Yale University, Department of Statistics and Data Science

# Synthetic categorical data generation via variational inference

Natalie Doss

July 13, 2019

## Contents

**1 Introduction** **1**

**2 Variational algorithm** **3**

**3 Implementation and comparison to other methods** **5**

**4 Alternatives** **6**

4.1 LLM-NP with Laplace method . . . . . 6
4.2 NLLM-NP with Laplace method . . . . . 7
4.3 LLM-NP with sampling . . . . . 8
4.3.1 Objective (ELBO) . . . . . 8
4.3.2 Variational algorithm . . . . . 9
4.4 LLM-NP with linear approximation and $\beta$ prior . . . . . 9
4.4.1 Objective (ELBO) . . . . . 10
4.4.2 Variational algorithm . . . . . 11
4.4.3 MGF calculation . . . . . 11

# 1 Introduction

Suppose we are given a dataset, and we wish to generate an infinite amount of synthetic data that "look like" the given data in the sense that the synthetic data have the same covariance structure as the original. In particular, we would like to do this for high-dimensional, categorical data. The problem discussed here is really only interesting in a high dimensional setting, and there are well known difficulties with high dimensional density estimation. Moreover, our chosen setting is one in which our covariates are categorical. So for instance, it is not possible to assume the data are Gaussian and estimate the covariance and generate normal data. We need to find a way to estimate the covariance for categorical data.

Here are some instances where solving this problem would be particularly useful. Suppose we had data on self-driving cars, in which accidents are rare. We might wish to ensure that any model we train is sufficiently trained on accidents, but since accidents are extremely rare, they might occur just once or twice in a finite dataset. If we could generate a larger dataset, we could ensure that enough accidents are in the training set so that our model would seek to perform well on accidents.

For example, say we took a survey of $U$ users, asking them about $I$ categorical items (favorite food, favorite music, favorite car type, and so on). They can only select one out of $K$ choices per item. We expect there to be correlation within and between items (if your favorite food is caviar, maybe your favorite music type is classical, and so on).

Here is some notation we will use throughout. Throughout, let $\phi_{\mu,\Sigma}$ indicate the density of the $N(\mu, \Sigma)$ distribution. I will use the indices:

$$\begin{aligned} u, v &\in [U] && \text{(user/sample)} \\ i, j &\in [I] && \text{(covariate/item)} \\ k, l &\in [K] && \text{(category)} \end{aligned}$$

We use $X$ to indicate the full dataset and $x_u$ to indicate a single draw. Let $z_u \in \mathbb{R}^d$.

We propose a latent embedding model as follows. Let the latent variable $z_u$ represents an embedding of user $u$, and $\beta_{ik}$ represents an embedding of category $k$ of item $i$. We want to allow for correlation both within and between the covariates. For instance, we might have several $z_u$ vectors that represent "healthy people." These probably have a large inner product with the $\beta_{ik}$'s associated with favorite food being vegetables and favorite activity being exercise. The model parameters are $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}, \beta_{ik} \in \mathbb{R}^p$. Let $\eta : \mathbb{R}^d \to \mathbb{R}^p$. The model is

$$\begin{aligned} &z_u \sim N(\mu, \Sigma). \\ &X_{ui} \sim_{i.i.d.} \mathit{Multinom}\left(\pi_{ui1}, \ldots, \pi_{uiK}\right), \text{ where } \pi_{uik} = \frac{\exp(\eta(z_u)'\beta_{ik})}{\sum_{l\in[K]} \exp(\eta(z_u)'\beta_{il})}. \end{aligned} \tag{1}$$

That is, $\mathbb{P}\{X_{ui} = k | z_u, \beta_{ik}\} = \pi_{uik}$. In general, the function $\eta$ may depend on many parameters $\theta$, and may have a complex form. In the simplest form of Model (3), we will let $\eta$ be the identity function.

For each $u \in [U]$, let $x_{ui} \in \mathbb{R}^K$ be the binary encoding of the categorical variable. I will write the observed data matrix as $X \in \mathbb{R}^{U\times IK}$. Let $B_i = (\beta_{i1}, \ldots, \beta_{iK}) \in \mathbb{R}^{p\times K}$, and let $B$ be collection of matrices $B_1, \ldots, B_I$. Let $x_{1:U}$ indicate the vectors $x_1, \ldots, x_U$, each in $\mathbb{R}^{IK}$, and let $z_{1:U}$ indicate the vectors $z_1, \ldots, z_U$, each in $\mathbb{R}^d$. Let $b_i = \sum_{l\in[K]} \exp(\eta(z)'\beta_{il})$. In summary, the model parameters are $\theta = (\theta_1, \theta_2) = ((\mu, \Sigma), B)$ when the function $\eta$ is linear. If we allow $\eta$ to have a more complex form, we would have additional parameters. For a single user $u$, the joint likelihood is:

$$p(x_u, z_u, |B, \mu, \Sigma) = \left( \prod_{i,k} \pi_{uik}^{x_{uik}} \right) \phi_{\mu,\Sigma}(z_u)$$

The joint across $U$ independent users is

$$p(x_{1:U}, z_{1:U}|B, \mu, \Sigma) = \left(\prod_{u,i,k} \pi_{uik}^{x_{uik}}\right)\left(\prod_u \phi_{\mu,\Sigma}(z_u)\right)$$

Now let $b_{ui} = \sum_{l \leq K} \exp\left(\eta(z_u)'\beta_{il}\right)$. Let $f_\theta(z) = \log p_\theta(x_{1:U}, z_{1:U})$. Then

$$f_\theta(z) \stackrel{c}{=} \sum_{u,i,k} x_{uik}\eta(z_u)'\beta_{ik} - \sum_{u,i} \log b_{ui} + \sum_u \log \phi_{\mu,\Sigma}(z_u). \tag{2}$$

In the second term, we used the fact that $\sum_{u,i,k} x_{uik} \log b_{ui} = \sum_{ui} \log b_{ui}$.

# 2 Variational algorithm

We seek to estimate the parameters $\theta$. In particular, we are interested in $\Sigma$. Then our estimate, we can generate data from the model (3). For likelihoods that are intractable, the Expectation Maximization (EM) algorithm is a good alternative. Since in our case, we cannot compute the $E$-step in closed form, we propose to use a variational EM algorithm. See my tutorial on variational inference for more information. We will assume that $q$ is the density for the $N(\lambda, V)$ density, where $\lambda \in \mathbb{R}^p$ and $V \in \mathbb{R}^{p \times p}$ is diagonal. This is the mean-field assumption; every $q(z) = \prod_{u \in U} q_u(z_u)$, where $q_u(z_u)$ is the $N(\lambda_u, v_u)$ density. Our objective is thus Here, as usual, the objective is

$$\begin{aligned}
\mathcal{L}(\theta) &= \mathbb{E}_q \log p(x, \theta) - \mathbb{E}_q \log q \\
&= \mathbb{E}_{q(z)} f_\theta(z) + \frac{1}{2}\log|V|,
\end{aligned}$$

where $f_\theta(z)$ is as in (2). We cannot directly calculate $\mathbb{E}_{q(z)} b_{ui}$; this is where we have an issue with non-conjugacy. We introduce a new set of variational parameters, $\zeta_u$. For any $x, \zeta > 0$,

$$\log x = \log(x/\zeta) + \log \zeta \leq (x/\zeta) - 1 + \log \zeta$$

Using this and Jensen,

$$\begin{aligned}
\mathbb{E}_q \log b_{ui} \leq \log(\mathbb{E}_q b_{ui}) &\leq \zeta_u^{-1}\mathbb{E}_q b_{ui} - 1 + \log \zeta_u \\
&= \left(\zeta_u^{-1} \sum_{l \leq K} \mathbb{E}_{q(z_u)} \exp(z_u'\beta_{il})\right) - 1 + \log \zeta_u \\
&= \left(\zeta_u^{-1} \sum_{l \in [K]} \exp\left(\lambda_u'\beta_{il} + \beta_{il}'V_u\beta_{il}/2\right)\right) - 1 + \log \zeta_u
\end{aligned}$$

So our objective is

$$\begin{aligned}\mathcal{L}(\lambda_{1:U}, V_{1:U}, B, \mu, \Sigma) \stackrel{c}{=} &\sum_{u,i,k} x_{uik}\lambda_u'\beta_{ik} - \sum_{u,i,k} \zeta_u^{-1} \exp\left(\lambda_u'\beta_{ik} + \beta_{ik}'V_u\beta_{ik}/2\right) - I\sum_u \log\zeta_u \\ &- \frac{1}{2}\sum_u \left((\lambda_u - \mu)'\Sigma^{-1}(\lambda_u - \mu) + tr(V_u\Sigma^{-1})\right) \\ &+ \frac{1}{2}\sum_u \left(\log|\Sigma|^{-1} + \log|V|\right)\end{aligned}$$

**$E$-step of variational EM** We can obtain $\hat{\zeta}$ in closed form, but not $\hat{\lambda}, \hat{\nu}_{ks}^2$. We do a coordinate ascent algorithm, maximizing in $\zeta$, then $\lambda$, then $\zeta$ again, then $V$. We repeat until convergence, using the thresholds given in the code. We maximize by doing the conjugate gradient algorithm for $\lambda$ and Newton's method in the log space for $V$. In this section, I drop the $u$ index on $x_u, \lambda_u, V_u$. Recall that $V$ has diagonal entries $v_1^2, \ldots, v_d^2$. I sometimes write this set of entries as a vector $V \in \mathbb{R}^d$.

$$\hat{\zeta} = \frac{\sum_{i,k} \exp\left(\beta_{ik}'\lambda + \beta_{ik}'V\beta_{ik}/2\right)}{I}$$

The gradients are

$$\nabla_\lambda\mathcal{L}(\lambda) = \sum_{i,k} x_{ik}\beta_{ik} - \zeta^{-1}\sum_{i,k} \beta_{i,k}\exp\left(\beta_{ik}'\lambda + \beta_{ik}'V\beta_{ik}/2\right) - \Sigma^{-1}(\lambda - \mu)$$

and

$$\frac{\partial\mathcal{L}(v_s^2)}{\partial v_s^2} = -\zeta^{-1}\sum_{i,k}\frac{\beta_{iks}^2}{2}\exp\left(\lambda'\beta_{ik} + \beta_{ik}'V\beta_{ik}/2\right) - \frac{\Sigma_{s,s}^{-1}}{2} + \frac{1}{2v_s^2}$$

And

$$\frac{\partial\mathcal{L}^2(v_s^2)}{\partial(v_s^2)^2} = -\zeta^{-1}\sum_{i,k}\frac{\beta_{iks}^4}{4}\exp\left(\beta_{ik}'\lambda + \beta_{ik}'V\beta_{ik}/2\right) - \frac{1}{2v_s^4}$$

We must restrict the $v_s^2$'s to be positive, so we will do Newton's method in the log space. Recall that Newton's algorithm's updates are:

$$x = x - \frac{f'(x)}{f''(x)}$$

Operating in the log space, let $z = \log x$, so $x = e^z$. So our function $f(x) = f(e^z)$ and we view $z$ as our function argument here.

$$\begin{aligned}\frac{\partial f(e^z)}{\partial z} &= e^z f'(e^z) = xf'(x) \text{ and} \\ \frac{\partial^2 f(e^z)}{\partial^2 z} &= e^{2z}f''(e^z) + e^z f'(e^z) = x^2 f''(x) + xf'(x)\end{aligned}$$

So Newton's algorithm is

$$\log x = \log x - \frac{xf'(x)}{x^2 f''(x) + xf'(x)} = \log x - \frac{f'(x)}{xf''(x) + f'(x)}$$

**$M$-step of variational EM**

$$\hat{\mu} = \frac{1}{U}\sum_u \lambda_u$$

$$\hat{\Sigma} = \frac{1}{U}\left(\sum_u V_u + (\lambda_u - \hat{\mu})(\lambda_u - \hat{\mu})'\right)$$

There is no closed-form solution for $\hat{\beta}_{ik}$. We will do a gradient ascent algorithm and will plug in all parameters that are already estimated. Now

$$\nabla_{\beta_{ik}} L(\beta_{ik}) = \sum_u \left(x_{uik}\lambda_u - \zeta_u^{-1}(\lambda_u + V_u\beta_{ik})\exp\left(\beta_{ik}'\lambda_u + \beta_{ik}'V_u\beta_{ik}/2\right)\right)$$

# 3 Implementation and comparison to other methods

Our model (3) for correlated categorical variables is closely related to the Correlated Topic Model (CTM) of Blei & Lafferty (2007). That model is as follows. I put it in my notation rather than that of the original paper, to make the correspondence clear. Here, let $\mu \in \mathbb{R}^K, \Sigma \in \mathbb{R}^{K\times K}$, with $\Sigma$ being positive definite. The latent dimension is now $p = K$, the number of categories. The correlated topic model assumes that a document with $U$ words is generated by the following process. Let there be $N$ words in the dictionary. Fix $\alpha_1, \ldots, \alpha_K$, all probability vectors in $\mathbb{R}^N$.

$$\begin{aligned} z_u &\sim N(\mu, \Sigma) \\ \pi_{uk} &= \frac{\exp z_{uk}}{\sum_{l\in[K]} \exp z_{ul}} \\ x_u &\sim_{i.i.d.} Multinom\left(\pi_1, \ldots, \pi_K\right) \quad \leftarrow \text{topic/category assignment} \\ W_u &\sim Multinom(\alpha_{x_u}) \quad \leftarrow \text{word} \end{aligned} \tag{3}$$

Note the correspondence between our model and the CTM. In (3), let $\beta_{ik} = I$ and $\eta$ be the identity function. Let $p = K$. And in the CTM, let $N = K$, and let $\alpha_1 = (1, 0, \ldots, 0)$, $\alpha_2 = (0, 1, 0, \ldots, 0)$, and so on. That is, whenever we draw topic $k$, we are guaranteed that the corresponding word is word $k$. The words correspond exactly to the topics; there is one word per topic. Now the CTM corresponds to our simplified model almost exactly, because we observe words that are actually just the topics themselves, just as in our model, we observe the categories/topics.

We implement the algorithm from Section 4.4.2 in $C$, building off of the code of Blei & Lafferty (2007). We generate data from the model and test the algorithm. Perhaps unsurprisingly, we find that our algorithm tends to underestimate the covariance matrix $\Sigma$ that we are seeking. This was worrisome, and we found a way to test whether it was simply a bug in the implementation or a fact about the proposed algorithm. It turns out to be the latter. To check it, I ran the original CTM code, but on data generated according to the process described above. The CTM correctly estimated the $\alpha$ vectors to be about 1 on one entry and zero on the others. It showed a similar shrinkage in the covariance matrix estimation.

We could also estimate Model (3) using a variational autoencoder (VAE). The main differences in what we implemented are that we use mean-field variational inference rather than the amortized mean field inference of VAE. That is, instead of estimating $q_u(z_u)$, the VAE estimates $q_\phi(z_u)$. See my variational inference sheet for a more complete discussion of VAE's and how they compare to non-amortized variational inference. When we estimated this model using the VAE, we found less shrinkage of the covariance matrix.

We similarly estimated our model using the CTM code from the Wang & Blei (2013) paper. This also implements the CTM model but now uses the Laplace and Delta methods (described therein) to estimate it. These methods showed less shrinkage of the covariance matrix.

# 4 Alternatives

This section provides some methods that we might also use to estimate this model. We didn't implement these. I use the abbreviate "LLM" for Linear logistic model, for when $\eta$ is the identity function, and "NLLM" for Non-linear logistic model, for when $\eta$ is some non linear function, as in a deep neural network. The abbreviation "P" is for parametric, as in amortized mean-field variational inference, and "NP" is for non-parametric, as in the non amortized (classical) mean-field variational inference.

## 4.1 LLM-NP with Laplace method

The calculations are as in Section 4.2, but now $\eta$ is the identity function. The following are the gradient and Hessian for a single $z = z_u$; I drop all subscripts $u$ for now. Now

$$\nabla f(z) = \sum_{i,k} x_{ik}\beta_{ik} - \sum_{i,k} \beta_{ik}\pi(z, \beta_{ik}) - \Sigma^{-1}(z - \mu)$$

And

$$\nabla^2 f(z)_{s,t} = -\sum_{i,k} \beta_{iks}\pi(z, \beta_{ik}) \left( \beta_{ikt} - \sum_{l \in [K]} \beta_{ilt}\pi(z, \beta_{il}) \right) - \Sigma^{-1}_{s,t}$$

And once we have $q$, we know that our objective is as follows. I let $\hat{\lambda}_u := \hat{\lambda}(x_u)$ and similarly for $\hat{V}_u$. I use $\hat{q}$ to indicate $\hat{q}(x_1), \ldots, \hat{q}(x_U)$. Let $\xi_u \sim_{i.i.d.} N(0, I_d)$. Using a single sample to approximate $\mathbb{E}_q \log b_{ui}$,

$$\begin{aligned} \mathcal{L}(\hat{q}) \approx \sum_{u,i,k} x_{uik}\hat{\lambda}_u'\beta_{ik} - \sum_{u,i} \mathbb{E}_{\hat{q}} \sum_{l \in [K]} \exp\left( (\hat{\lambda}_u + \hat{V}_u^{1/2}\xi_u)'\beta_{il} \right) - \\ \frac{1}{2} \sum_u \left( (\hat{\lambda}_u - \mu)'\Sigma^{-1}(\hat{\lambda}_u - \mu) + tr(V\Sigma^{-1}) + \log|V| \right) \end{aligned}$$

*E*-**step:** For each $u \in [U]$, $\hat{q}(z_u)$ is the $N(\hat{\lambda}_u, \hat{V}_u)$ density, where

$$\begin{aligned}\hat{\lambda}_u &= \hat{z}_u = \operatorname{argmax} f(z_u)\\ \hat{V}_u &= -\nabla^2 f(\hat{z}_u)^{-1}\end{aligned}$$

These are both found using the gradient and Hessian, calculated above.

*M*-**step:** Using the objective after finding $\hat{q}$, we see that:

$$\begin{aligned}\hat{\mu} &= \frac{1}{U}\sum_u \hat{\lambda}_u\\ \hat{\Sigma} &= \frac{1}{U}\sum_u (\hat{\lambda}_u - \hat{\mu})(\hat{\lambda}_u - \hat{\mu})' + \frac{1}{U}\sum_u \hat{V}_u\end{aligned}$$

For $\beta_{ik}$, we don't have an analytic solution, but we can do gradient ascent. The gradient is:

$$\begin{aligned}\nabla_{\beta_{ik}}\mathcal{L} &= \sum_u x_{uik}\hat{\lambda}_u - \sum_u \frac{(\hat{\lambda}_u + \hat{V}_u^{1/2}\xi_u)\exp\left((\hat{\lambda}_u + \hat{V}_u^{1/2}\xi_u)'\beta_{ik}\right)}{\sum_{l\le K}\exp\left((\hat{\lambda}_u + \hat{V}_u^{1/2}\xi_u)'\beta_{il}\right)}\\ &= \sum_u x_{uik}\hat{\lambda}_u - \sum_u \hat{a}_u \pi(\hat{a}_u, \beta_{ik})\end{aligned}$$

where $\hat{a}_u = \hat{\lambda}_u + \hat{V}_u^{1/2}\xi_u$.

If we have the identity $\beta$, everything is just as in Section 4.1, except now the gradients are as follows. Note how this matches the CTM calculations of Wang & Blei (2013) for the latent variable in that model. That is, replace their $t(z)$ with our $\sum_i x_i$ where $x_i \in \mathbb{R}^K$; everything is just the same.

$$\begin{aligned}\nabla f(z) &= \sum_i x_i - I\pi - \Sigma^{-1}(z-\mu)\\ \nabla^2 f(z)_{st} &= \pi_s(\mathbf{1}\{s=t\} - \pi_t) - \Sigma^{-1}_{st}\end{aligned}$$

Now we don't have $B$ in the model anymore; we just have $\mu, \Sigma$. And their updates will be as in Section 4.1.

## 4.2 NLLM-NP with Laplace method

Let $J(\eta)$ be the Jacobian of $\eta$; note $J(\eta) \in \mathbb{R}^{p\times d}$. And let

$$H(\eta) = (H(\eta_1), \ldots, H(\eta_p))$$

be a tensor that is the array of the Hessians of the components of $\eta$. So we write

$$H(\eta)'_{s,t}\beta := \sum_{j\le p} \frac{\partial^2 \eta_j(z)}{\partial z_s \partial z_t}\beta_j$$

I sometimes abbreviate $J(\eta), H(\eta)$ to just $J, H$. And I sometimes drop the $\eta(z)$ and just write $\eta$. Now

$$\nabla f(z) = \sum_{i,k} x_{ik} J(\eta)'\beta_{ik} - \sum_{ik} \frac{J(\eta)'\beta_{ik}\exp(\eta'\beta_{ik})}{b_i} - \Sigma^{-1}(z-\mu)$$

We cannot find $\hat{z}$ in closed form, but we can do gradient ascent or some other algorithm to find it. And for the Hessian, first recall that for any function $h(z)$,

$$\frac{\partial^2(\log h(z))}{\partial z^2} = \frac{h''(z)}{h(z)} - \left(\frac{h'(z)}{h(z)}\right)^2$$

Now

$$\begin{aligned}\nabla^2 f(z)_{s,t} = &\sum_{i,k} x_{ik} H(\eta)'_{s,t}\beta_{ik} \\ &- \sum_i \left( \sum_k \frac{\exp(\eta'\beta_k)\left(H'_{s,t}\beta_k + (J'_{[,s]}\beta_k * J'_{[,t]}\beta_k)\right)}{b_i} - \frac{\sum_k J'_{[,s]}\beta_k \exp(\eta'\beta_l)\sum_l J'_{[,t]}\beta_l \exp(\eta'\beta_l)}{b_i^2}\right) \\ &- \Sigma^{-1}_{s,t}\end{aligned}$$

Now if we do a variational-EM algorithm, the $M$-step will involve taking derivatives of $f_{\theta,x}(z)$ with respect to the parameters $\theta$. If we have the $N(\mu,\Sigma)$ prior, then $\theta_1 = (\mu,\Sigma)$, and the updates are the sufficient statistics as in Section 2. For $\tilde{\theta}_2$, we will need the derivatives of $\eta$ with respect to these parameters. And as in the discussion in the variational inference sheet, we can approximate the integral via sampling, since we have from the $E$-step the $\hat{\lambda}, \hat{V}$ for $q$. Suppose we approximate the integral via one sample $\xi$. Write $\tilde{\eta}_u = \eta(\hat{\lambda}(x_u) + \hat{V}^{1/2}(x_u)\xi_u)$.

$$\nabla_{\beta_{ik}}\mathcal{L} = \sum_u x_{uik}\tilde{\eta}_u - \sum_u \frac{\tilde{\eta}_u \exp(\tilde{\eta}'_u\beta_{ik})}{\sum_l \exp(\tilde{\eta}'_u\beta_{il})}$$

## 4.3 LLM-NP with sampling

### 4.3.1 Objective (ELBO)

Now instead of introducing $\zeta$ to compute $\mathbb{E}_{q(z_u)}\log b_{ui}$, we do the following.

$$\begin{aligned}\mathbb{E}_{q(z_u)}\log\sum_{k\le K}\exp\left(z'_u\beta_{ik}\right) &= \mathbb{E}_{\xi_u\sim N(0,I_d)}\log\sum_{k\le K}\exp\left(\lambda'_u\beta_{ik} + \xi'_u V_u^{1/2}\beta_{ik}\right) \\ &\approx \log\sum_{k\le K}\exp\left(\lambda'_u\beta_{ik} + \xi'_u V_u^{1/2}\beta_{ik}\right)\end{aligned}$$

where $\xi_u \sim N(0, I_d)$. That is, I'm approximating the integral with a single draw from the distribution. We could use more draws to get a better approximate. Our full ELBO now is:

$$\begin{aligned}\mathcal{L}(\lambda_{1:U}, V_{1:U}, B, \mu, \Sigma) = &\sum_{u,i,k} x_{uik}\lambda_u'\beta_{ik} - \sum_{u,i} \log \sum_{k\leq K} \exp\left(\lambda_u'\beta_{ik} + \xi_u' V_u^{1/2}\beta_{ik}\right) \\ &+ U\log|\Sigma^{-1}| - \frac{1}{2}\sum_u \left((\lambda_u - \mu)'\Sigma^{-1}(\lambda_u - \mu) + tr(V_u^{1/2}\Sigma^{-1}V_u^{1/2})\right) \\ &+ \frac{\sum_u \log|V_u|}{2}\end{aligned}$$

### 4.3.2 Variational algorithm

The gradients are (dropping indices for now):

$$\nabla_\lambda \mathcal{L}(\lambda) = \sum_{i,k} x_{ik}\beta_{ik} - \sum_{i\leq I} \frac{\sum_{k\leq K} \beta_{ik} \exp\left(\lambda'\beta_{ik} + \xi' V^{1/2}\beta_{ik}\right)}{\sum_{l\leq K} \exp\left(\lambda'\beta_{il} + \xi' V^{1/2}\beta_{ik}\right)} - \frac{1}{2}\Sigma^{-1}(\lambda - \mu)$$

$$\begin{aligned}\frac{\partial \mathcal{L}(v_s^2)}{\partial v_s^2} &= -\sum_{i\leq I} \frac{\sum_{k\leq K} \frac{\xi_s\beta_{iks}}{2v_s} \exp\left(\lambda'\beta_{ik} + \xi' V^{1/2}\beta_{ik}\right)}{\sum_{l\leq K} \exp\left(\lambda'\beta_{il} + \xi' V^{1/2}\beta_{ik}\right)} - \frac{\Sigma_{s,s}^{-1}}{2} + \frac{1}{2v_s^2} \\ &= -\frac{\xi_s}{2v_s}\sum_{i\leq I} \frac{\sum_{k\leq K} \beta_{iks} \exp\left(\lambda'\beta_{ik} + \xi' V^{1/2}\beta_{ik}\right)}{\sum_{l\leq K} \exp\left(\lambda'\beta_{il} + \xi' V^{1/2}\beta_{ik}\right)} - \frac{\Sigma_{s,s}^{-1}}{2} + \frac{1}{2v_s^2}\end{aligned}$$

For the $M$-step, the solutions for $\mu, \Sigma$ are the same as in previous sections. And

$$\nabla_{\beta_{ik}} \mathcal{L}(\beta_{ik}) = \sum_u x_{uik}\lambda_u - \sum_u \frac{\left(\lambda_u + V_u^{1/2}\xi_u\right) \exp\left(\lambda_u'\beta_{ik} + \xi_u' V_u^{1/2}\beta_{ik}\right)}{\sum_{l\leq K} \exp\left(\lambda_u'\beta_{il} + \xi_u' V_u^{1/2}\beta_{ik}\right)}$$

## 4.4 LLM-NP with linear approximation and $\beta$ prior

Let $z_u, \beta_{ik} \in \mathbb{R}^d$. And let $\nu, \mu \in \mathbb{R}^d$ and $\Omega, \Sigma \in \mathbb{R}^{d\times d}$. Now we place a prior on $\beta$; here is the data-generating process.

$$z_u \sim N(\mu, \Sigma) \tag{4}$$

$$\beta_{ik} \sim N(0, \gamma^2 I_d) \tag{5}$$

$$\mathbb{P}\{X_{ui} = k | z_u, \beta_{ik}\} = \frac{\exp(z_u'\beta_{ik})}{\sum_{l\in[K]} \exp(z_u'\beta_{il})} \tag{6}$$

For variational inference, we use the families:

$$q(z_u) = N(\lambda_u, V_u)$$
$$q(\beta_{ik}) = N(\psi_{ik}, W_{ik})$$

where $W_u, V_{ik}$ are all diagonal matrices with entries $v_{iks}, w_{us}$ for $s \in [d]$. We find the parameters of $q$ to maximize

$$\mathbb{E}_{z\sim q} \log p(x, \theta, B) - \mathbb{E}_q \log q$$

#### 4.4.1 Objective (ELBO)

The joint across $U$ independent users is

$$p(x_{1:U}, z_{1:U}, B) = \left(\prod_{u,i,k} \pi(z_u, \beta_{ik})^{x_{uik}}\right)\left(\prod_{u} \phi_{\mu,\Sigma}(z_u)\right)\left(\prod_{i,k} \phi_{0,\gamma^2 I_d}(\beta_{ik})\right)$$

Now let $b_{ui} = \sum_{l\le K} \exp(z_u'\beta_{il})$. We have

$$\log p(x, \theta, B) \stackrel{c}{=} \sum_{u,i,k} x_{uik} z_u' \beta_{ik} - \sum_{u,i} \log b_{ui} + \sum_{u} \log \phi_{\mu,\Sigma}(z_u) + \sum_{i,k} \log \phi_{0,\gamma^2 I_d}(\beta_{ik})$$

Note that in the second term, we used the fact that $\sum_{u,i,k} x_{u,i,k} b_{ui} = \sum_{ui} b_{ui}$. Now to help handle the expectation of this term, we introduce a new set of variational parameters, $\zeta_u$. Note that for any $x, \zeta > 0$,

$$\log x = \log(x/\zeta) + \log \zeta \le (x/\zeta) - 1 + \log \zeta$$

Using this and Jensen,

$$\begin{aligned}\mathbb{E}_q \log b_{ui} &\le \log(\mathbb{E}_q b_{ui}) \le \zeta_u^{-1} \mathbb{E}_q b_{ui} - 1 + \log \zeta_u \\ &= \left(\zeta_u^{-1} \sum_{l\le K} \mathbb{E}_q \exp(z_u' \beta_{il})\right) - 1 + \log \zeta_u\end{aligned}$$

We have by the diagonality of the variance matrices and the independence of $\theta_u, \beta_{ik}$,

$$\mathbb{E}_{q(\beta_{ik})} \mathbb{E}_{q(z_u)} e^{z_u' \beta_{il}} = \prod_{s\in[d]} f_{uils}$$

where $f_{uils} = \mathbb{E}_{q(z_u)} \mathbb{E}_{q(\beta_{ik})} \exp(z_u[s]\beta_{il}[s])$. See Section 4.4.3 for its full form.

$$\mathbb{E}_q \log b_{ui} \le \left(\zeta_u^{-1} \sum_{l\in[K]} \prod_{s\in[d]} f_{uils}\right) - 1 + \log \zeta_u$$

And

$$\begin{aligned}H(q(\theta_u))) &= \frac{-\log|V_u^{-1}|}{2} = \frac{\log|V_u|}{2} \\ H(q(\beta_{ik})) &= \frac{-\log|W_{ik}^{-1}|}{2} = \frac{\log|W_{ik}|}{2}\end{aligned}$$

So our full ELBO is:

$$\begin{aligned}&\sum_{u,i,k} x_{uik} \lambda_u' \lambda_{ik} - \left(\sum_{u,i} \zeta_u^{-1} \sum_{l\in[K]} \prod_{s\in[d]} f_{uils}\right) - \sum_{u,i} \log \zeta_u + U \log|\Sigma^{-1}| + IKd \log \frac{1}{\gamma^2} \\ &- \frac{1}{2} \sum_u \left((\lambda_u - \mu)' \Sigma^{-1} (\lambda_u - \mu) + tr(V_u^{1/2} \Sigma^{-1} V_u^{1/2})\right) - \frac{1}{2} \sum_{i,k} \left(\frac{\|\psi_{ik}\|_2^2}{\gamma^2} + \frac{tr(W_{ik})}{\gamma^2}\right) \\ &+ \frac{\sum_u \log|V_u| + \sum_{ik} \log|W_{ik}|}{2}\end{aligned}$$

### 4.4.2 Variational algorithm

**Variational E-step updates** I'm letting $v_u, w_{ik}$ be the vectors in question for the diagonal matrices. The elements are $v_{us}$ or $v_u[s]$, either way.

$$\partial_{\lambda_{ut}} L(\lambda_{ut}) = \sum_{i,k} x_{uik}\psi_{ik}[t] - \Sigma^{-1}(\lambda_u - \mu)[t] - \sum_{i,k} \frac{\partial f(\lambda_{ut}, v_{ut}^2, \psi_{ikt}, w_{ikt}^2)}{\partial \lambda_{ut}} \prod_{s \neq t} f(\lambda_{us}, v_{us}^2, \psi_{iks}, w_{iks}^2)$$

$$\partial_{\psi_{ikt}} L(\psi_{ikt}) = \sum_{u} x_{uik}\lambda_u[t] - \frac{1}{\gamma^2}\psi_{ik}[t] - \sum_{u} \frac{\partial f(\lambda_{ut}, v_{ut}^2, \psi_{ikt}, w_{ikt}^2)}{\partial \psi_{ikt}} \prod_{s \neq t} f(\lambda_{us}, v_{us}^2, \psi_{iks}, w_{iks}^2)$$

$$\partial_{v_{ut}} L(v_{ut}^2) = -\frac{diag(\Sigma^{-1})[t]}{2} + \frac{1}{2v_u^2[t]} - \sum_{i,k} \frac{\partial f(\lambda_{ut}, v_{ut}^2, \psi_{ikt}, w_{ikt}^2)}{\partial v_{ut}^2} \prod_{s \neq t} f(\lambda_{us}, v_{us}^2, \psi_{iks}, w_{iks}^2)$$

$$\partial_{w_{ikt}} L(w_{ikt}^2) = -\frac{1}{2} + \frac{1}{2w_{ik}^2[t]} - \sum_{u} \frac{\partial f(\lambda_{ut}, v_{ut}^2, \psi_{ikt}, w_{ikt}^2)}{\partial w_{ikt}^2} \prod_{s \neq t} f(\lambda_{us}, v_{us}^2, \psi_{iks}, w_{iks}^2)$$

And in closed form, for each $u \in [U]$,

$$\hat{\zeta}_u = \frac{-\sum_{i,k} \prod_{s \leq d} f_{uiks}}{I}$$

**Sufficient statistics for $M$ step**

$$\hat{\mu} = \frac{1}{U}\sum_u \lambda_u$$
$$\hat{\Sigma} = \frac{1}{U}\sum_u (\lambda_u - \hat{\mu})(\lambda_u - \hat{\mu})' + \frac{1}{U}\sum_u V_u$$
$$\hat{\gamma}^2 = \frac{1}{IKd}\sum_{i,k} \left( \|\psi_{ik}\|_2^2 + tr(W_{ik}) \right)$$

### 4.4.3 MGF calculation

To evaluate $\mathbb{E}_{z_u \sim q}\mathbb{E}_{\beta_{ik} \sim q)} \exp(z_u'\beta_{ik})$, first consider the following. Let

$$z \sim N(\lambda, v^2)$$
$$\beta \sim N(\psi, w^2)$$

Then

$$
\begin{aligned}
\mathbb{E}_\beta \mathbb{E}_z \exp(z\beta) &= \mathbb{E}_\beta \exp\left(\lambda\beta + v^2\beta^2/2\right) \\
&= \mathbb{E}_\beta \exp\left(\frac{v^2}{2}\left(\beta^2 + \frac{2\lambda}{v^2}\beta + \frac{\lambda^2}{v^4} - \frac{\lambda^2}{v^4}\right)\right) \\
&= \mathbb{E}_\beta \exp\left(\frac{v^2}{2}\left(\beta + \frac{\lambda}{v^2}\right)^2 - \frac{\lambda^2}{2v^2}\right) \\
&= e^{-\lambda^2/2v^2}\mathbb{E}_\beta \exp\left(\frac{v^2}{2}\left(\beta + \frac{\lambda}{v^2}\right)^2\right)
\end{aligned}
$$

Now $\beta + \lambda/v^2 \sim N(\psi + \lambda/v^2, w^2)$, so

$$
\begin{aligned}
\beta + \frac{\lambda}{v^2} &\sim wN\left(\frac{\psi v^2 + \lambda}{wv^2}, 1\right) \Rightarrow \\
\left(\beta + \frac{\lambda}{v^2}\right)^2 &\sim w^2\chi_1^2\left(\left(\frac{\psi v^2 + \lambda}{wv^2}\right)^2\right) \\
&\sim \frac{1}{v^4}\chi_1^2\left(\psi^2 v^4 + \lambda^2 + 2\psi\lambda v^2\right)
\end{aligned}
$$

Using the moment-generating function for the non-central chi square distribution, we have:

$$
\begin{aligned}
f(\lambda, v^2, \psi, w^2) = \mathbb{E}_\beta \mathbb{E}_z \exp(z\beta) &= \frac{1}{\sqrt{1 - v^2w^2}}\exp\left(-\frac{\lambda^2}{2v^2}\right)\exp\left(\frac{v^2}{2v^4}\frac{\psi^2v^4 + \lambda^2 + 2\psi\lambda v^2}{1 - v^2w^2}\right) \\
&= \frac{1}{\sqrt{1 - v^2w^2}}\exp\left(\frac{\psi^2v^4 + \lambda^2 + 2\psi\lambda v^2 - \lambda^2 + \lambda^2v^2w^2}{2v^2(1 - v^2w^2)}\right) \\
&= \frac{1}{\sqrt{1 - v^2w^2}}\exp\left(\frac{\psi^2v^2 + 2\psi\lambda + \lambda^2w^2}{2(1 - v^2w^2)}\right)
\end{aligned}
$$

Let $g(\lambda, v^2, \psi, w^2) = \frac{\psi^2v^2 + 2\psi z + \lambda^2w^2}{2(1 - v^2w^2)}$. The gradients of this with respect to each parameter are:

$$
\begin{aligned}
\frac{\partial f}{\partial \lambda} &= \frac{2\lambda w^2 + 2\psi}{2(1 - v^2w^2)^{3/2}}\exp\left(g(\lambda, v^2, \psi, w^2)\right) \\
\frac{\partial f}{\partial \psi} &= \frac{2\psi v^2 + 2\lambda}{2(1 - v^2w^2)^{3/2}}\exp\left(g(\lambda, v^2, \psi, w^2)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial f}{\partial v^2} &= \frac{\partial g/\partial v^2}{2(1 - v^2w^2)^{1/2}}\exp\left(g(\lambda, v^2, \psi, w^2)\right) + \frac{w^2}{2(1 - v^2w^2)^{3/2}}\exp\left(g(\lambda, v^2, \psi, w^2)\right) \\
\frac{\partial f}{\partial w^2} &= \frac{\partial g/\partial w^2}{2(1 - v^2w^2)^{1/2}}\exp\left(g(\lambda, v^2, \psi, w^2)\right) + \frac{v^2}{2(1 - v^2w^2)^{3/2}}\exp\left(g(\lambda, v^2, \psi, w^2)\right)
\end{aligned}
$$

where

$$\frac{\partial g}{\partial v^2} = \frac{(1 - v^2w^2)\lambda^2 + (\psi^2 v^2 + 2\psi\lambda + \lambda^2 w^2)w^2}{2(1 - v^2w^2)^2}$$
$$\frac{\partial g}{\partial w^2} = \frac{(1 - v^2w^2)\lambda^2 + (\psi^2 v^2 + 2\psi\lambda + \lambda^2 w^2)v^2}{2(1 - v^2w^2)^2}$$

# References


Blei, David M., & Lafferty, John D. 2007. A correlated topic model of science. *Annals of Applied Statistics*, **1**(1), 17–35.

Gopalan, Prem, Hofman, Jake M., & Blei, David M. 2014. Scalable Recommendation with Hierarchical Poisson Factorization.

Wang, Chong, & Blei, David M. 2013. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, **14**, 1005–1031.