Supplementary Material: General Table Completion using a Bayesian Nonparametric Model

Isabel Valera
Department of Signal Processing
and Communications
University Carlos III in Madrid
`ivalera@tsc.uc3m.es`

Zoubin Ghahramani
Department of Engineering
University of Cambridge
`zoubin@eng.cam.ac.uk`

1 Accelerated Gibbs Sampler

In [1], the authors presented a linear-time accelerated Gibbs sampler for conjugate IBP models that effectively marginalized over the latent factors. The per-iteration complexity of this algorithm is $\mathcal{O}(N(K^2 + KD))$, which is comparable to the uncollapsed linear-Gaussian IBP sampler that has per-iteration complexity $\mathcal{O}(NDK^2)$ but does not marginalize over the weighting factors, and as a result, presents slower convergence rate.

This algorithm exploits the Bayes rule to avoid the cubic complexity with N due to the computation of the marginal likelihood in the Collapsed Gibbs sampler. In particular, it applies the Bayes rule to obtain the probability of each element in the latent feature matrix $\mathbf{Z}$ being active as

$$p(z_{nk} = 1 | \{\mathbf{Y}^d\}_{d=1}^D, \mathbf{Z}_{\neg nk}) \propto \frac{m_{\neg n,k}}{N} \prod_{d=1}^{D} \prod_{r=1}^{S_d} \int_{\mathbf{b}_r^d} p(y_{nr}^d | \mathbf{z}_n, \mathbf{b}_r^d) p(\mathbf{b}_r^d | \mathbf{y}_{\neg nr}^d \mathbf{Z}_{\neg n}) d\mathbf{b}_r^d, \tag{1}$$

where S_d is the number of columns in matrices $\mathbf{Y}^d$ and $\mathbf{B}^d$ (being S_d the number of categories R_d for those dimension d that contains categorical attributes, and $S_d = 1$ otherwise), $\mathbf{Z}_{\neg n}$ corresponds to matrix $\mathbf{Z}$ after removing the the n-th row, the vector $\mathbf{y}_{\neg nr}^d$ is the $r-$th column of matrix $\mathbf{Y}^d$ without the element y_{nr}^d, and $p(\mathbf{b}_r^d | \mathbf{x}_{\neg n}^d, \mathbf{Z}_{\neg n})$ is the posterior of $\mathbf{b}_r^d$ computed without taking the n-th datapoint into account, i.e.,

$$p(\mathbf{b}_r^d | \mathbf{y}_{\neg nr}^d, \mathbf{Z}_{\neg n}) = \mathcal{N}(\mathbf{b}_r^d | \mathbf{P}_{\neg n}^{-1} \boldsymbol{\lambda}_{\neg nr}^d, \mathbf{P}_{\neg n}^{-1}), \tag{2}$$

where $\mathbf{P}_{\neg n} = \mathbf{Z}_{\neg n}^\top \mathbf{Z}_{\neg n} + 1/\sigma_B^2 \mathbf{I}_K$ and $\boldsymbol{\lambda}_{\neg ny}^d = \mathbf{Z}_{\neg n}^\top \mathbf{y}_{\neg nr}^d$ are the natural parameters of the Gaussian distribution.

Note that, opposite to the notation in [1], we here resort to the natural parameters for the Gaussian distribution over the posterior of $\mathbf{b}_r^d$ instead of the mean and the covariance matrix. This formulation allows us to compute the full posterior over the weighting factors as

$$p(\mathbf{b}_r^d | \mathbf{y}_r^d, \mathbf{Z}) = \mathcal{N}(\mathbf{b}_r^d | \mathbf{P}^{-1} \boldsymbol{\lambda}_r^d, \mathbf{P}^{-1}), \tag{3}$$

where $\mathbf{P} = \mathbf{P}_{\neg n} + \mathbf{z}_n^\top \mathbf{z}_n$ and $\boldsymbol{\lambda}_r^d = \boldsymbol{\lambda}_{\neg nr}^d + \mathbf{z}_n^\top y_{nr}^d$ are the natural parameters of the Gaussian distribution.

The Accelerated Gibbs sampling algorithm iteratively samples the value of each element z_{nk} according to

$$p(z_{nk} = 1 | \{\mathbf{Y}^d\}_{d=1}^D, \mathbf{Z}_{\neg nk}) \propto \frac{m_{\neg n,k}}{N} \prod_{d=1}^{D} \prod_{r=1}^{S_d} \mathcal{N}(y_{nr}^d | \mathbf{z}_n \boldsymbol{\lambda}_{\neg nr}^d, \mathbf{z}_n \mathbf{P}_{\neg n} \mathbf{z}_n^\top + \sigma_y^2). \tag{4}$$

After having sampled all elements z_{nk} for the K_+ non-zero columns in $\mathbf{Z}$ for each data point n, the algorithm samples from a distribution (where the prior is a Poisson distribution with mean α/N) a number of new features necessary to explain that data point.

2 Posterior distribution over $\mathbf{Y}^d$

As previously described, in the 5-th step of Algorithm **??**, we need to sample from the auxiliary Gaussian variables y^d_{nr} from the posterior distribution $p(y^d_{nr}|x^d_n, \mathbf{z}_n, \mathbf{b}^d)$. The posterior distribution y^d_{nr} for all the considered types of data are given given by:

1. For real-valued observation:

$$p(y^d_{n1}|x^d_n, \mathbf{z}_n, \mathbf{B}^d) = \delta(x^d_n) \tag{5}$$

2. For positive real-valued observations:

$$\begin{aligned} &p(y^d_{n1}|x^d_n, \mathbf{z}_n, \mathbf{B}^d) \\ &= \mathcal{N}\left(y^d_{n1}\middle| \left(\frac{(\mathbf{z}_n\mathbf{b}^d_1)}{\sigma^2_y} + \frac{f^{-1}(x^d_n)}{\sigma^2_u}\right)\left(\frac{1}{\sigma^2_y} + \frac{1}{\sigma^2_u}\right)^{-1}, \left(\frac{1}{\sigma^2_y} + \frac{1}{\sigma^2_u}\right)^{-1}\right). \end{aligned} \tag{6}$$

3. For categorical observations:

$$\begin{aligned} &p(y^d_{nr}|x^d_n = T, \mathbf{z}_n, \mathbf{B}^d) \\ &= \left\{ \begin{array}{ll} \mathcal{N}(y^d_{nr}|\mathbf{z}_n\mathbf{b}^d_r, \sigma^2_y)\mathbb{I}(y^d_{nr} > \max_{j\neq r}(y^d_{nj})) & \text{If} \quad r = T \\ \mathcal{N}(y^d_{nr}|\mathbf{z}_n\mathbf{b}^d_r, \sigma^2_y)\mathbb{I}(y^d_{nr} < y^d_{nT}) & \text{If} \quad r \neq T \end{array} \right. \end{aligned} \tag{7}$$

   In words, if $x^d_n = T = r$ we sample y^d_{nr} from a Gaussian truncated by the left by $\max_{j\neq r}(y^d_{nj})$ and, otherwise, we sample from a Gaussian truncated by the right by y^d_{nr} with $r = x^d_n$. Note that sampling from the variables y^d_{nr} corresponds to solve a multinomial probit regression problem. To achieve identifiability we assume, without loss of generality, that the regression function $f_{R_d}(\mathbf{z}_n)$ is identically zero, and therefore, we fix $b^d_{kR_d} = 0$ for all k.

4. For ordinal observations:

$$p(y^d_{n1}|x^d_n = r, \mathbf{z}_n, \mathbf{B}^d) \sim \mathcal{N}(y^d_{n1}|\mathbf{z}_n\mathbf{b}^d_1, \sigma^2_y)\mathbb{I}(\theta^d_{r-1} < y^d_{n1} \leq \theta^d_r). \tag{8}$$

   Note that in this case, we also need to sample the values for the thresholds θ^d_r with $r = 1, \ldots, R_d - 1$ as

$$\begin{aligned} p(\theta^d_r|y^d_{n1}) \sim &\mathcal{N}(\theta^d_r|0, \sigma^2_\theta)\mathbb{I}(\theta^d_r > \max(\theta^d_{r-1}, \max_n(y^d_{n1}|x^d_n = r)) \\ &\times \mathbb{I}(\theta^d_r < \min(\theta^d_r, \min_n(y^d_{n1}|x^d_n = r + 1)). \end{aligned} \tag{9}$$

   In this case, sampling from the variables y^d_{n1} corresponds to solve an ordered probit regression problem, where the thresholds $\{\theta_r\}^{R_d}_{r=1}$ are unknown. Hence, to achieve identifiability we need to set the one of the thresholds, θ_1 in our case, to a fixed value.

5. For count observations:

$$p(y^d_{n1}|x^d_n, \mathbf{z}_n, \mathbf{B}^d) = \mathcal{N}(y^d_{n1}|\mathbf{z}_n\mathbf{b}^d_1, \sigma^2_y)\mathbb{I}(f^{-1}(x^d_n) \leq y^d_{n1} < f^{-1}(x^d_n + 1)), \tag{10}$$

   where $f^{-1} : \Re_+ \rightarrow \Re$ is the inverse function of f, i.e., $f^{-1}(f(y)) = y$. Therefore, y^d_{n1} from a Gaussian truncated by the left by $f^{-1}(x^d_n)$ and by the right by $f^{-1}(x^d_n + 1)$.

References

[1] F. Doshi-Velez and Z. Ghahramani. Accelerated sampling for the indian buffet process. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 273–280, New York, NY, USA, 2009. ACM.