Doubly Non-Central Beta Matrix Factorization for DNA Methylation Data (Supplementary material)

Aaron Schein[1] **Anjali Nagulpally**[2] **Hanna Wallach**[3] **Patrick Flaherty**[2]

[1]Data Science Institute, Columbia University
[2]Department of Mathematics and Statistics, University of Massachusetts Amherst
[3]Microsoft, New York City, NY

A THE DNCB DISTRIBUTION

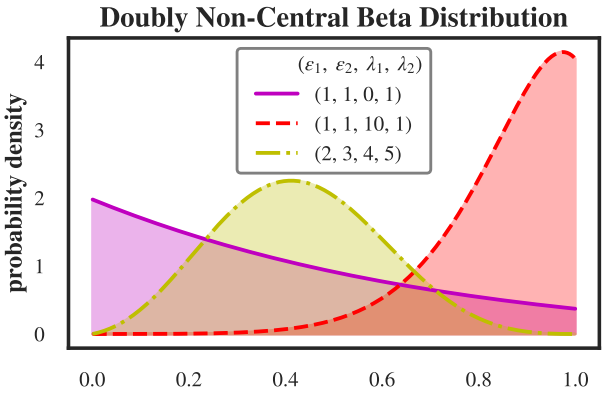
The DNCB distribution is defined in definition 1 of the main paper. It can take the same set of shapes over the $(0, 1)$ interval as the beta distribution (see fig. 1a), as well as tri-modal shapes when the shape parameters $\epsilon_1, \epsilon_2 < 1$ (see fig. 1b).

Ongaro and Orsi [2015] provide a general formula for the moments of the DNCB distribution. Its first moment is

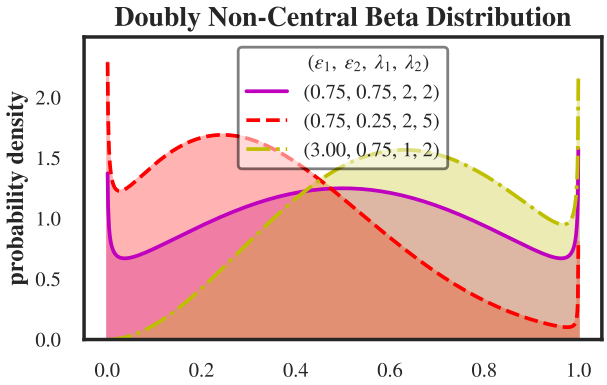
$$\mathbb{E}[\beta] = \frac{\epsilon_1}{\epsilon_\bullet e^{\lambda_\bullet}} \left[{}_1F_1\left(\epsilon_\bullet; \epsilon_\bullet{+}1; \lambda_\bullet\right) + \frac{\epsilon_\bullet \lambda_1}{\epsilon_1(\epsilon_\bullet{+}1)} {}_1F_1\left(\epsilon_\bullet{+}1; \epsilon_\bullet{+}2; \lambda_\bullet\right) \right]$$

where ${}_1F_1(\cdot;\cdot;\cdot)$ denotes Kummer's confluent hypergeometric function. The second moment is more involved, but also does not involve any special functions beyond ${}_1F_1(\cdot;\cdot;\cdot)$.

Computing the mean and variance of the DNCB is easy because there are many efficient open-source implementations of ${}_1F_1(\cdot;\cdot;\cdot)$—e.g., in the Python library scipy [Virtanen et al., 2020]. On the other hand, computing the DNCB density, which we need to assess out-of-sample predictive performance (see section 5 of the main paper), requires computing Humbert's confluent hypergeometric function $\Psi_2[\cdot;\cdot,\cdot;\cdot,\cdot]$ for which we know of no open-source implementations. We therefore implemented the algorithm of Orsi [2017] in Cython. We have released our code for this, along with our implementations of DNCB-MF and BG-NMF and the real and synthetic datasets that we used for our experiments.[1]



(a) With $\epsilon_1, \epsilon_2 \geq 1$, the DNCB distribution takes uni-modal or bi-modal shapes over $(0, 1)$, similar to the beta distribution; with $\lambda_1 = \lambda_2 = 0$, the DNCB distribution coincides with the beta distribution.



(b) The DNCB distribution can additionally take tri-modal shapes when $\epsilon_1 < 1$ or $\epsilon_2 < 1$.

[1] https://github.com/aschein/dncb-mf

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

B POSTERIOR INFERENCE

Here, we provide a complete summary of our entire Gibbs sampler. As we described in section 4 of the main paper, the first step is to sample the gamma-distributed auxiliary variables:

$$\left(\gamma_{ij}^{(\bullet)} \mid -\right) \sim \text{Gam}\left(\epsilon_0^{(\bullet)} + y_{ij}^{(\bullet)}, 1\right), \tag{1}$$

$$\gamma_{ij}^{(1)} = \beta\gamma_{ij}^{(\bullet)} \quad \text{and} \quad \gamma_{ij}^{(2)} = (1-\beta)\gamma_{ij}^{(\bullet)}. \tag{2}$$

The Poisson-distributed auxiliary variables are then conditionally independent Bessel random variables—i.e.,

$$(y_{ij}^{(r)}|-) \sim \text{Bess}\left(\epsilon_0^{(r)} - 1,\, 2\sqrt{\gamma_{ij}^{(r)} \sum_{k=1}^{K} \theta_{ik}^{(r)} \phi_{kj}}\right) \tag{3}$$

for $r \in \{1, 2\}$. Conditioned on these auxiliary counts, the updates for the latent factors follow from gamma–Poisson matrix factorization. First, we represent each count as the sum of K subcounts—i.e., $y_{ij}^{(r)} = \sum_{k=1}^{K} y_{ijk}^{(r)}$. By Poisson additivity, each of these subcounts is Poisson distributed and their complete conditional is a multinomial distribution:

$$\left(\left(y_{ijk}^{(r)}\right)_{k=1}^{K} \mid -\right) \sim \text{Multi}\left(y_{ij}^{(r)},\, \left(\frac{\theta_{ik}^{(r)}\phi_{kj}}{\sum_{k'=1}^{K}\theta_{ik'}^{(r)}\phi_{kj}}\right)_{k=1}^{K}\right). \tag{4}$$

By Poisson–gamma conjugacy, the complete conditionals of the latent factors, conditioned on the subcounts, are

$$\left(\theta_{ik}^{(r)} \mid -\right) \sim \text{Gam}\left(a_0 + \sum_{j=1}^{M} y_{ijk}^{(r)},\, b_0 + \sum_{j=1}^{M} \phi_{kj}\right), \tag{5}$$

$$\left(\phi_{kj} \mid -\right) \sim \text{Gam}\left(e_0 + \sum_{i=1}^{N}\sum_{r=1}^{2} y_{ijk}^{(r)},\, f_0 + \sum_{i=1}^{N}\sum_{r=1}^{2} \theta_{ik}^{(r)}\right). \tag{6}$$

Equations (1) to (6) summarize the entire Gibbs sampler for DNCB-MF. Iteratively following these steps is asymptotically guaranteed to sample from the exact posterior.

References

Andrea Ongaro and Carlo Orsi. Some results on non-central beta distributions. *Statistica*, 75(1):85–100, 2015.

Carlo Orsi. New insights into non-central beta distributions. *arXiv preprint arXiv:1706.08557*, 2017.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272, 2020.