sigma_B^2(g,h)}\}}{\sqrt{2\pi}\sigma_B(g,h)} \exp\{-B^2(g,h)(\frac{n_{gh}}{2\sigma^2} + \frac{1}{2\sigma_B^2(g,h)})\} \exp\{B(g,h)(\frac{Y_{gh}}{\sigma^2} + \frac{\mu_B(g,h)}{\sigma_B^2(g,h)})\}. \end{aligned}$$

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

$$\begin{aligned} p(Y|\vec{D}, \vec{E}) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{N_1 N_2} \exp\{-\frac{\sum_{j,k} Y^2(j,k)}{2\sigma^2}\} \prod_{g,h} \sqrt{\frac{\sigma^2}{n_{gh}\sigma_B^2(g,h) + \sigma^2}} \\ &\times \exp\{-\frac{1}{2}[\frac{\mu_B^2(g,h)}{\sigma_B^2(g,h)} - \frac{(\frac{Y_{gh}}{\sigma^2} + \frac{\mu_B(g,h)}{\sigma_B^2(g,h)})^2}{(\frac{n_{gh}}{\sigma^2} + \frac{1}{\sigma_B^2(g,h)})}]\}. \end{aligned}$$

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

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

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

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

where

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

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

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

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).

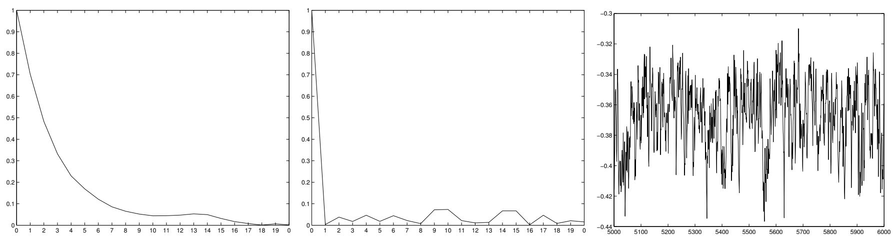

FIG 1. *Illustration of the properties of MCMC inference for one component of $\hat{B}$. From left to right: absolute ACF for the entire sample chain, absolute ACF for thinned samples after burn-in, trace for the last 1000 iterations.*

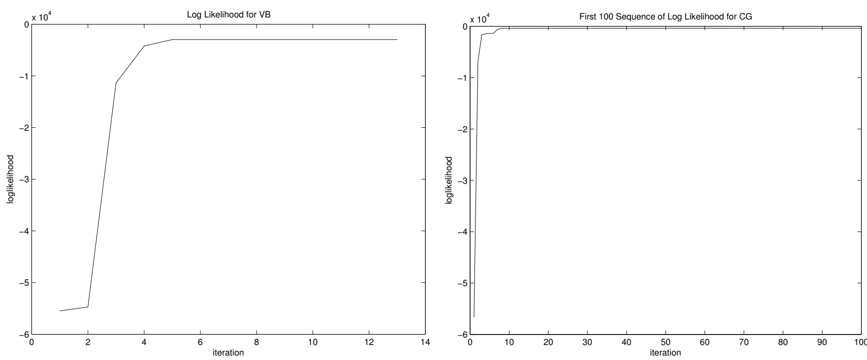

FIG 2. *Log-likelihood of variational Bayes (all iterations before convergence) and MCMC (the first 100 iterations).*

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.