Weighted Tensor Decomposition for Learning Latent Variables with Partial Data

Supplementary Material

Appendices

A Generative models and empirical moments

Spherical Gaussian Mixtures The spherical Gaussian mixture model posits that the data matrix, $\mathbf{X} \in \mathbb{R}^{D\times N}$, consists of N data points represented as D dimensional vectors. The generative process for the n^{th} data point, $\mathbf{x}_n$, is

$$h_n \sim \text{Multinomial}(1, \boldsymbol{\pi}),$$
$$x_n|h_n, \mathbf{A} \sim \mathcal{N}(\mathbf{a}_{h_n}, \sigma^2).$$

where $\mathbf{a}_{h_n}$ is the $(h_n)^{th}$ column (topic) in the topics matrix $\mathbf{A} \in \mathbb{R}^{D\times K}$, and $\boldsymbol{\pi} \in \mathbb{R}^K$ represents the probability of data points to be drawn from each topic ($\sum_{k=1}^K \pi_k = 1$). A schematic illustration is presented in Figure 1.

Hsu and Kakade [2013] showed that if we estimate the variance of the Gaussians, σ^2, as the smallest eigenvalue of the covariance matrix, $\mathbb{E}[\mathbf{x}\otimes\mathbf{x}] - \mathbb{E}[\mathbf{x}]\otimes\mathbb{E}[\mathbf{x}]$, the empirical estimates

$$\hat{\mathbf{S}} = \mathbb{E}[\mathbf{x}\otimes\mathbf{x}] - \sigma^2\mathbf{I} \tag{1}$$

$$\hat{\mathbf{T}} = \mathbb{E}[\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}] - \sigma^2\sum_{i=1}^{D}(\mathbb{E}[\mathbf{x}]\otimes\mathbf{e}_i\otimes\mathbf{e}_i + \mathbf{e}_i\otimes\mathbb{E}[\mathbf{x}]\otimes\mathbf{e}_i + \mathbf{e}_i\otimes\mathbf{e}_i\otimes\mathbb{E}[\mathbf{x}]), \tag{2}$$

converge to the theoretical moments of the model -

$$\mathbf{S} = \sum_k \pi_k \mathbf{a}_k\mathbf{a}_k^T, \tag{3}$$

$$\mathbf{T} = \sum_k \pi_k \mathbf{a}_k\otimes\mathbf{a}_k\otimes\mathbf{a}_k. \tag{4}$$

Gamma-Poisson Model The second model we will focus on is the the gamma-Poisson (GP) generative model described in Podosinnikova et al. [2015]. The GP model is closely related to latent Dirichlet allocation (LDA) [Podosinnikova et al., 2015]. In addition to its popular use for modeling text corpora [Blei et al., 2003], the GP model is also relevant for capturing structure in applications where not all counts may be recorded (such as what parts of the genome are sequenced in genomics). A schematic illustration is presented in Figure 1.

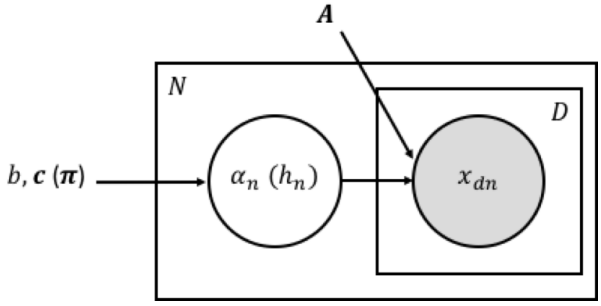

Figure 1: Schematic illustration of the gamma-Poisson and mixture of Gaussians generative models.

Formally, we represent the data as a matrix $\mathbf{X} \in \mathbb{N}_0^{D\times N}$ (that is, the data $\mathbf{X}$ is a matrix of non-negative integers), with every column $\mathbf{x}_n$ sampled according to

$$\alpha_{nk} \sim \text{Gamma}(c_k, b),$$
$$x_d|\alpha_n, \mathbf{A} \sim \text{Poisson}([\mathbf{A}\boldsymbol{\alpha}_n]_d).$$

Here, the global topics matrix $\mathbf{A} \in \mathbb{R}_+^{D\times K}$ can be interpreted as the collection of rates of word d in topic k; following Podosinnikova et al. [2015] and without loss of generality we constrain the columns in $\mathbf{A}$ to sum to 1. The observation-specific vector $\alpha_n \in \mathbb{R}^K$ determines the relative contribution of each of the K topics for a particular observation n. The parameters b and $\mathbf{c} \in \mathbb{R}^K$ are constants that encode our prior about both the length and relative popularities of the topics. In the context of text modeling, every element x_{dn} can be thought of as representing the number of times the d^{th} word in the vocabulary appears in the n^{th} document, with the mean document length being $L = \sum_k c_k/b$.

In this work, we will use the following second and third or-

der tensors, first introduced by Podosinnikova et al. [2015]:

$$\hat{\mathbf{S}} = \text{cov}(\mathbf{x}, \mathbf{x}) - \text{diag}(\mathbb{E}(\mathbf{x})) \tag{5}$$

$$\begin{aligned}\hat{T}_{d_1,d_2,d_3} &= \text{cum}(x_{d_1}, x_{d_2}, x_{d_3}) + 2\delta_{d_1 d_2 d_3}\mathbb{E}(x_{d_1}) \\ &- \delta_{d_2 d_3}\text{cov}(x_{d_1}, x_{d_2}) - \delta_{d_1 d_3}\text{cov}(x_{d_1}, x_{d_2}) \\ &- \delta_{d_1 d_2}\text{cov}(x_{d_1}, x_{d_3})\end{aligned} \tag{6}$$

where δ is the Kronecker delta and the second and third cumulants are defined as

$$\begin{aligned}\text{cov}(x_{d_1}, x_{d_2}) &= \mathbb{E}[(x_{d_1} - \mathbb{E}[x_{d_1}])(x_{d_2} - \mathbb{E}[x_{d_2}])], \\ \text{cum}(x_{d_1}, x_{d_2}, x_{d_3}) &= \mathbb{E}[(x_{d_1} - \mathbb{E}[x_{d_1}]) \\ &\quad (x_{d_2} - \mathbb{E}[x_{d_2}])(x_{d_3} - \mathbb{E}[x_{d_3}])].\end{aligned}$$

These empirical tensors $\hat{\mathbf{S}}$ and $\hat{\mathbf{T}}$ will converge to

$$\mathbf{S} = \sum_k s_k \mathbf{a}_k \mathbf{a}_k^T, \tag{7}$$

$$\mathbf{T} = \sum_k t_k \mathbf{a}_k \otimes \mathbf{a}_k \otimes \mathbf{a}_k. \tag{8}$$

where $\mathbf{a}_k$ is the k^{th} column of $\mathbf{A}$, $s_k = \text{var}(\alpha_k)$ and $t_k = \text{cum}(\alpha_k, \alpha_k, \alpha_k)$.

Note that we use the *moments* for the Gaussian mixtures, while for the GP model we calculate the *central moments*.

B Estimation of variance for Gaussian mixtures

When applying the WTPM to Gaussian mixtures, some attention must be paid to correctly estimating the variance of the mixtures. If the estimate for $cov(\mathbf{x}, \mathbf{x})$ is exact, its $D - K$ smallest eigenvalues are equal to σ^2, while all other eigenvalues are strictly larger than σ^2. Therefore Hsu and Kakade [2013] suggest using the smallest eigenvalue of $cov(\mathbf{x}, \mathbf{x})$ as an estimate for the variance of the Gaussians. In practice, we find that the mean of the $D - K$ smallest eigenvalues yields better estimate, and that the quality of inference is very sensitive to this estimate.

Furthermore, poorly estimated moments lead to a large estimation error for σ^2, and therefore to perform the WTPM on Gaussian mixtures, we first calculate the variance of the $D_c - K$ smallest eigenvalues of $cov(\mathbf{x}_c, \mathbf{x}_c)$, where $\mathbf{x}_c$ is the data vector containing only the D_c complete dimensions.

Adopting the interpretation of weighting the moments estimates as a rescaling of the dimensions, such rescaling would deform the spherical Gaussian mixtures into elliptical Gaussians. In appendix C we show that a weighting of the form given in 6 naturally leads to the correct form of the moments, and therefore once σ^2 is calculated, no further modification is needed to apply the WTPM to Gaussian mixtures.

C Moments for elliptical Gaussian mixtures

In this appendix we discuss the modifications to the moments estimates for elliptical Gaussian mixtures. We adopt the interpretation introduced earlier of viewing the weighting of the moments according to 6 as a rescaling of the data, $x^*_{dn} = x_{dn} w_d$. Following a similar derivation to that presented in Hsu and Kakade [2013] for non-spherical Gaussian mixtures, the estimates for the moments of the rescaled data are

$$\begin{aligned}\hat{S}^*_{d_1 d_2} &= \mathbb{E}[x^*_{d_1} x^*_{d_2}] - \sigma^{*2}_{d_1} I \\ \hat{T}^*_{d_1 d_2 d_3} &= \mathbb{E}[x^*_{d_1} x^*_{d_2} x^*_{d_3}] - \sigma^{*2}_{d_1}\mathbb{E}[x^*_{d_1}]\delta_{d_2}\delta_{d_3} \\ &\quad - \sigma^{*2}_{d_2}\delta_{d_1}\mathbb{E}[x^*_{d_2}]\delta_{d_3} - \sigma^{*2}_{d_3}\delta_{d_1}\delta_{d_2}\mathbb{E}[x^*_{d_3}],\end{aligned}$$

where σ^{*2}_d is the standard deviation of the rescaled dimension d. The difficulty in applying the tensor decomposition method to non-spherical Gaussian mixtures lies in the fact that we don't know of a straight forward way compute σ^{*2}_d, whereas in the case of spherical mixtures ($\sigma^2_d = \sigma^2$ for all d), the variance is simply the smallest eigenvalue of the matrix $\mathbb{E}[\mathbf{x} \otimes \mathbf{x}] - \mathbb{E}[\mathbf{x}] \otimes \mathbb{E}[\mathbf{x}]$. However, if we know that the original data is generated by a spherical Gaussian mixture with standard deviation σ^2, the standard deviation of every dimension after rescaling by w_d is $\sigma^{*2}_d = w^2_d \sigma^2$.

This implies that the rescaled moments can be written as

$$\begin{aligned}\hat{S}^*_{d_1 d_2} &= w_{d_1} w_{d_2} \hat{S}_{d_1 d_2} \\ \hat{T}^*_{d_1 d_2 d_3} &= w_{d_1} w_{d_2} w_{d_3} \hat{T}_{d_1 d_2 d_3}.\end{aligned}$$

This result, similar to equations 4 and 5 implies that the same weighting scheme used in our algorithm can be applied to Gaussian mixtures as well.

D Insensitivity to structure of topics

In this appendix we demonstrate that our results are insensitive to the exact structure of the topics. In Figure 2 we show results similar to the results shown in Figure 4 for experiments performed with randomly generated topics. Each data point is an average over 25 experiments, where for each experiment $\mathbf{A} \in \mathbb{R}_+^{D \times K}$ was generated by sampling each of the $K = 4$ columns in $\mathbf{A}$ from Dir($\mathbf{1}$). The qualitative effect of a transition between the full dimensionality method to the partial dimensionality method being optimal is still observed, as well as the ability of the WTPM to perform at least as well as the better of the two methods for the entire range of parameter space studied.

E Optimal weights for minimization of inference error

Gamma-Poisson model We wish to find the weights w_d which minimize the topics reconstruction error. Us-

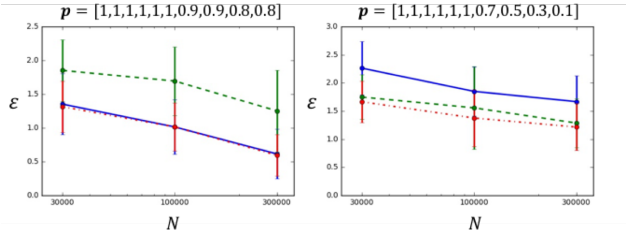

Figure 2: Reconstruction error for the complete dimensions of synthetic data, ε_c, vs. N, where every incomplete dimension has a different probability, p_d, to be observed. A different topics matrix, $\mathbf{A}$, is sampled for each test. Each data point is an average over 25 runs.

ing results from Anandkumar et al. [2012], Podosinnikova et al. [2015] showed that the inference error is bounded by the sum of two contributions, one originating in the uncertainty in estimating $\hat{\mathbf{S}}$, which scales as $\mathbb{E}\left[||\hat{\mathbf{S}}-\mathbf{S}||_F\right]/(\sigma_K(A)L)^2$, and one from $\hat{\mathbf{T}}$, which scaling as $\mathbb{E}\left[||\hat{\mathbf{T}}-\mathbf{T}||_F\right]/(\sigma_K(A)L)^3$, where $\sigma_K(A)$ is the K-th largest singular value of A.

In the following we shall make the assumption that D is significantly larger than K and thus the singular values of the topics matrix will be well approximated by the dimensions with no missing values, that is, we may treat $\sigma_K(A)$ as constant with respect to our choice of weights.

We derive weights that minimize an upper bound of $\mathbb{E}\left[||\hat{\mathbf{S}}-\mathbf{S}||_F\right]/L^2$. The same set of weights $w_d \propto p_d$ approximately minimize $\mathbb{E}\left[||\hat{\mathbf{T}}-\mathbf{T}||_F\right]/L^3$, and the derivation is similar. Thus, we minimize an a quantity that is bounded away from the actual inference error by the sum of the Jensen gaps:

$$\mathbb{E}\left[||\hat{\mathbf{S}}-\mathbf{S}||_F\right] < \sqrt{\mathbb{E}\left[||\hat{\mathbf{S}}-\mathbf{S}||_F^2\right]},$$
$$\mathbb{E}\left[||\hat{\mathbf{T}}-\mathbf{T}||_F\right] < \sqrt{\mathbb{E}\left[||\hat{\mathbf{T}}-\mathbf{T}||_F^2\right]}.$$

We first observe that the weighting of $\hat{\mathbf{S}}$ rescales the Frobenius error in the following way:

$$\begin{aligned}
\mathbb{E}\left[||\hat{\mathbf{S}}^*-\mathbf{S}^*||_F\right] &< \sqrt{\mathbb{E}\left[||\hat{\mathbf{S}}^*-\mathbf{S}^*||_F^2\right]} \\
&= \sqrt{\mathbb{E}\left[\sum_{d_1,d_2=1}^{D}(\hat{S}^*_{d_1d_2}-S^*_{d_1d_2})^2\right]} \\
&= \sqrt{\mathbb{E}\left[\sum_{d_1,d_2=1}^{D}w_{d_1}^2w_{d_2}^2(\hat{S}_{d_1d_2}-S_{d_1d_2})^2\right]} \\
&= \sqrt{\sum_{d_1,d_2}^{D}w_{d_1}^2w_{d_2}^2\mathbb{E}\left[(\hat{S}_{d_1d_2}-S_{d_1d_2})^2\right]}
\end{aligned}$$

We observe that the uncertainty in $\hat{\mathbf{S}}$ scales as $\frac{1}{N_{d_1d_2}}$, where $N_{d_1d_2}$ is the number of times an estimate for $S_{d_1d_2}$ can be calculated from the data, i.e. the number of samples for which both x_{d_1n} and x_{d_2n} are observed - $N_{d_1d_2} = Np_{d_1}p_{d_2}$. Thus, we can choose γ_{d_1,d_2} such that

$$\mathbb{E}\left[\left(\hat{\mathbf{S}}-\mathbf{S}\right)^2_{d_1,d_2}\right] \leq \frac{\gamma_{d_1,d_2}}{Np_{d_1}p_{d_2}}.$$

Note that γ_{d_1,d_2} is independent of the weighting. We can then write:

$$\mathbb{E}\left[||\hat{\mathbf{S}}^*-\mathbf{S}^*||_F\right] \leq \sqrt{\sum_{d_1,d_2}^{D}\frac{\gamma_{d_1,d_2}w_{d_1}^2w_{d_2}^2}{Np_{d_1}p_{d_2}}} \equiv E_S^*.$$

The mean document length, $L = \mathbb{E}\left[\sum_{d_1}^{D}\mathbb{E}\left[X\right]_{d_1}\right]$, is also rescaled by the weights:

$$\begin{aligned}
L^* &= \mathbb{E}\left[\sum_{d_1}^{D}\mathbb{E}\left[X\right]^*_{d_1}\right] \\
&= \mathbb{E}\left[\sum_{d_1}^{D}w_{d_1}\mathbb{E}\left[X\right]_{d_1}\right] \\
&= b\mathbf{w}^\top\mathbf{A}\mathbf{c}
\end{aligned}$$

The goal is to minimize $E_S^*/(L^*)^2$, which is an upper bound for $\mathbb{E}\left[||\hat{\mathbf{S}}^*-\mathbf{S}^*||_F\right]/(L^*)^2$, over the closed unit hypercube in $\mathbb{R}^D$.

In general, we observe that $E_S^*/(L^*)^2$ (and respectively $E_T^*/(L^*)^3$) may not be convex in the choice of weights, thus an analytic derivation for the optima may not be feasible. However, under our existing assumptions, we see that $E_S^*/(L^*)^2$ (and respectively $E_T^*/(L^*)^3$) is coordinate-wise convex and hence we may seek local minima by computing the stationary points of $E_S^*/(L^*)^2$:

$$0 = \frac{\partial}{\partial w_{d^*}} \left(\frac{E_S^*}{(L^*)^2} \right)$$
$$= \frac{\partial}{\partial w_{d^*}} \left(\frac{\sqrt{\sum_{d_1,d_2}^{D} \frac{\gamma_{d_1,d_2} w_{d_1}^2 w_{d_2}^2}{N p_{d_1} p_{d_2}}}}{b \mathbf{w}^\top \mathbf{A} \mathbf{c}} \right).$$

Differentiating, we get

$$0 = \frac{4 \left(\sum_{d_1=1}^{D} \frac{\gamma_{d_1,d^*} w_{d_1}^2 w_{d^*}}{N p_{d_1} p_{d^*}} \right)}{2b \mathbf{w}^\top \mathbf{A} \mathbf{c} \sqrt{\sum_{d_1,d_2}^{D} \frac{\gamma_{d_1,d_2} w_{d_1}^2 w_{d_2}^2}{N p_{d_1} p_{d_2}}}}$$
$$- \frac{b \left[\mathbf{A} \mathbf{c} \right]_{d^*} \sqrt{\sum_{d_1,d_2}^{D} \frac{\gamma_{d_1,d_2} w_{d_1}^2 w_{d_2}^2}{N p_{d_1} p_{d_2}}}}{\left(b \mathbf{w}^\top \mathbf{A} \mathbf{c} \right)^2}$$

which simplifies to

$$\frac{w_{d^*}}{p_{d^*}} = \frac{[\mathbf{A}\mathbf{c}]_{d^*}}{\mathbf{w}^\top \mathbf{A} \mathbf{c}} \left(\frac{\sum_{d_1,d_2}^{D} \frac{\gamma_{d_1,d_2} w_{d_1}^2 w_{d_2}^2}{N p_{d_1} p_{d_2}}}{2 \sum_{d_1=1}^{D} \frac{\gamma_{d_1,d^*} w_{d_1}^2}{N p_{d_1}}} \right).$$

For a high dimensional problem, choosing D to be sufficiently large, we may consider the contributions from w_{d^*} in RHS of the above negligible—in the numerator there are D terms including w_{d^*} in a sum of D^2 terms and in the denominator there is one term out of D. Therefore for large D the expression within the RHS brackets can be considered constant. Thus, we obtain the scaling $w_d^* \propto p_d^*$.

Should we make an additional choice of a constant γ such that

$$\left[\hat{\mathbf{S}} - \mathbf{S} \right]_{d_1,d_2} \leq \frac{\gamma}{N p_{d_1} p_{d_2}},$$

reflecting disregard for the structure in the data, we obtain the following expression for $E_S^*/(L^*)^2$

$$E_S^*/(L^*)^2 = \frac{\sqrt{\frac{\gamma}{N}} \sum_d \frac{w_d^2}{\sqrt{p_d}}}{(b \mathbf{w}^\top \mathbf{A} \mathbf{c})^2}.$$

The above formulation of $E_S^*/(L^*)^2$ is convex and yields a globally optimal weighting with the choice: $w_d \propto \sqrt{p_d}$.

In experimentations we find that the two choices of weights, $w_d \propto \sqrt{p_d}$ and $w_d \propto p_d$ perform indistinguishably.

Generalization to other models In our approach, we expect the optimal weights to depend on the specific model used. Computing the optimal weights for different models requires complexity bounds results to determine how the inference error depends on the moments estimation error and any other parameters which might change with the weighting (an equivalent result to the $\varepsilon \sim \mathbb{E}[||\hat{\mathbf{S}} - \mathbf{S}||_F]/L^2$ scaling presented in the beginning of this appendix). Given these model dependent results, the optimal wrights for every moment can be easily computed by following the same straight-forward method used in this appendix, namely calculating the scaled inference errors and differentiating with respect to the weights.

References

Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

Anastasia Podosinnikova, Francis Bach, and Simon Lacoste-Julien. Rethinking lda: moment matching for discrete ica. In *Advances in Neural Information Processing Systems*, pages 514–522, 2015.