Foundations of Graphical Models - Final project December 21, 2019

Achille Nazaret (aon2018), David Cheikhi (dc3430)

Abstract

The majority of graphical models studied rely on independence assumptions between parts of the model: the *plates*. From topic models to standard variational auto-encoder (VAE), the independence hypothesis leads to efficient inference algorithms such as stochastic variational inference (SVI) with reduced variance or even amortized inference. However, real data often comes with an additional graph dependency information between observations that challenges independence. We choose to leverage this information and gain in both interpetability and modelization. We review the limits of independence and propose a simple model to overcome them. Then we design two inference algorithms and shape a path toward SVI. Our model can identify mixed populations and recover interactions between them.

1 Introduction and motivations

In this project, we extend a typical class of graphical models to non iid settings. First, we recall what is this usual modelization, then we question its assumptions and finally present our contributions. In the classical setting, a scientist collects covariates x_i corresponding to n different entities like individuals or cells. Then, she designs a graphical model to explain the dataset by doing (conditional) **independence** assumptions between each entities i, given global latent variables Θ. A representation is in figure 1.

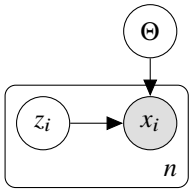


Figure 1: Typical model with observations for n entities.

Mixture models, matrix factorization, mixed membership models and Bayesian linear regression fall in this class. The assumptions of such graphical models are the following:

A1: Internal states: $z_i \overset{\text{iid}}{\sim} f(\alpha)$ – Each entity state is independent from the other

A2: Observed states: $x_i \mid (z_i, \Theta) \overset{\text{iid}}{\sim} g(z_i)$ – Observed states are conditionally independent

Now, we assume we have **relational information** $\mathcal{R}$ about the entities for potential interactions between them. It can be spatial locations or social network relations. Thus, we want to relax the independence assumption to capture interpretable interactions. We review the two assumptions.

Independent internal states z_i Assumption A1 is related to organizational patterns between entities in G. For instance, entities of type $z = a$ and entities of type $z = b$ can be more likely connected such that $(z_i, z_j) \mid \mathcal{R}$ are no longer independent. A large domain about it are Hidden Random Markov Fields, that can be used to cluster spatial data [1]. Others like Relational Topic Models [2] aim at predicting missing links between the entities. Finally, a recent article [3] suggests to use a joint prior and variational family to build a correlated VAE. However, none of them leverage the relations to detect and disentangle the correlations.

Conditionally independent observed states x_i Assumption A2 is more subtle. The z_i could be independent, but then, given the neighbors of entity i, the observed state x_i would be different. For instance if a biological cell of type a is next to a cell of type b then its gene expression can be down-regulated [4]. This dependency is about local influence on the observed state.

Our Goal Our goal is to challenge the second assumption and keep the first one: we assume that the observed entities have an intrinsic independent state but that their observed state can be influenced by their relationships.

2 Our approach

Let's have n entities with observed states x_i, and hidden states z_i. We suppose the z_i are **independent**. We also have a graph $G = (V, E)$ where an edge $(i, j) \in E$ represents a possible interaction between entity i and j that impacts x_i, x_j.

Therefore, the general model can be written in the following form with $P, f(.), g(.)$ distributions:

$$z_i \overset{\text{iid}}{\sim} f(\alpha) \qquad \Theta \sim P$$
$$x_i \mid z_{N(i)}, z_i \overset{\text{iid}}{\sim} g(z_i, z_{N(i)}, \Theta)$$

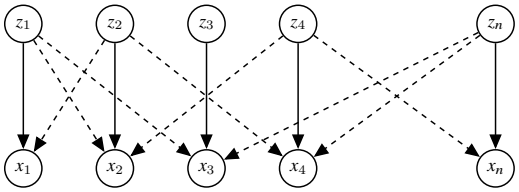


Figure 2: The studied graphical model with dashed edges only if nodes are neighbors in G

Our goal is to infer the latent parameters Θ as well as the latent space z. This is non-identifiable and we decide to design a very constrained model to analyze this new type of algorithms. We choose to focus on model where the neighbors perturbation stays small.

2.1 The model

We consider a model which extends the mixture model by adding interactions ρ between clusters. We use conjugate priors from the exponential family to enable the derivation of a Gibbs Sampler and specifically normal priors for centered perturbations. Finally, the perturbation intensity ϵ is a hyperparameter of the model.

$$\beta_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \qquad z_v \overset{\text{iid}}{\sim} \text{Cat}(\pi)$$
$$\rho_{i,k,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \qquad w_v \mid (\beta, \rho, z_{N(v)}, z_v) = \beta_{z_v} + \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{z_v, z_w}$$
$$\pi \overset{\text{iid}}{\sim} \text{Dirichlet}(\mathfrak{d}) \qquad x_{v,j} | w_v \overset{\text{iid}}{\sim} \mathcal{N}(w_{v,j}, S^2)$$

It is easy to see that the model is non identifiable: shifting all β by s and ρ by $-\frac{s}{\epsilon}$ leads to the same distribution over x. However, we think it is not an issue for our preliminary analysis and it could be overcame by introducing non linearities in the interactions, such as a zero-inflation with Bernoulli variables. The derivation of inference algorithms stays fairly similar in that case.

(a) Data of 5 clusters with $\epsilon = 0$ (no interaction)

(b) Same clusters but with interactions $\epsilon = 0.4$

Figure 3: Example of generated data and influence of epsilon (two first features). An algorithm ignoring the underlying graph structure would struggle to disentangle overlapping clusters.

3 Inference

3.1 Limits of probabilistic programming tools

We first tried a probabilistic programming approach with Pyro and PyMC3. To the best of our knowledge and after long experimentation with Pyro, we cannot marginalize discrete assignments that are not iid. For Pyro, the recommended approach for discrete variables is to enumerate in parallel[1] the possible values of each z_i. If the joint doesn't factorize then the enumeration is over n-uplets (exponentially large). However, the potential sparsity of the graph suggests a reduction of the enumeration complexity, but Pyro doesn't handle it. We propose an approach in 3.4.

3.2 Gibbs sampler

Derivation of the Gibbs Sampler We derived a Gibbs Sampler by computing the complete conditionals: the details can be found in the appendix. Here are some parameters updates:

$$n_i = \sum_v \mathbb{1}_{z_v=i}, \quad m_{v,k} = \sum_{w\in N(v)} \mathbb{1}_{z_w=k}, \quad \alpha_{i,j} = \sum_{v;z_v=i}^{n} (x_{v,j} - \tfrac{\epsilon}{|N(v)|} \sum_{w\in N(v)} \rho_{i,z_w,j}), \quad \eta_{i,k} = \frac{1}{\frac{S^2}{\sigma^2} + \epsilon^2 \sum_{v,z_v=i} \frac{m_{v,k}^2}{|N(v)|^2}}$$

$$\mu_{i,k,j} = \sum_{v,z_v=i} \frac{\epsilon m_{v,k}}{|N(v)|} \left(x_{v,j} - \beta_{i,j} - \tfrac{\epsilon}{|N(v)|} \sum_{w\in N(v);z_w\neq k} \rho_{i,z_w,j} \right)$$

$$\boxed{\pi \sim \text{Dirichlet}(f+n_1,\ldots,f+n_c)}, \quad \boxed{\boldsymbol{\beta}_{i,\bullet} \sim \mathcal{N}\left(\frac{\sigma^2}{S^2+n_i\sigma^2}\boldsymbol{\alpha}_{i,\bullet}, \frac{\sigma^2 S^2}{S^2+n_i\sigma^2} I_m \right)}, \quad \boxed{\rho_{i,k,j} \sim \mathcal{N}(\eta_{i,k}\mu_{i,k,j}, S^2\eta_{i,k})}$$

We can check that if $\epsilon = 0$, we recover the classic Gaussian mixture update. An important note is that z (resp. ρ) variables can't be updated all at once because the updates depend on the values of other z (resp. ρ).

Convergence analysis We ran the Gibbs sampler with the following procedure: start 20 Markov chains with different initializations (seeds), run a burn-in (50 epochs) and only keep the three chains with highest log joint to run 1000 epochs (or more). See figure 4. Then for each metric of interest, we sample multiple times from the chains and average the results into a Monte Carlo estimation.

3.3 Coordinate Ascent with Variational Inference

Derivation of CAVI We derive CAVI using the formula of [5] which leverage the calculations done for the Gibbs Sampling. If we set the variational family to be mean-field with factors in the same family as the corresponding complete conditional, then the optimal factor with all other factor fixed is given by:

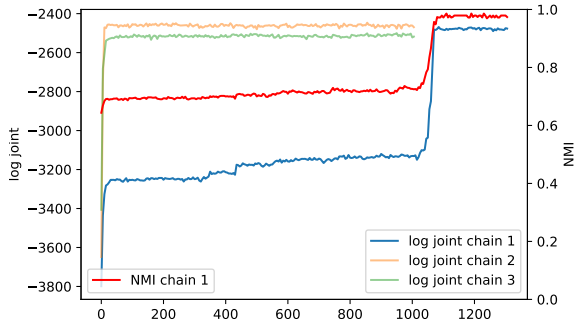
$$q(\theta) \propto \exp\left(\mathbb{E}_{-\theta}\left[\log p(\theta \mid \Theta \backslash \{\theta\}\right]\right)$$

We introduce the following variational parameters:

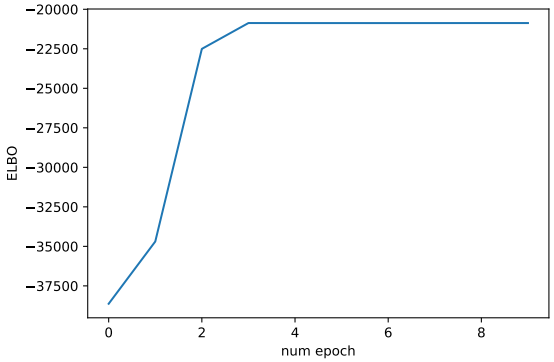
$$q(\pi_i) \sim \text{Dirichlet}(\delta), \quad q(z_v) \sim \text{Cat}(\phi_v), \quad q(\beta_{i,j}) \sim \mathcal{N}(\mu_{\beta_{i,j}}, \sigma_{\beta_{i,j}}), \quad q(\rho_{i,k,j}) \sim \mathcal{N}(\mu_{\rho_{i,k,j}}, \sigma_{\rho_{i,k,j}})$$

The computations were very tedious and we used optimized tensor sums `numpy.einsum(..., "optimal")` to have an incredible speed-up, leveraged to update independent variables in parallel. We computed the ELBO by decomposing it into KL divergence and reconstruction loss, and estimated the latter with MCMC of 100 samples.

[1] `https://pyro.ai/examples/enumeration.html`



(a) Log joint + constant, of 3 Gibbs samplers. The cluster NMI score (plotted for chain 1) is highly correlated to the joint.



(b) Convergence of the ELBO for CAVI

Figure 4: SBC histograms for w for the Gibbs sampler and CAVI

Convergence Analysis CAVI reaches convergence way faster than Gibbs sampling (15s compared to 2 minutes) but get stuck in local minima after a few iterations. We restarted the optimizer at multiple points. We are not surprised by this behavior as a Gaussian mixture model contains many local optimas due to discrete variables and CAVI cannot escape them as a deterministic algorithm. However, we believe our implementation might have a bug as the ELBO sometimes decreases for particular initializations.

3.4 Stochastic Variational Inference

An approach which will be faster and also probably avoid more local maxima is to add stochasticity with SVI. The objective function doesn't factorize entirely but we can still marginalize z_i out of the neighbor influence and write the ELBO as:

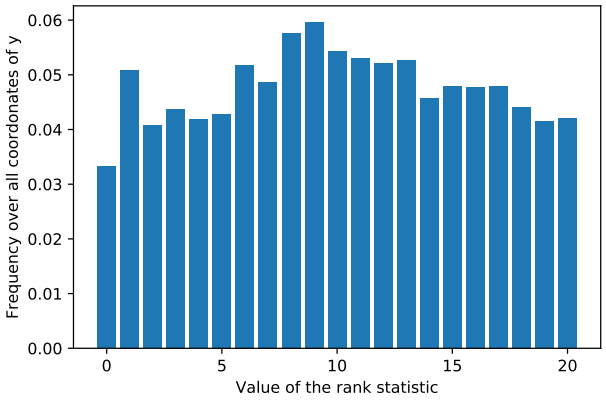
$$ELBO(q) = \sum_v \mathbb{E}_{q(z_v, z_{N(v)}, \Theta)} \left[\log p\left(x_v | z_v, z_{N(v)}, \Theta\right) \right] + \sum_v KL\left(q(z_v) \,||\, p(z_v)\right) + \sum_{\theta \in \Theta} KL\left(q(\theta) \,||\, p(\theta)\right)$$

In this form, the Rao-Blackwellisation of BBVI [6] is still promising to reduce the variance as the sampling space is reduced to a few neighbors in each expectation. Additionally, the reparametrization of the gaussian and the categorical (via a Gumbel trick) could lead to lower variance. We expect Pyro to be able to keep track of the neighbors dependency graph not with its probabilistic core but with its automatic differentiation core.

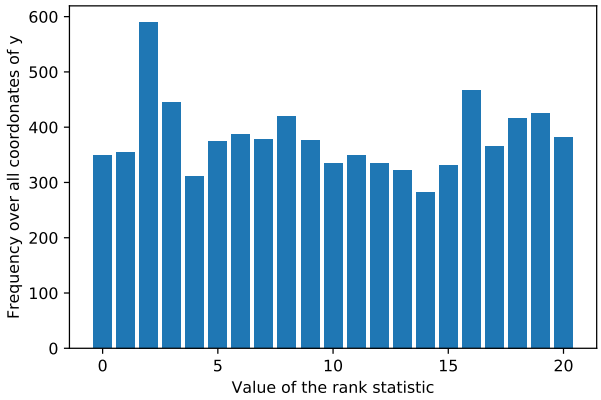
3.5 Analysis

SBC To check the consistency of our posteriors, we used Simulation-Based Calibration as suggested in [7]. For each inference method, we generate $\rho^{(l)}, \beta^{(l)}, z^{(l)}$ from the prior and generate $x^{(l)}$ accordingly. For each of these datasets, we sample $\rho_1^{(l)}, \beta_1^{(l)}, z_1^{(l)}, \ldots, \rho_L^{(l)}, \beta_L^{(l)}, z_L^{(l)}$ from the approximated posterior given by the inference methods. For a given latent variable (let's say β for clarity), for each coordinate we compute the rank statistic $r(\{(\beta_1^{(l)})_{i,k,j}, \ldots, (\beta_L^{(l)})_{i,k,j}\}, (\beta^{(l)})_{i,k,j})$ and average this quantity over all coordinates (i, k, j) and all runs l. If the posterior inference is not auto correlated, we expect to see a uniform histogram. That is what we observe. Especially, the distributions seems uniform for CAVI, which doesn't help to diagnose the suspected issue for this

algorithm. The histograms for statistic w are in figure 5. The other histograms can be found in the appendix.



(a) Histogram of rank statistics averaged over all coordinates of w for Gibbs sampling



(b) Histogram of rank statistics averaged over all coordinates of w for CAVI

Figure 5: SBC histograms for w for the Gibbs sampler and CAVI

4 Benchmark

We benchmark our model performance in different tasks, on synthetic data. We focused on the case where the observed clusters overlap to demonstrate the superiority of our model over a regular mixture of gaussians (in the case where clusters are well separated, the two models perform similarly). We generate data with our generative process and attempt to retrieve the clusters. The dependency graph $G = (V, E)$ is generated as a k-Nearest Neighbor graphs on randomly generated locations for the entities. We benchmark against K-Means, which is equivalent to our algorithm with $\epsilon = 0$. We run two benchmarks:

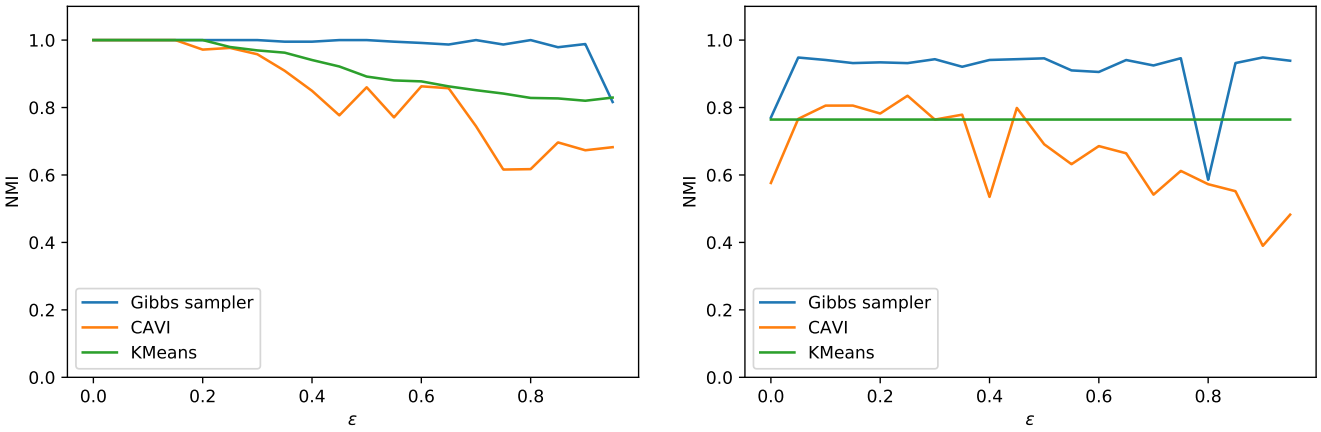
1. We generate data and run inference for multiple values of ϵ, using the same value of ϵ to generate the data and to infer it. This way we measure the robustness of the clustering with increasing perturbation.

2. For the second benchmark, we generate a fixed dataset ($\epsilon = 0.4$) and infer over this dataset with different values of ϵ (that may not correspond to the value which was used to generate the data). Especially, we check the ability to recover the clustering without knowing the value of ϵ.

The results of the benchmark, given by Figure 6, show that our model with Gibbs Sampling recovers clusters more accurately than without accounting for the perturbation and that we don't need to know the value of ϵ to be able to correctly fit the data. Our inference method with CAVI seems to be struggling with local minima.

5 Criticism

Our model is very constrained which has the advantage of having interpretable data (meaningful clusters), but lead to a tedious inference procedure with the discrete latent variables. Also, it



(a) Evolution of clusters recovery when increasing interaction

(b) Fixed dataset with $\epsilon = 0.4$; clusters recovery with various values of ϵ

Figure 6: Benchmark of the cluster recovery, using Normalized Mutual Information

may be hard in practice to access the "true" interaction graph, which would need to be set as prior knowledge. In this case, it may be necessary to formulate hypothesis on the structure of the graph (the same way we formulate hypothesis on the structure of the distributions of the model). Even when knowing such a graph (like social networks), it is not clear whether two elements are connected because they are similar (which corresponds to the first independence hypothesis) or if they are similar because they are connected (which corresponds to the second independence hypothesis). Finally, the model is not completely identifiable, which doesn't allow us to recover exactly the perturbations.

6 Conclusion

We introduce a model that accounts for non iid mixtures, where each entity has an independent hidden state and the observation is a perturbed version of this hidden state controlled by its neighbors. We show that this model outperforms classic mixture of gaussians when recovering clusters from perturbed data. The model can compensate a mispecified ϵ by scaling accordingly the values of the latent variables β and ρ. Randomly activating or deactivating some perturbations to break the linearity would enable to recover latent variables.

The code can be found at `https://github.com/ANazaret/correlations`

References

[1] Olivier François, Sophie Ancelet, and Gilles Guillot. Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174(2):805–816, 2006.

[2] Jonathan Chang and David Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

[3] Da Tang, Dawen Liang, Tony Jebara, and Nicholas Ruozzi. Correlated variational auto-encoders, 2019.

[4] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235, 2019.

[5] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Feb 2017.

[6] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. *arXiv preprint arXiv:1401.0118*, 2013.

[7] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

A Gibbs sampler derivation

We recall the joint:

$$\mathbb{P}\left[x, z, \rho, \beta, \pi\right] \propto \mathbb{P}\left[\pi\right] \prod_{v=1}^{n} \mathbb{P}\left[z_v \mid \pi\right] \prod_{i=1}^{c} \prod_{j=1}^{m} \mathbb{P}\left[\beta_{i,j}\right] \prod_{i=1}^{c} \prod_{k=1}^{c} \prod_{j=1}^{m} \mathbb{P}\left[\rho_{i,k,j}\right] \prod_{v=1}^{n} \prod_{j=1}^{m} \mathbb{P}\left[x_{v,j} \mid z, \beta, \rho\right]$$

$$\propto \mathbb{P}\left[\pi\right] \prod_{v=1}^{n} \pi_{z_v} \prod_{i=1}^{c} \prod_{j=1}^{m} e^{-\frac{\beta_{i,j}^2}{\sigma^2}} \prod_{i=1}^{c} \prod_{k=1}^{c} \prod_{j=1}^{m} e^{-\frac{\rho_{i,k,j}^2}{\sigma^2}} \prod_{v=1}^{n} \prod_{j=1}^{m} e^{-\frac{(x_{v,j} - \beta_{z_v,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{z_v,z_w,j})^2}{S^2}}$$

We write: $n_i = \sum_v \mathbb{1}\left\{z_v = i\right\}$

Derivation of $\pi \mid z$

$$\boxed{\pi \sim \text{Dirichlet}(f + n_1, \ldots, f + n_c)}$$

Derivation of $\beta_{i,j} \mid x, z, \rho, \beta_{-(i,j)}$

$$\Pr[\beta_{i,j} | x, z, \rho, \beta_{-(i,j)}, \pi] \propto \Pr[x, z, \rho, \beta, \pi]$$

$$\propto e^{-\frac{\beta_{i,j}^2}{2\sigma^2}} \prod_{v; z_v = i}^{n} e^{-\frac{(x_{i,j} - \beta_{i,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{i,z_w,j})^2}{2S^2}}$$

Switching to log notation

$$-\log[\Pr[\beta_{i,j}] \;\dot{\propto}\; \frac{\beta_{i,j}^2}{2\sigma^2} + \sum_{v; z_v = i}^{n} \frac{(x_{i,j} - \beta_{i,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{i,z_w,j})^2}{2S^2}$$

$$\dot{\propto}\; \beta_{i,j}^2 \left[\frac{1}{2\sigma^2} + \frac{n_i}{2S^2}\right] - 2\beta_{i,j} \sum_{v; z_v = i}^{n} \frac{(x_{i,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{i,z_w,j})}{2S^2}$$

Using $\alpha_{i,j} = \sum_{v; z_v = i}^{n} (x_{v,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{i,z_w,j})$, we have:

$$-\log \Pr[\beta_{i,j} | x, z, \rho, \beta_{-(i,j)}, \pi] \;\dot{\propto}\; \frac{S^2 + n_i\sigma^2}{2\sigma^2 S^2} [\beta_{i,j}^2 - 2\beta_{i,j} \frac{\sigma^2}{S^2 + n_i\sigma^2} \alpha_{i,j}]$$

Therefore:

$$\beta_{i,j} \mid x, z, \rho, \beta_{-(i,j)}, \pi \sim \mathcal{N}\left(\frac{\sigma^2}{S^2 + n_i\sigma^2} \alpha_{i,j}, \frac{\sigma^2 S^2}{S^2 + n_i\sigma^2}\right)$$

We deal with vector of normal random variables so $\beta_{i,j}$ can be compacted into a random vector:

$$\boxed{\boldsymbol{\beta}_{i,\bullet} \mid x, z, \rho, \beta_{-i}, \pi \sim \mathcal{N}\left(\frac{\sigma^2}{S^2 + n_i\sigma^2} \boldsymbol{\alpha}_{i,\bullet}, \frac{\sigma^2 S^2}{S^2 + n_i\sigma^2} I_m\right)}$$

Derivation of $\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)}, \beta$

$$
\begin{aligned}
-\log \mathbb{P}\left[\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)}\right] &\dot{\propto} -\log \mathbb{P}\left[x, z, \rho, \beta, \pi\right] \\
&\dot{\propto} \frac{\rho_{i,k,j}^2}{2\sigma^2} + \frac{1}{2S^2} \sum_{v, z_v = i} \left(\beta_{i,j} - x_{v,j} + \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{i,z_w,j} \right)^2
\end{aligned}
$$

Defining $m_{v,k} = \sum_{w \in N(v)} \mathbb{1}\{z_w = k\}$

$$
\begin{aligned}
-\log \mathbb{P}\left[\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)}\right] &\dot{\propto} \left(\frac{1}{2\sigma^2} + \frac{\epsilon^2}{2S^2} \sum_{v, z_v = i} \frac{m_{v,k}^2}{|N(v)|^2} \right) \rho_{i,k,j}^2 \\
&\quad + 2\rho_{i,k,j} \frac{1}{2S^2} \sum_{v, z_v = i} \frac{\epsilon m_{v,k}}{|N(v)|} \left(\beta_{i,j} - x_{v,j} + \frac{\epsilon}{|N(v)|} \sum_{w \in N(v); z_w \neq k} \rho_{i,z_w,j} \right)
\end{aligned}
$$

Setting:

$$
\begin{aligned}
\eta_{i,k} &= \frac{1}{\frac{S^2}{\sigma^2} + \epsilon^2 \sum_{v, z_v = i} \frac{m_{v,k}^2}{|N(v)|^2}} \\
\mu_{i,k,j} &= \sum_{v, z_v = i} \frac{\epsilon m_{v,k}}{|N(v)|} \left(x_{v,j} - \beta_{i,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v); z_w \neq k} \rho_{i,z_w,j} \right)
\end{aligned}
$$

Then we have:

$$
-\log \mathbb{P}\left[\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)}\right] \dot{\propto} \frac{1}{2S^2 \eta_{i,k}} \left[\rho_{i,k,j}^2 - 2\rho_{i,k,j} \eta_{i,k} \mu_{i,k,j} \right]
$$

Therefore:

$$
\boxed{\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)} \sim \mathcal{N}(\eta_{i,k} \mu_{i,k,j}, S^2 \eta_{i,k})}
$$

$$
\eta_{i,k} = \frac{1}{\frac{S^2}{\sigma^2} + \epsilon^2 \sum_{v, z_v = i} \frac{m_{v,k}^2}{|N(v)|^2}} \quad , \quad \boldsymbol{\mu}_{i,k,\bullet} = \sum_{v, z_v = i} \frac{\epsilon m_{v,k}}{|N(v)|} \left(\boldsymbol{x}_{v,\bullet} - \boldsymbol{\beta}_{i,\bullet} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v); z_w \neq k} \boldsymbol{\rho}_{i,z_w,\bullet} \right)
$$

$$
\boxed{\boldsymbol{\rho}_{i,k,\bullet} \mid x, z, \rho_{-(i,k,\bullet)} \sim \mathcal{N}\left(\eta_{i,k} \boldsymbol{\mu}_{i,k,\bullet} , S^2 \eta_{i,k} I_m \right)}
$$

No correlation: $\epsilon = 0$ $\eta_{i,k} = \frac{\sigma^2}{S^2}$ and $\mu_{i,k,j} = 0$: we get the prior back.
m_{kv} uniform gives $\frac{m_{kv}}{|N(v)|} = \frac{1}{c}$ then for $c \to \infty$ we have no interaction (cancels out).

Derivation of $z_v \mid x, z_{-v}, \rho, \beta$

$$
\begin{aligned}
-\log \mathbb{P}\left[z_v \mid x, z_{-v}, \rho, \beta\right] &\dot{\propto} -\log \Pr[x, z, \rho, \beta, \pi] \\
-\log \mathbb{P}\left[z_v \mid x, z_{-v}, \rho, \beta\right] + \log(\pi_{z_v}) &\dot{\propto} \sum_{j=1}^{m} \left(\beta_{z_v,j} + \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{z_v,z_w,j} - x_{v,j} \right)^2 \\
&\quad + \sum_{w \in N_{out}(v)} \sum_{j=1}^{m} \left(\beta_{z_w,j} + \frac{\epsilon}{|N(w)|} \sum_{u \in N(w)} \rho_{z_w,z_u,j} - x_{w,j} \right)^2
\end{aligned}
$$

In a nutshell:

$$
n_i = \sum_v \mathbb{1}\{z_v = i\} \qquad \alpha_{i,j} = \sum_{v; z_v = i}^{n} (x_{v,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v)} \rho_{i,z_w,j})
$$

$$
\eta_{i,k} = \frac{1}{\frac{S^2}{\sigma^2} + \epsilon^2 \sum_{v, z_v = i} \frac{m_{v,k}^2}{|N(v)|^2}} \qquad \mu_{i,k,j} = \sum_{v, z_v = i} \frac{\epsilon m_{v,k}}{|N(v)|} \left(x_{v,j} - \beta_{i,j} - \frac{\epsilon}{|N(v)|} \sum_{w \in N(v); z_w \neq k} \rho_{i,z_w,j} \right)
$$

$$
\boxed{\pi \sim \text{Dirichlet}(f + n_1, \ldots, f + n_c)}
$$

$$
\boxed{\boldsymbol{\beta}_{i,\bullet} \mid x, z, \rho, \beta_{-i}, \pi \sim \mathcal{N}\left(\frac{\sigma^2}{S^2 + n_i \sigma^2} \boldsymbol{\alpha}_{i,\bullet}, \frac{\sigma^2 S^2}{S^2 + n_i \sigma^2} I_m \right)}
$$

$$
\boxed{\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)} \sim \mathcal{N}(\eta_{i,k} \mu_{i,k,j}, S^2 \eta_{i,k})}
$$

B CAVI derivation

We derive CAVI using the formula of [5] which leverage the calculations done for the Gibbs Sampling. We can take all expressions and replace introduced quantities with expectations. Some quantities are denominators of fractions or products, but it is due to factorization and thus we can simply expand to check that it is correct to take expectation.
We recall:

$$
\begin{aligned}
-\log \mathbb{P}\left[\beta_{i,j} | x, z, \rho, \beta_{-(i,j)}, \pi\right] &\dot{\propto} \frac{S^2 + n_i \sigma^2}{2\sigma^2 S^2} [\beta_{i,j}^2 - 2\beta_{i,j} \frac{\sigma^2}{S^2 + n_i \sigma^2} \alpha_{i,j}] \\
-\log \mathbb{P}\left[\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)}\right] &\dot{\propto} \frac{1}{2S^2 \eta_{i,k}} \left[\rho_{i,k,j}^2 - 2\rho_{i,k,j} \eta_{i,k} \mu_{i,k,j} \right]
\end{aligned}
$$

and taking the expectation gives:

$$
-\mathbb{E}_{-\beta_{i,j}} \left[\log \mathbb{P}\left[\beta_{i,j} | x, z, \rho, \beta_{-(i,j)}, \pi\right]\right] \dot{\propto} \frac{S^2 + \overline{n_i} \sigma^2}{2\sigma^2 S^2} \left[\beta_{i,j}^2 - 2\beta_{i,j} \frac{\sigma^2}{S^2 + \overline{n_i} \sigma^2} \overline{\alpha_{i,j}} \right] \tag{1}
$$

$$
-\mathbb{E}_{-\rho_{i,k,j}} \left[\log \mathbb{P}\left[\rho_{i,k,j} \mid x, z, \rho_{-(i,k,j)}\right]\right] \dot{\propto} \frac{1}{2S^2 \overline{\eta_{i,k}}} \left[\rho_{i,k,j}^2 - 2\rho_{i,k,j} \overline{\eta_{i,k}}\ \overline{\mu_{i,k,j}} \right] \tag{2}
$$

Such that the optimal variational parameters are:

$$(\mu_{\beta_{i,j}}, \sigma_{\beta_{i,j}}) = \left(\frac{\sigma^2}{S^2 + \overline{n_i}\sigma^2} \overline{\alpha_{i,j}}, \frac{\sigma^2 S^2}{S^2 + \overline{n_i}\sigma^2} \right) \tag{3}$$

$$(\mu_{\rho i,k,j}, \sigma_{\rho_{i,k,j}}) = \left(\overline{\eta_{i,k}}\ \overline{\mu_{i,k,j}}, S^2 \overline{\eta_{i,k}} \right) \tag{4}$$

$$\tag{5}$$

where:

$$\overline{n_i} = \mathbb{E}\left[n_i\right] = \sum_v \phi_{v,i} \tag{6}$$

$$\overline{\alpha_{i,j}} = \mathbb{E}\left[\alpha_{i,j}\right] \tag{7}$$

$$= \mathbb{E}\left[\sum_v^n \mathbb{1}\left\{z_v = i\right\} \left(x_{v,j} - \epsilon' \sum_{w \in N(v)} \rho_{i,z_w,j} \right) \right] \tag{8}$$

$$\overset{z \perp\!\!\!\perp \rho}{=} \sum_v^n \phi_{vi} \left(x_{v,j} - \epsilon' \sum_{w \in N(v)} \sum_{k=1}^c \phi_{wk} \mathbb{E}_\rho\left[\rho_{i,k,j}\right] \right) \tag{9}$$

$$= (\phi^\top x)_{ij} - \epsilon' \sum_{k=1}^c \mu(\rho_{i,k,j}) \sum_v \sum_{w \in N(v)} \phi_{vi}\phi_{wk} \tag{10}$$

$$= (\phi^\top x)_{ij} - \epsilon' \sum_{k=1}^c \mu(\rho_{i,k,j}) (\phi^T A \phi)_{ik} \tag{11}$$

$$\tag{12}$$

$$\overline{m_{vk}} = \sum_{w \in N(v)} \phi_{wk} = (A\phi)_{vk} \tag{13}$$

$$\overline{m_{vk}^2} = \left(\sum_{w \in N(v)} \phi_{w,k} \right)^2 + \sum_{w \in N(v)} \phi_{wk}(1 - \phi_{wk}) = (\overline{m_{vk}})^2 + (A(\phi \odot (1 - \phi)))_{vk} \tag{14}$$

$$1/\overline{\eta_{i,k}} = \frac{S^2}{\sigma^2} + \epsilon'^2 \sum_v \mathbb{E}\left[\mathbb{1}\left\{z_v = i\right\} m_{vk}^2\right] \tag{15}$$

$$\overset{z_v \perp\!\!\!\perp z_w}{=} \frac{S^2}{\sigma^2} + \epsilon'^2 \sum_v \phi_{vi} \overline{m_{vk}^2} \quad = \quad S^2/\sigma^2 + \epsilon'^2 \left(\phi^\top \overline{m^2} \right)_{ik} \tag{16}$$

$$\overline{\mu_{i,k,j}} = \mathbb{E}\left[\epsilon' \sum_v \mathbb{1}\{z_v = i\} m_{v,k}\left(x_{v,j} - \beta_{i,j} - \epsilon' \sum_{w \in N(v)} \mathbb{1}\{z_w \neq k\} \rho_{i,z_w,j}\right)\right] \tag{17}$$

$$= \epsilon' \sum_v \phi_{vi}\left(\overline{m_{v,k}} x_{v,j} - \overline{m_{v,k}}\mu(\beta_{i,j}) - \epsilon' \mathbb{E}\left[m_{v,k} \sum_{w \in N(v)} \mathbb{1}\{z_w \neq k\} \rho_{i,z_w,j}\right]\right) \tag{18}$$

$$\overline{\mu_{i,k,j}}/\epsilon' = \underbrace{\left(\sum_v \phi_{vi}\overline{m_{v,k}} x_{v,j}\right)}_{a_{i,k,j}} - \underbrace{\left(\sum_v \phi_{vi}\overline{m_{v,k}}\mu(\beta_{i,j})\right)}_{b_{i,k,j}} - \epsilon' \underbrace{\sum_v \phi_{vi}\mathbb{E}\left[m_{v,k} \sum_{w \in N(v)} \mathbb{1}\{z_w \neq k\} \rho_{i,z_w,j}\right]}_{c_{i,k,j}} \tag{19}$$

Computation of $c_{i,k,j}$

$$c_{i,k,j} = \sum_v \phi_{vi} \sum_{w' \in N(v)} \sum_{w \in N(v)} \mathbb{E}\left[\mathbb{1}\{z_{w'} = k\} \mathbb{1}\{z_w \neq k\} \rho_{i,z_w,j}\right]$$
$$= \sum_v \phi_{vi} \sum_{w' \in N(v)} \phi_{w',k} \sum_{w \neq w'} \sum_{l \neq k} \phi_{w,l}\mu(\rho_{i,l,j})$$

Recall $B = A\phi$

$$= \sum_{l \neq k} \mu(\rho_{i,l,j}) \sum_v \phi_{vi} \sum_{w' \in N(v)} \phi_{w',k}\left(\sum_{w \in N(v)} \phi_{w,l} - \phi_{w',l}\right)$$
$$= \sum_{l \neq k} \mu(\rho_{i,l,j}) \sum_v \phi_{vi} \sum_{w' \in N(v)} \phi_{w',k} \sum_{w \in N(v)} \phi_{w,l} - \sum_{l \neq k} \mu(\rho_{i,l,j}) \sum_v \phi_{vi} \sum_{w' \in N(v)} \phi_{w',k}\phi_{w',l}$$
$$= \sum_{l \neq k} \mu(\rho_{i,l,j}) \sum_v \phi_{vi}(A\phi)_{vk}(A\phi)_{vl} - \sum_{l \neq k} \mu(\rho_{i,l,j}) \sum_v \phi_{vi} \sum_{w' \in N(v)} \phi_{w',k}\phi_{w',l}$$
$$= \sum_l \sum_v \mu(\rho_{i,l,j})\phi_{vi}B_{vk}B_{vl}\mathbb{1}\{l \neq k\} - \sum_l \sum_v \sum_w \mu(\rho_{i,l,j})A_{vw}\phi_{vi}\phi_{w,k}\phi_{w,l}\mathbb{1}\{l \neq k\}$$

Using

$$\sum_v \sum_{w \in N(v)} \phi_{vi}\phi_{wj} = (\phi^T A\phi)_{ij}$$

At line 17, m_{vk} and $\rho_{i,z_w,j}$ are not independent.

The calculations for the latent variables, cornerstone of dependence of our model, are more tedious. We'll write $\epsilon' = \epsilon/|N(v)|$. We recall,

$$\log \mathbb{P}\left[z_v \mid x, z_{-v}, \rho, \beta\right] \;\dot{\propto}\; \log(\pi_{z_v}) - \sum_{j=1}^{m}\left(\beta_{z_v,j} + \frac{\epsilon}{|N(v)|}\sum_{w\in N(v)}\rho_{z_v,z_w,j} - x_{v,j}\right)^2 - \sum_{w\in N_{out}(v)}\sum_{j=1}^{m}\left(\beta_{z_w,j} + \frac{\epsilon}{|N(w)|}\sum_{u\in N(w)}\rho_{z_w,z_u,j} - x_{w,j}\right)^2$$

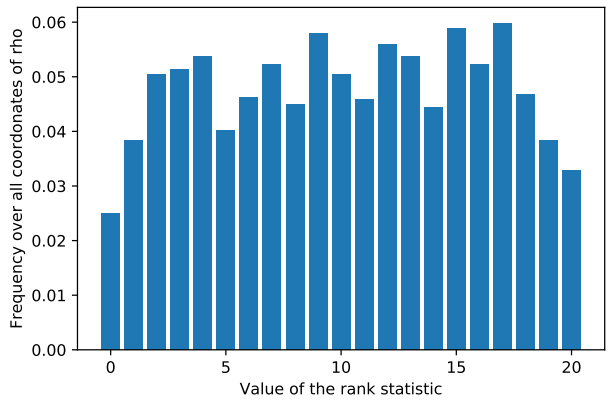
For $z_v = k$ we have:

$$\mathbb{E}\left[\log \pi_k\right] = \psi(\delta_k) - \psi\left(\textstyle\sum_l \delta_l\right)$$

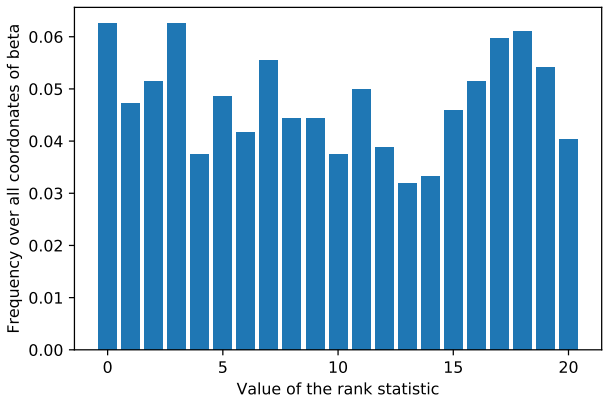
$$\begin{aligned}\mathbb{E}\left[\left(\beta_{z_v,j} - x_{v,j} + \epsilon' \textstyle\sum_{w\in N(v)} \rho_{z_v,z_w,j}\right)^2\right] &= \mathbb{E}\left[(\beta_{z_v,j} - x_{v,j})^2\right] + 2\epsilon'\mathbb{E}\left[(\beta_{z_v,j} - x_{v,j})\sum_{w\in N(v)}\rho_{z_v,z_w,j}\right] \\ &\quad + \epsilon'^2\mathbb{E}\left[\sum_{w,w'\in N(v)}\rho_{k,z_w,j}\rho_{k,z_{w'},j}\right] \\ &= \sigma^2_{\beta_{kj}} + \left(\mu_{\beta_{kj}} - x_{vj}\right)^2 \\ &\quad + 2\epsilon'\left(\mu_{\beta_{kj}} - x_{vj}\right)\sum_{w\in N(v)}\phi_w^\top \mu_{\rho_{k,\bullet,j}} \\ &\quad + \epsilon'^2\sum_{w\in N(v)}\left[\phi_w^\top(\mu^2_{\rho_{k,\bullet,j}} + \sigma^2_{\rho_{k,\bullet,j}}) - (\phi_w^\top\mu_{\rho_{k,\bullet,j}})^2\right] + \epsilon'^2\left(\sum_{w\in N(v)}\phi_w^\top\mu_{\rho_{k,\bullet,j}}\right)^2\end{aligned}$$

Using $\phi'_{v,i} = \mathbb{1}\{i = k\}$ in the previous formula, we can easily compute the second sum and finish the computation.

C SBC histograms

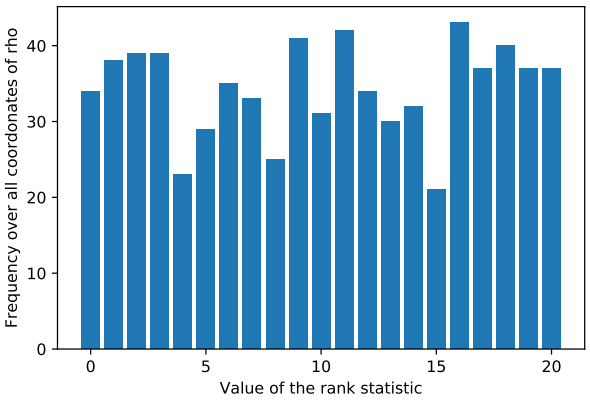


(a) Histogram of rank statistics averaged over all coordinates of ρ

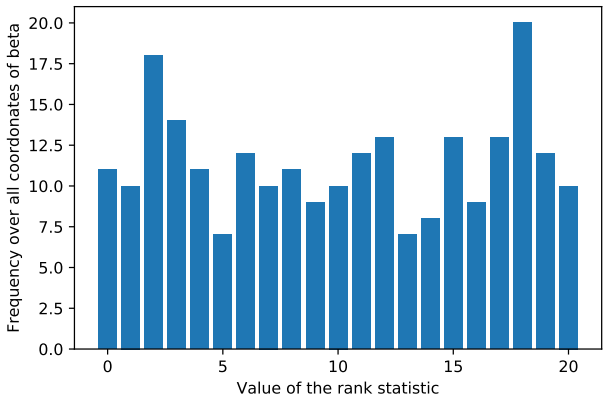


(b) Histogram of rank statistics averaged over all coordinates of β

Figure 7: SBC histograms for β, ρ for the Gibbs sampler



(a) Histogram of rank statistics averaged over all coordinates of ρ



(b) Histogram of rank statistics averaged over all coordinates of β

Figure 8: SBC histograms for β, ρ for CAVI