JMLR: Workshop and Conference Proceedings 1:1–12, 2018. Symposium on Advances in Approximate Bayesian Inference

A Variational Inference Approach for Locally Private Inference of Poisson Factorization Models

Alexandra Schofield xanda@cs.cornell.edu
Aaron Schein aschein@cs.umass.edu
Zhiwei Steven Wu zsw@umn.edu
Hanna Wallach hanna@dirichlet.net

1. Introduction

Poisson factorization models are a powerful tool in social science and machine learning for understanding data generated by people. For instance, a corporate email corpus can surface count data such as interactions between parties, counts of words in documents, and logs of types of events and their actors recorded in these messages. However, these data may also encode interactions and content that their human generators wish to keep private, which risks leaking information specific to a few interactions in the final low-rank model.

Most existing methods for private inference of exponential family models operate in the *central model* (Park et al., 2016; Bernstein and Sheldon, 2018), in which a trusted aggregator collects all the data and perturbs it to ensure privacy. Recent work by Schein et al. (2018) provides a privacy-preserving method for Poisson factorization models in the *local model*, in which individual users perturb their own data to ensure privacy without relying on any trusted party. The privacy guarantee uses a generalization of local differential privacy (Warner, 1965) called *limited-precision local privacy*, which allows privacy protection on observations of variable scales (e.g. ranging from an entire document to a single word token). A corresponding MCMC procedure for inference of Poisson factorization models relies on the recharacterization of the traditional data generative process to include the addition of perturbation through two-sided geometric noise (Ghosh et al., 2012), with the true data characterized a random variable generated using the Skellam and Bessel distributions.

The implementation of this method in prior work of (Schein et al., 2018) applies an MCMC approach to iteratively resample the latent variables, such that after an initial burn-in period of resampling, these samples may be aggregated to estimate the true model parameters. However, this approach to private Bayesian Poisson matrix factorization is prohibitively slow for two primary reasons. First, it requires sampling from the Bessel distribution, which is much slower than sampling for nonprivate Poisson factorization. Second, the addition of private noise produces a denser matrix of observations, in which no observed zero entry in the data is necessarily a true zero in the original data. This addition converts inference from a sparse problem, where zero counts could be ignored, to a dense one, where every entry in the observed data requires an expensive sampling procedure.

We propose a new mean-field variational inference (VI) procedure to infer these models. To do this, we first prove that the Bessel distribution is in the exponential family if its parameter ν is fixed, and that it is the optimal Q-distribution for the latent count variables of the true data. However, traditional mean-field inference fails, as the resulting expectation disrupts downstream inference due to the non-integer value of the expectation of the Bessel. We overcome this obstacle by defining the Q-distribution over the count to be a delta-spike at the *mode* of the optimal Bessel distribution (which is integral, by definition). To show that this mode approximation is close to the value obtained using the mean, we prove that the absolute difference between the mean and mode of a Bessel distribution is bounded by one. Finally, we show that in a synthetic experiment, this method converges to reasonable estimates of the true model parameters 20 times faster than the MCMC approach.

2. Background

Poisson factorization (Titsias, 2008; Cemgil, 2009; Zhou and Carin, 2012; Gopalan and Blei, 2013; Paisley et al., 2014) is a broad class of models for learning latent structure from discrete data, containing many of the most widely used models in the social sciences. Poisson factorization assumes that each observed count $y_{\boldsymbol{i}}$ is a Poisson random variable—$y_{\boldsymbol{i}} \sim \text{Poisson}\left(\mu_{\boldsymbol{i}}\right)$—with unknown rate parameter $\mu_{\boldsymbol{i}}$ that is a deterministic function of shared model parameters. We use the multi-index notation $\boldsymbol{i}$ for generality; in the special case of Poisson *matrix* factorization, each count is indexed by a row and a column—e.g., $\boldsymbol{i}=(d,v)$—and its latent rate is defined as the dot product of corresponding row and column parameters—i.e., $\mu_{\boldsymbol{i}} = \mu_{\boldsymbol{i}} = \boldsymbol{\theta}_d^\top \boldsymbol{\phi}_v$. In many cases, the latent rate is defined to be a *linear* function of shared model parameters—i.e., $\mu_{\boldsymbol{i}} = \sum_{k=1}^K \mu_{\boldsymbol{i}k}$, where $\mu_{\boldsymbol{i}k}$ is a function of latent parameters specific to latent component k. In Poisson matrix factorization, we define $\mu_{\boldsymbol{i}k}$ as the product of the latent parameters associated with $\boldsymbol{i}=(d,v)$, or $\mu_{\boldsymbol{i}k} = \theta_{dk}\phi_{kv}$.

MCMC for Poisson factorization. When the latent rate is a linear function, the observed count can be interpreted as the sum of latent sub-counts—i.e., $y_{\boldsymbol{i}} = \sum_{k=1}^K y_{\boldsymbol{i}k}$—where each sub-count is an independent Poisson random variable $y_{\boldsymbol{i}k} \sim \text{Poisson}\left(\mu_{\boldsymbol{i}k}\right)$. The first step in MCMC inference is to sample the vector of sub-counts $\boldsymbol{y_i} = \left(y_{\boldsymbol{i}k}\right)_{k=1}^K$ from its complete conditional, which is a multinomial whose priors are determined by $\mu_{\boldsymbol{i}}$:

$$\left(\boldsymbol{y_i} \mid -\right) \sim \text{Multinom}\left(y_{\boldsymbol{i}},\, \boldsymbol{\mu_i}\right) \tag{1}$$

where $\boldsymbol{\mu_i} = \left(\mu_{\boldsymbol{i}k}\right)_{k=1}^K$. The multinomial distribution is often parameterized with a *probability* vector $\boldsymbol{p_i}$. In this paper, we parameterize it using a *proportionality* vector $\boldsymbol{\mu_i}$ which leaves implicit the normalization step–i.e., $p_{\boldsymbol{i}k} = \frac{\mu_{\boldsymbol{i}k}}{\sum_{k'=1}^K \mu_{\boldsymbol{i}k'}}$. Conditioned on samples of the sub-counts, updates to the other latent variables are model-specific—i.e., depend on the particular structure of the rate function and the priors over the parameters. However, in what follows, we derive a general approach that applies to all models with a linear rate $\mu_{\boldsymbol{i}}$.

CAVI for Poisson factorization. The optimal variational family for $\boldsymbol{y_i}$ is proportional to the geometric expectation of its complete conditional (given in Equation 1) and equal to:

$$Q^*\left(\boldsymbol{y_i}\right) = \text{Multinom}\left(\boldsymbol{y_i}; y_{\boldsymbol{i}},\, \left(\mathbb{G}_Q\left[\mu_{\boldsymbol{i}k}\right]\right)_{k=1}^K\right) \tag{2}$$

The geometric expectations $\mathbb{G}_Q\left[\mu_{\boldsymbol{i}k}\right] = e^{\mathbb{E}_Q[\ln \mu_{\boldsymbol{i}k}]}$ can be understood as messages from the factors $Q(\mu_{\boldsymbol{i}k})$ in the context of a message-passing algorithm between variational parameters. In standard cases, these expectations are available in closed-form. The factors $Q(\mu_{\boldsymbol{i}k})$ depend in turn on messages from $Q(\boldsymbol{y_i})$ expressed via the expectation of $y_{\boldsymbol{i}}$:

$$\mathbb{E}_Q\left[y_{\boldsymbol{i}k}\right] = y_{\boldsymbol{i}} \frac{\mathbb{G}_Q\left[\mu_{\boldsymbol{i}k}\right]}{\sum_{k'=1}^{K} \mathbb{G}_Q\left[\mu_{\boldsymbol{i}k'}\right]}. \tag{3}$$

Locally private MCMC for Poisson factorization. In locally private Poisson factorization, we do not directly observe the count $y_{\boldsymbol{i}}$. Instead, we observe a noised version of it $\tilde{y}_{\boldsymbol{i}}^{(\pm)} = \tau_{\boldsymbol{i}} + y_{\boldsymbol{i}}$ where $\tau_{\boldsymbol{i}} \sim 2\text{Geo}\left(\alpha_{\boldsymbol{i}}\right)$ is a two-sided geometric random variable generated based on privacy parameter $\alpha_{\boldsymbol{i}} \in [0, 1]$. Schein et al. (2018) show that such perturbation mechanism satisfies $(N, N \ln(1/\alpha_{\boldsymbol{i}}))$-limited-precision local privacy.[1] MCMC in this case proceeds by treating the true data $y_{\boldsymbol{i}}$ itself as a latent variable and re-sampling it from its complete conditional. To do this, one can augment standard Poisson factorization with a set of auxiliary variables for each data point $\mathcal{A}_{\boldsymbol{i}} = \left\{\lambda_{\boldsymbol{i}}^{(+)}, \lambda_{\boldsymbol{i}}^{(-)}, g_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}, \tilde{y}_{\boldsymbol{i}}^{(+)}, m_{\boldsymbol{i}}\right\}$, related to a reinterpretation of the generative process for two-sided geometric noise as given in Appendix A. The key feature of this auxiliary variable scheme is the two-step sampling procedure of a value of the sensitive count $y_{\boldsymbol{i}}$ as a latent variable:

$$\left(m_{\boldsymbol{i}} \mid -\right) \sim \text{Bessel}\left(|\tilde{y}_{\boldsymbol{i}}^{(\pm)}|,\, 2\sqrt{\lambda_{\boldsymbol{i}}^{(-)}\left(\lambda_{\boldsymbol{i}}^{(+)} + \mu_{\boldsymbol{i}}\right)}\right) \tag{4}$$

$$\left(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{(+)} \mid -\right) \sim \text{Multinom}\left(\tilde{y}_{\boldsymbol{i}}^{(+)}, (\lambda_{\boldsymbol{i}}^{(+)}, \mu_{\boldsymbol{i}1}, \ldots, \mu_{\boldsymbol{i}K})\right) \tag{5}$$

Here, $\tilde{y}_{\boldsymbol{i}}^{(+)}$ is defined deterministically as the maximum of $m_{\boldsymbol{i}}$ and $m_{\boldsymbol{i}} + \tilde{y}_{\boldsymbol{i}}^{(\pm)}$. The noisy observation $\tilde{y}_{\boldsymbol{i}}^{(\pm)}$ and $\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{(+)} = (g_{\boldsymbol{i}}^{(+)}, y_{\boldsymbol{i}1}, \ldots, y_{\boldsymbol{i}K})$ is a vector of latent sub-counts that sum to $\tilde{y}_{\boldsymbol{i}}^{(+)}$: the first of these sub-counts $g_{\boldsymbol{i}}^{(+)}$ represents the positive noise added to the sensitive data $y_{\boldsymbol{i}} = \sum_{k=1}^{K} y_{\boldsymbol{i}k}$, while the latter sub-counts constitute the vector of sub-counts corresponding to each latent component in non-private Poisson factorization (see Equation 1).

3. Locally private variational inference for Poisson factorization

In this section we describe the challenges to deriving CAVI for locally private Poisson factorization and sketch our solutions. In CAVI, we look to impose a factorized variational distribution over all latent variables which, in this case, includes the set of auxiliary variables $\mathcal{A}_{\boldsymbol{i}}$ for each data point mentioned in the last section. The main technical challenge lies in finding a variational families for $m_{\boldsymbol{i}}$ and $\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{(+)}$ that yield a good approximation and closed-form messages to the other factors. Towards this goal, we prove the following two related propositions:

Proposition 1 *(Proven in Appendix B.) The Bessel distribution $m \sim$ Bessel(ν, a) for fixed ν is an exponential family with sufficient statistic $T_\nu(m) = 2m + \nu$, natural parameter $\eta_\nu(m) = \log(\frac{a}{2})$, and base measure $h_\nu(m) = \frac{1}{m!\Gamma(m+\nu+1)}$.*

[1] A local randomizer R satisfies (N, ϵ)-limited-precision local privacy if for any two observations $y, y' \in \mathbb{N}^d$ such that $\|y - y'\|_1 \leq N$, R satisfies $\Pr[R(y) = r] \leq \exp(\epsilon) \Pr[R(y') = r]$ for any r in the output range.

Proposition 2 *The optimal variational distribution for $m_{\boldsymbol{i}}$ is a Bessel distribution:*

$$Q^*(m_{\boldsymbol{i}}) = \textit{Bessel}\left(m_{\boldsymbol{i}};\, |\tilde{y}_{\boldsymbol{i}}^{(\pm)}|, 2\sqrt{\mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(-)}\right]\mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(+)}+\mu_{\boldsymbol{i}}\right]}\right). \tag{6}$$

Unfortunately, selecting the Bessel distribution as the family for $Q(m_{\boldsymbol{i}})$ prevents a closed-form solution to $Q^*(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{(+)})$: specifically, it results in the expression of the integer number of samples in a binomial distribution downstream in the generative process as a non-integer expectation of $m_{\boldsymbol{i}}$. To overcome this, we instead select the variational family for $Q(m_{\boldsymbol{i}})$ to be a Dirac delta function at the *mode* of the optimal family (in Equation 6):

$$Q(m_{\boldsymbol{i}}) = \delta\left[m_{\boldsymbol{i}} = \text{mode}(Q^*(m_{\boldsymbol{i}}))\right] \tag{7}$$

This choice allows a closed-form solution for the optimal variational distribution over $\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{(+)}$:

$$Q^*(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{(+)}) = \text{Multinom}\left(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{(+)};\, \mathbb{E}_Q\left[\tilde{y}_{\boldsymbol{i}}^{(+)}\right],\, \left(\mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(+)}\right], \mathbb{G}_Q\left[\mu_{\boldsymbol{i}1}\right], \ldots, \mathbb{G}_Q\left[\mu_{\boldsymbol{i}K}\right]\right)\right) \tag{8}$$

Furthermore, we can prove that this degenerate choice variational family is in fact nearly optimal in practice. Since the other factors depend on $Q(m_{\boldsymbol{i}})$ only through the message $\mathbb{E}_Q\left[m_{\boldsymbol{i}}\right]$, we need only show that the difference between the optimal message and ours is small. Under the optimal $Q^*(m_{\boldsymbol{i}})$, this message equals the *expected value* under the Bessel distribution in Equation 6 while in our solution, it equals the *mode* of that same distribution. To bound the error between these two, we prove the following proposition:

Proposition 3 *(Proven in Appendix C.) The absolute difference between the mean and mode of the Bessel distribution is bounded by 1:*

$$|\ \mathbb{E}_{\textit{Bessel}(m;\nu,a))}[m] - \textit{mode}(\textit{Bessel}(m;\nu,a))\ | \leq 1.$$

Equations 6 and 8 provide solutions to the main challenge of deriving a tractable and nearly optimal variational family. The closed-form solutions to the optimal variational families for all other auxiliary variables are more straightforward to derive and we give them in the supplementary material (see Appendix D). There is one final challenge: the log term inside the expectation $\mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(+)}+\mu_{\boldsymbol{i}}\right] = e^{\mathbb{E}_Q\left[\ln \lambda_{\boldsymbol{i}}^{(+)}+\mu_{\boldsymbol{i}}\right]}$ in Equation 6 prevents the derivation of a closed form. However, we can derive a first-order Taylor approximation via the delta method (Ver Hoef, 2012).

4. Discussion

We present a new CAVI algorithm for inference of Bayesian Poisson factorization models under local privacy. Our method relies on two key theoretical insights about the Bessel distribution: first, that with fixed ν, it belongs to the exponential family, and second, that its mode is an integer neighbor of its mean. Using these, we implement a tool with significant performance improvements over even an optimized version of the prior MCMC algorithm. On a synthetic test case of inferring a rank-50 factorization of an 1000 by 1000 data observation matrix, we obtain a 20x speedup in model inference. We detail the construction of this synthetic performance experiment in Appendix E.

Acknowledgments

We would like to thank David Mimno, and collaborators at Microsoft Research and in the Cornell Natural Language Processing group for their support and feedback. We would also like to thank the reviewers for their helpful comments.

References

Garrett Bernstein and Daniel Sheldon. Differentially private bayesian inference for exponential families. In *Advances in Neural Information Processing Systems 31*, 2018.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

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

Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

Luc Devroye. Simulating Bessel random variables. *Statistics & Probability Letters*, 57(3): 249–257, 2002.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

Prem K. Gopalan and David M. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

John Paisley, David M. Blei, and Michael I. Jordan. Bayesian nonnegative matrix factorization with stochastic variational inference. In Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*, pages 203–222. 2014.

Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. Private topic modeling. *arXiv:1609.04120*, 2016.

Aaron Schein, Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, and Hanna Wallach. Locally private bayesian inference for count models. *arXiv preprint arXiv:1803.08471*, 2018.

Michalis K. Titsias. The infinite gamma–Poisson feature model. In *Advances in Neural Information Processing Systems 21*, pages 1513–1520, 2008.

Jay M Ver Hoef. Who invented the delta method? *The American Statistician*, 66(2): 124–127, 2012.

Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Lin Yuan and John D. Kalbfleisch. On the Bessel distribution and related problems. *Annals of the Institute of Statistical Mathematics*, 52(3):438–447, 2000.

Mingyuan Zhou and Lawrence Carin. Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems 25*, pages 2546–2554, 2012.

Appendix A. Different factorizations of the joint posterior

To find a variational inference algorithm, we leverage existing equivalent characterizations of the generative process for Poisson factorization models under differential privacy.[2] We consider in this case the variable $y_{\boldsymbol{i}}$ to be a true observed data point, while $\tilde{y}_{\boldsymbol{i}}^{(\pm)}$ is the observed version with random noise. We treat two-sided geometric noise as being the difference of two Poisson variables $g_{\boldsymbol{i}}^{(+)}$ and $g_{\boldsymbol{i}}^{(-)}$, generated with Gamma-distributed priors $\lambda_{\boldsymbol{i}}^{(+)}$ and $\lambda_{\boldsymbol{i}}^{(-)}$, respectively/ We define $\tilde{y}_{\boldsymbol{i}}^{(+)} = y_{\boldsymbol{i}} + g_{\boldsymbol{i}}^{(+)}$ and $\tilde{y}_{\boldsymbol{i}}^{(\pm)} = y_{\boldsymbol{i}} + g_{\boldsymbol{i}}^{(+)} - g_{\boldsymbol{i}}^{(-)}$. We can determine whether $\tilde{y}_{\boldsymbol{i}}^{(\pm)}$ is positive or negative by comparing the values of $\tilde{y}_{\boldsymbol{i}}^{(+)}$ and $g_{\boldsymbol{i}}^{(-)}$; we refer to the minimum of these two as $m_{\boldsymbol{i}}$.

$$P\left(g_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}, \left(y_{\boldsymbol{i}k}\right)_{k=1}^{K}, y_{\boldsymbol{i}}, \tilde{y}_{\boldsymbol{i}}^{(+)}, \tilde{y}_{\boldsymbol{i}}^{(\pm)}, m_{\boldsymbol{i}}\right). \tag{9}$$

The most straightforward factorization of this joint first generates all the Poisson random variables, then computes the remaining variables given their deterministic relationships to the underlying Poissons:

$$\begin{aligned} &P\left(g_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}, \left(y_{\boldsymbol{i}k}\right)_{k=1}^{K}, y_{\boldsymbol{i}}, \tilde{y}_{\boldsymbol{i}}^{(+)}, \tilde{y}_{\boldsymbol{i}}^{(\pm)}, m_{\boldsymbol{i}}\right) \\ &\quad = \text{Poisson}\left(g_{\boldsymbol{i}}^{(+)}; \lambda_{\boldsymbol{i}}^{(+)}\right) \text{Poisson}\left(g_{\boldsymbol{i}}^{(-)}; \lambda_{\boldsymbol{i}}^{(-)}\right) \left(\prod_{k=1}^{K} \text{Poisson}\left(y_{\boldsymbol{i}k}; \theta_{dk}\phi_{kv}\right)\right) \mathbb{1}\left(y_{\boldsymbol{i}} = \sum_{k=1}^{K} y_{\boldsymbol{i}k}\right) \\ &\qquad \mathbb{1}\left(\tilde{y}_{\boldsymbol{i}}^{(+)} = y_{\boldsymbol{i}} + g_{\boldsymbol{i}}^{(+)}\right) \mathbb{1}\left(\tilde{y}_{\boldsymbol{i}}^{(\pm)} = \tilde{y}_{\boldsymbol{i}}^{(+)} - g_{\boldsymbol{i}}^{(-)}\right) \mathbb{1}\left(m_{\boldsymbol{i}} = \min\{\tilde{y}_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}\}\right). \end{aligned} \tag{10}$$

We can equivalently first generate the sums of Poissons and then thin them using multinomial and binomial draws. In the following equation, the delta functions are implicitly present in the multinomial and binomial PMFs. Note that we write the probability parameters in the multinomial and binomial PMFs as unnormalized vectors. Also note that

[2] Parts of this and Appendix D have been excerpted from supplementary material in a prior preprint version of Schein et al. (2018).

$\mu_{\boldsymbol{i}} = \sum_k \prod_d \theta^{(i)}_{i_d,k}$, where i_d is the index into the dth dimension of index vector $\boldsymbol{i}$.

$$\begin{aligned}
&P\left(g^{(+)}_{\boldsymbol{i}}, g^{(-)}_{\boldsymbol{i}}, \left(y_{\boldsymbol{i}k}\right)_{k=1}^{K}, y_{\boldsymbol{i}}, \tilde{y}^{(+)}_{\boldsymbol{i}}, \tilde{y}^{(\pm)}_{\boldsymbol{i}}, m_{\boldsymbol{i}}\right) \\
&\quad= \text{Poisson}\left(g^{(-)}_{\boldsymbol{i}}; \lambda^{(-)}_{\boldsymbol{i}}\right)\ \text{Poisson}\left(\tilde{y}^{(+)}_{\boldsymbol{i}}; \lambda^{(+)}_{\boldsymbol{i}} + \mu_{\boldsymbol{i}}\right) \\
&\qquad \text{Binom}\left((y_{\boldsymbol{i}}, g^{(+)}_{\boldsymbol{i}});\ \tilde{y}^{(+)}_{\boldsymbol{i}}, (\mu_{\boldsymbol{i}}, \lambda^{(+)}_{\boldsymbol{i}})\right)\ \text{Mult}\left(\left(y_{\boldsymbol{i}k}\right)_{k=1}^{K};\ y_{\boldsymbol{i}}, \left(\prod_d \theta^{(i)}_{i_d,k}\right)_{k=1}^{K}\right) \\
&\qquad \mathbb{1}\left(\tilde{y}^{(\pm)}_{\boldsymbol{i}} = \tilde{y}^{(+)}_{\boldsymbol{i}} - g^{(-)}_{\boldsymbol{i}}\right)\ \mathbb{1}\left(m_{\boldsymbol{i}} = \min\{\tilde{y}^{(+)}_{\boldsymbol{i}}, g^{(-)}_{\boldsymbol{i}}\}\right).
\end{aligned} \tag{11}$$

We can equivalently first generate the difference $\tilde{y}^{(\pm)}_{\boldsymbol{i}}$ and minimum $m_{\boldsymbol{i}}$ as Skellam and Bessel random variables. Conditioned on these variables, we can then compute $\tilde{y}^{(+)}_{\boldsymbol{i}}$ and $g^{(-)}_{\boldsymbol{i}}$ via their deterministic relationship and, finally, thin $\tilde{y}^{(+)}_{\boldsymbol{i}}$ using binomial and multinomial draws:

$$\begin{aligned}
&P\left(g^{(+)}_{\boldsymbol{i}}, g^{(-)}_{\boldsymbol{i}}, \left(y_{\boldsymbol{i}k}\right)_{k=1}^{K}, y_{\boldsymbol{i}}, \tilde{y}^{(+)}_{\boldsymbol{i}}, \tilde{y}^{(\pm)}_{\boldsymbol{i}}, m_{\boldsymbol{i}}\right) \\
&\quad= \text{Skel}\left(\tilde{y}^{(\pm)}_{\boldsymbol{i}};\ \lambda^{(+)}_{\boldsymbol{i}} + \mu_{\boldsymbol{i}},\ \lambda^{(-)}_{\boldsymbol{i}}\right)\ \text{Bes}\left(m_{\boldsymbol{i}};\ |\tilde{y}^{(\pm)}_{\boldsymbol{i}}|,\ 2\sqrt{\lambda^{(-)}_{\boldsymbol{i}}(\lambda^{(+)}_{\boldsymbol{i}} + \mu_{\boldsymbol{i}})}\right) \\
&\quad \mathbb{1}\left(\tilde{y}^{(+)}_{\boldsymbol{i}} = m_{\boldsymbol{i}}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{\boldsymbol{i}} \leq 0)}\ \mathbb{1}\left(g^{(-)}_{\boldsymbol{i}} = m_{\boldsymbol{i}}\right)^{\mathbb{1}(\tilde{y}^{(\pm)}_{\boldsymbol{i}} > 0)}\ \mathbb{1}\left(\tilde{y}^{(\pm)}_{\boldsymbol{i}} = \tilde{y}^{(+)}_{\boldsymbol{i}} - g^{(-)}_{\boldsymbol{i}}\right) \\
&\quad \text{Binom}\left((y_{\boldsymbol{i}}, g^{(+)}_{\boldsymbol{i}});\ \tilde{y}^{(+)}_{\boldsymbol{i}}, (\mu_{\boldsymbol{i}}, \lambda^{(+)}_{\boldsymbol{i}})\right)\ \text{Mult}\left(\left(y_{\boldsymbol{i}k}\right)_{k=1}^{K};\ y_{\boldsymbol{i}}, \left(\prod_d \theta^{(i)}_{i_d,k}\right)_{k=1}^{K}\right).
\end{aligned} \tag{12}$$

It is this last factorization that enables us to derive the variational distribution.

Appendix B. The Bessel distribution as exponential family

Theorem 1 *The Bessel distribution $m \sim Bes(\nu, a)$ for fixed ν is an exponential family with sufficient statistic $T_\nu(m) = 2m + \nu$, natural parameter $\eta_\nu(m) = \log\left(\frac{a}{2}\right)$, and base measure $h_\nu(m) = \frac{1}{m!\,\Gamma(m+\nu+1)}$.*

Proof The Bessel distribution (Yuan and Kalbfleisch, 2000) is a two-parameter distribution over the non-negative integers:

$$f(n; a, \nu) = \frac{\left(\frac{a}{2}\right)^{2n+\nu}}{n!\,\Gamma(n+\nu+1) I_\nu(a)}, \tag{13}$$

where the normalizing constant $I_\nu(a)$ is a modified Bessel function of the first kind—i.e.,

$$I_\nu(a) = \sum_{n=0}^{\infty} \frac{\left(\frac{a}{2}\right)^{2n+\nu}}{n!\,\Gamma(n+\nu+1)}. \tag{14}$$

For fixed and known ν, we can rewrite the Bessel PMF as

$$f(n; a, \nu) = \frac{1}{n!\,\Gamma(n+\nu+1)} \exp\left((2n+\nu)\ \log\left(\frac{a}{2}\right) - \log I_\nu(a)\right). \tag{15}$$

We can then define the following functions:

$$h_\nu(n) = \frac{1}{n!\,\Gamma(n{+}\nu{+}1)} \tag{16}$$

$$T_\nu(n) = 2n + \nu \tag{17}$$

$$\eta_\nu(a) = \log\left(\frac{a}{2}\right) \tag{18}$$

$$A_\nu(a) = \log I_\nu(a). \tag{19}$$

Finally, we can rewrite the Bessel PMF in the exponential-family form:

$$f(n; a, \nu) = h_\nu(n) \exp\left(\eta_\nu(a) \cdot T_\nu(n) - A_\nu(a)\right). \tag{20}$$

■

Appendix C. The mode and mean of the Bessel distribution

Theorem 2 *The mode of a Bessel distribution for parameters a, ν can be a constant-bounded approximation of the mean:*

$$\left|\mathbb{E}_{\mathrm{Bessel}(m;\nu,a))}[m] - \mathrm{mode}(\mathrm{Bessel}\,(m; a, \nu))\right| \leq 1.$$

The intuitive meaning of this is that the mode of the Bessel distribution is guaranteed to be one of the two integers closest to the mean of the distribution. Given one of these two integers will always be the best integer approximation of this number, we can also say that there cannot exist an integer approximation of the mean of the Bessel that is strictly between the mode and the mean of the Bessel distribution.

Proof

A Bessel distribution takes two arguments, which we refer to as its order, ν and coordinate, a. The distribution is defined as:

$$p(x = n | x \sim \mathrm{Bessel}\,(\nu, a)) = \frac{1}{I_\nu(a) n! \Gamma(n + \nu + 1)} \left(\frac{a}{2}\right)^{2n+\nu},$$

where $I_\nu(a)$ is a modified Bessel function of the first kind. The arithmetic mean of the distribution is

$$\mathbb{E}_{\mathrm{Bessel}(m;\nu,a))}[m] = \frac{a}{2} R_\nu(a),$$

with $R_\nu(a)$ referring to the ratio of two Bessel functions:

$$\frac{I_{\nu+1}(a)}{I_\nu(a)}.$$

The Bessel distribution has one or two neighboring integer modes. The mode can be computed directly from the parameters of the distribution without Bessel functions:

$$\mathrm{mode}(\mathrm{Bessel}\,(\nu, a)) = \left\lfloor \frac{\sqrt{a^2 + \nu^2} - \nu}{2} \right\rfloor.$$

Unlike the mean, which can take arbitrary non-negative real values, the mode is guaranteed by the floor function to be a non-negative integer.

We use the following bound on the mean of a Bessel ratio from Devroye (2002):

$$\frac{a}{\nu + 1 + \sqrt{a^2 + (\nu+1)^2}} \leq R_\nu(a) \leq \frac{a}{\nu + \sqrt{a^2 + \nu^2}}. \tag{21}$$

We multiply through by $\frac{a}{2}$ to bound the mean of the Bessel distribution:

$$\frac{a^2}{2(\nu + 1 + \sqrt{a^2 + (\nu+1)^2})} \leq \mathbb{E}_{\text{Bessel}(m;\nu,a))}[m] \leq \frac{a^2}{2(\nu + \sqrt{a^2 + \nu^2})}. \tag{22}$$

We can rewrite these bounds using a difference-of-squares:

$$\begin{aligned}
\frac{a^2}{2(\nu + \sqrt{a^2 + \nu^2})} &= \frac{a^2(\sqrt{a^2 + \nu^2} - \nu)}{2(\nu + \sqrt{a^2 + \nu^2})(\sqrt{a^2 + \nu^2} - \nu)} \\
&= \frac{a^2(\sqrt{a^2 + \nu^2} - \nu)}{2a^2 + \nu^2 - \nu^2} \\
&= \frac{\sqrt{a^2 + \nu^2} - \nu}{2}.
\end{aligned}$$

This upper bound coincides with the unrounded formulation of the mode. Because the mode is the floor of this quantity, we know it is less than or equal to this upper bound with a difference of less than 1.

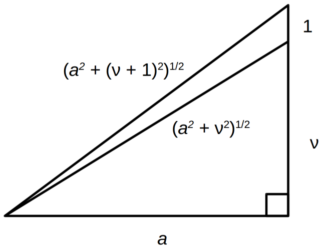
We can convert the lower bound in the same way:

$$\frac{a^2}{2((\nu+1) + \sqrt{a^2 + (\nu^2 + 1)})} = \frac{\sqrt{a^2 + (\nu+1)^2} - (\nu+1)}{2}.$$

We are interested in bounding the difference between the upper and lower bounds:

$$\frac{\sqrt{a^2 + \nu^2} - \nu}{2} - \frac{\sqrt{a^2 + (\nu+1)^2} - (\nu+1)}{2} = \frac{\sqrt{a^2 + \nu^2} + 1 - \sqrt{a^2 + (\nu+1)^2}}{2}. \tag{23}$$

Knowing that ν is positive, we can state that $\sqrt{a^2 + \nu^2} < \sqrt{a^2 + (\nu+1)^2}$, or $\sqrt{a^2 + (\nu+1)^2} - \sqrt{a^2 + \nu^2} > 0$. Based on the upper slice of the triangle in the figure below, the triangle inequality also gives us another bound, that $\sqrt{a^2 + \nu^2} + 1 > \sqrt{a^2 + (\nu+1)^2}$, or $1 > \sqrt{a^2 + (\nu+1)^2} - \sqrt{a^2 + \nu^2}$.



Together, these imply that

$$0 \leq \sqrt{a^2+\nu^2} + 1 - \sqrt{a^2+(\nu+1)^2} \leq 1.$$

Substituting this in to our computation of the distance between the upper and lower bounds of the mean, we find that

$$\frac{\sqrt{a^2+\nu^2} + 1 - \sqrt{a^2+(\nu+1)^2}}{2} < \frac{1}{2}, \tag{24}$$

or that the inferred upper and lower bounds for the mean of the Bessel distribution produce an interval no larger than $\frac{1}{2}$. The upper bound of this $\frac{1}{2}$ interval is the same as the upper bound of the length-1 open interval of the mode of the Bessel. In the most extreme case when the mode is at the bottom end of its range and the mean is at the top, we have

$$\mathbb{E}_{\text{Bessel}(m;\nu,a))}[m] - \text{mode}(\text{Bessel}\,(\nu,a)) < 1.$$

In the opposite case, we have

$$\text{mode}(\text{Bessel}\,(\nu,a)) - \mathbb{E}_{\text{Bessel}(m;\nu,a))}[m] \leq \frac{1}{2} < 1.$$

■

Appendix D. Deriving the CAVI updates for the variational distribution

Recall that the observed data consists of the difference variables $\tilde{y}_{\boldsymbol{i}}^{(\pm)}$.

$$Q\left(m_{\boldsymbol{i}}, \tilde{y}_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}, y_{\boldsymbol{i}}, (y_{\boldsymbol{i}k})_{k=1}^{K}\right) \propto \mathbb{G}_Q\left[P(\tilde{y}_{\boldsymbol{i}}^{(\pm)}, m_{\boldsymbol{i}}, \tilde{y}_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}, y_{\boldsymbol{i}}, (y_{\boldsymbol{i}k})_{k=1}^{K} \mid -)\right]. \tag{25}$$

Because the likelihood (i.e., the Skellam term) in equation 12 does not depend on any of these latent variables, it disappears entirely. We can then rewrite the right-hand side of equation 25 as:

$$\begin{aligned}
&\mathbb{G}_Q\left[P(m_{\boldsymbol{i}}, \tilde{y}_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}, y_{\boldsymbol{i}}, (y_{\boldsymbol{i}k})_{k=1}^{K} \mid \tilde{y}_{\boldsymbol{i}}^{(\pm)} -)\right] \\
&\quad = \mathbb{G}_Q\left[\text{Bes}\left(m_{\boldsymbol{i}};\, |\tilde{y}_{\boldsymbol{i}}^{(\pm)}|,\, 2\sqrt{\lambda_{\boldsymbol{i}}^{(-)}(\lambda_{\boldsymbol{i}}^{(+)} + \mu_{\boldsymbol{i}})}\right)\right] \\
&\qquad \mathbb{1}\left(\tilde{y}_{\boldsymbol{i}}^{(+)} = m_{\boldsymbol{i}}\right)^{\mathbb{1}(\tilde{y}_{\boldsymbol{i}}^{(\pm)} \leq 0)} \mathbb{1}\left(g_{\boldsymbol{i}}^{(-)} = m_{\boldsymbol{i}}\right)^{\mathbb{1}(\tilde{y}_{\boldsymbol{i}}^{(\pm)} > 0)} \mathbb{1}\left(\tilde{y}_{\boldsymbol{i}}^{(\pm)} = \tilde{y}_{\boldsymbol{i}}^{(+)} - g_{\boldsymbol{i}}^{(-)}\right) \\
&\qquad \mathbb{G}_Q\left[\text{Binom}\left((y_{\boldsymbol{i}}, g_{\boldsymbol{i}}^{(+)});\, \tilde{y}_{\boldsymbol{i}}^{(+)}, (\mu_{\boldsymbol{i}}, \lambda_{\boldsymbol{i}}^{(+)})\right)\, \text{Mult}\left((y_{\boldsymbol{i}k})_{k=1}^{K}\,;\, y_{\boldsymbol{i}},\, \left(\prod_d \theta_{i_d,k}^{(i)}\right)_{k=1}^{K}\right)\right].
\end{aligned} \tag{26}$$

Theorem 6.1 states that the Bessel distribution for fixed first parameter is an exponential family. We can therefore use standard results to push in the geometric expectations:

$$\begin{aligned}
&\mathbb{G}_Q\left[P(m_{\boldsymbol{i}}, \tilde{y}_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(+)}, g_{\boldsymbol{i}}^{(-)}, y_{\boldsymbol{i}}, (y_{\boldsymbol{i}k})_{k=1}^{K} \mid \tilde{y}_{\boldsymbol{i}}^{(\pm)} -)\right] \\
&= \text{Bes}\left(m_{\boldsymbol{i}};\, |\tilde{y}_{\boldsymbol{i}}^{(\pm)}|,\, 2\sqrt{\mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(-)}\right](\mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(+)} + \mu_{\boldsymbol{i}}\right])}\right) \\
&\qquad \mathbb{1}\left(\tilde{y}_{\boldsymbol{i}}^{(+)} = m_{\boldsymbol{i}}\right)^{\mathbb{1}(\tilde{y}_{\boldsymbol{i}}^{(\pm)} \leq 0)} \mathbb{1}\left(g_{\boldsymbol{i}}^{(-)} = m_{\boldsymbol{i}}\right)^{\mathbb{1}(\tilde{y}_{\boldsymbol{i}}^{(\pm)} > 0)} \mathbb{1}\left(\tilde{y}_{\boldsymbol{i}}^{(\pm)} = \tilde{y}_{\boldsymbol{i}}^{(+)} - g_{\boldsymbol{i}}^{(-)}\right) \\
&\qquad \text{Binom}\left((y_{\boldsymbol{i}}, g_{\boldsymbol{i}}^{(+)});\, \mathbb{E}_Q\left[\tilde{y}_{\boldsymbol{i}}^{(+)}\right], (\mathbb{G}_Q\left[\mu_{\boldsymbol{i}}\right], \mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(+)}\right])\right) \\
&\qquad \text{Mult}\left((y_{\boldsymbol{i}k})_{k=1}^{K};\, \mathbb{E}_Q\left[y_{\boldsymbol{i}}\right], \left(\mathbb{G}_Q\left[\prod_d \theta_{i_d,k}^{(i)}\right]\right)_{k=1}^{K}\right). \qquad (27)
\end{aligned}$$

There are two expectations that do not have an analytic form:

$$\mathbb{G}_Q\left[\lambda_{\boldsymbol{i}}^{(+)} + \mu_{\boldsymbol{i}}\right] = \exp\left(\mathbb{E}_Q\left[\ln\left(\lambda_{\boldsymbol{i}}^{(+)} + \sum_{k=1}^{K}\prod_d \theta_{i_d,k}^{(i)}\right)\right]\right) \qquad (28)$$

and

$$\mathbb{G}_Q\left[\mu_{\boldsymbol{i}}\right] = \exp\left(\mathbb{E}_Q\left[\ln\left(\sum_{k=1}^{K}\prod_d \theta_{i_d,k}^{(i)}\right)\right]\right); \qquad (29)$$

however, both can be very closely approximated using the delta method (Ver Hoef, 2012), which has been previously used in variational inference schemes to approximate intractable expectations (Braun and McAuliffe, 2010; Wang and Blei, 2013). In particular, for some variable $Y = f(X)$, expectation $\mathbb{E}[Y]$ is approximately:

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] \approx f\left(\mathbb{E}[X]\right) + \frac{1}{2} f''\left(\mathbb{E}[X]\right)\mathbb{V}[X]. \qquad (30)$$

In our case, we can consider the two-dimensional Poisson matrix factorization case where $\boldsymbol{i}$ are represented by row d and column v. We therefore have

$$\begin{aligned}
\mathbb{E}_Q\left[\ln \mu_{dv}\right] = \mathbb{E}_Q\left[\ln\left(\sum_{k=1}^{K}\theta_{dk}\phi_{kv}\right)\right] &\approx \ln\left(\mathbb{E}_Q\left[\sum_{k=1}^{K}\theta_{dk}\phi_{kv}\right]\right) - \frac{\mathbb{V}_Q\left[\sum_{k=1}^{K}\theta_{dk}\phi_{kv}\right]}{2\left(\mathbb{E}_Q\left[\sum_{k=1}^{K}\theta_{dk}\phi_{kv}\right]\right)^2} \qquad (31) \\
&= \ln\left(\sum_{k=1}^{K}\mathbb{E}_Q\left[\theta_{dk}\right]\mathbb{E}_Q\left[\phi_{kv}\right]\right) - \frac{\sum_{k=1}^{K}\mathbb{V}_Q\left[\theta_{dk}\phi_{kv}\right]}{2\left(\sum_{k=1}^{K}\mathbb{E}_Q\left[\theta_{dk}\right]\mathbb{E}_Q\left[\phi_{kv}\right]\right)^2}. \qquad (32)
\end{aligned}$$

Finally, because θ_{dk} and ϕ_{kv} are independent, we have

$$\mathbb{V}_Q\left[\theta_{dk}\phi_{kv}\right] = \mathbb{V}_Q\left[\theta_{dk}\right]\mathbb{V}_Q\left[\phi_{kv}\right] + \mathbb{V}_Q\left[\theta_{dk}\right]\left(\mathbb{E}_Q\left[\phi_{kv}\right]\right)^2 + \mathbb{V}_Q\left[\phi_{kv}\right]\left(\mathbb{E}_Q\left[\theta_{dk}\right]\right)^2. \qquad (33)$$

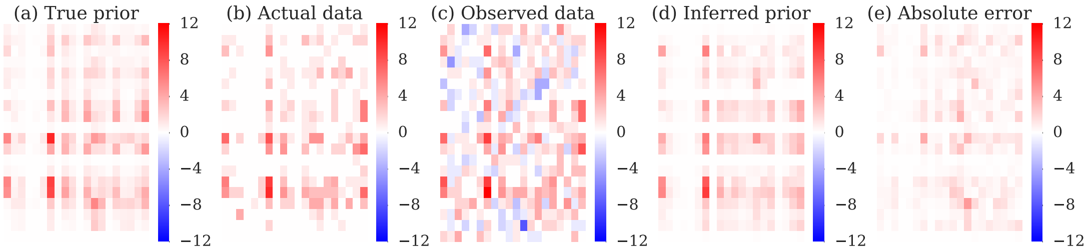


Figure 1: Demonstration of the results of the VI inference process for a 20-word, 20-document synthetic dataset with 3 latent topics. The true data parameters (a) are recovered well by our inference procedure (d), even though the noisy data (d) is much denser than the true data (c). The mean average error (MAE) between the true model parameters and the inferred parameters under privacy is 0.507, with the individual errors pictured in (e).

Appendix E. Validation

In order to test this result, we generated synthetic count data using a Poisson matrix factorization formulation analogous to the LDA topic model (Blei et al., 2003) with D documents, V unique terms in a document vocabulary, and K latent topics. We first used a Gamma prior to generate two matrices of latent parameters, θ of dimension $D \times K$ and ϕ of dimension $K \times V$. We then compute the product of these, $\theta\phi = \boldsymbol{Y}$, as the Poisson prior of our data generation process. Finally, we add two-sided geometric noise scaled to $\epsilon/N = 1$, a ratio that applies when the privacy budget ϵ and the maximum allowed difference between documents N are equal (e.g., a privacy budget of 2 to privatize 2-word spans). We then test our inference procedure to see how closely it estimates the true parameters of the original model. We find that our model successfully converges to within a reasonable estimate of the true model parameters given the data, as demonstrated in a small example in Figure 1.

Using a larger example of a synthetic 1000-by-1000 matrix of count observations, we test the performance of this algorithm as compared to the performance using inference with MCMC (Schein et al., 2018). We observe that a single-threaded version of the code from the MCMC implementation, even with optimizations such as reducing the frequency of parameter resampling for the noise distributions, each iteration of inference takes approximately 0.8 seconds. Using 5000 iterations of burnin and 2500 to collect samples, this takes a combined 1.67 hours to infer, with a final mean average error (MAE) of 0.36. In contrast, our variational inference model takes approximately 2.8 seconds per iteration, and typically requires 100 iterations to converge, resulting in a combined 5 minutes of inference to reach error of 0.52.