From Exponential Family (Probabilistic) PCA to Probabilistic Fenchel-Young PCA

Sophia Sklaviadis

Abstract

Principal component analysis (PCA) is a popular dimension reduction, visualization and denoising method, most commonly applied under the assumption of Normally distributed data. Tipping and Bishop [1999] derive classical PCA as a limiting case of probabilistic PCA (PPCA).

To use PCA with discrete and non-Normally distributed data, Collins et al. [2001] extend classical PCA to the Exponential family, while Li and Tao [2010] describe a general Expectation Maximization approach for probabilistic Exponential family PCA. To accomplish variable selection alongside dimension reduction Lu et al. [2016] use a sparsity inducing penalty on the loading matrix. Recently Zeng et al. [2022] use 0-inflated Logistic Normal Multinomial (LNM) PPCA to circumvent sparse data, while Fang and Subedi [2023] use a LNM mixture for clustering.

As an alternative, we model sparsity directly by further extending classical Exponential family PCA to deformed exponential families through Fenchel-Young (FY) losses (Blondel et al. [2020]). We compare LNM PPCA with classical Multinomial PCA and with Fenchel-Young PCA on low dimensional data reconstruction and clustering in variably sparse data settings, and show that FY PCA is a simple and efficient approach to modeling sparse, discrete data, outperforming LNM PPCA. Finally, we derive a Fenchel-Young ELBO for probabilistic FY PCA.

We first contextualize Collins et al. [2001]'s classical Exponential family PCA, implemented through a degenerate version of EM (Murphy [2012], p.947), within the Probabilistic PCA setting by first reviewing the probabilistic model from which Tipping and Bishop [1999] derive classical PCA as a limiting case. We are then in a position to extend Collins et al. [2001]'s Bregman divergence objective to Fenchel-Young losses where we can use Tsallis entropy with $a = 2$ to fit classical Exponential family PCA to a sparse data matrix $\mathbf{X}$ with columns whose (deformed Exponential family) densities lead to the Sparsemax loss (Martins and Astudillo [2016]). Further, in the probabilistic setting of Exponential family PCA described abstractly by Li and Tao [2010], we consider adding Normal priors on the columns of the latent factor matrix $\mathbf{Z}$, leading to Gaussian-Sparsemax latent variables and the FY ELBO objective.

FY PCA and probabilistic FY PCA are alternatives to the approaches previously proposed in the literature on sparse PCA (classical: Zou et al. [2006], prob: Guan and Dy [2009], structured: Jenatton et al. [2010], Exponential family: Lu et al. [2016], Logistic Normal Multinomial: Zeng et al. [2022], etc.), which apply naturally to sparse (count) data.[1]

[1] In passing we mention the connection between Multinomial PCA and skipgram word embedding models, e.g. Cotterell et al. [2017] interpret the skipgram model as Exponential family PCA without situating their interpretation in a probabilistic PCA framework.

1 Preliminaries

Suppose the observed data $\mathbf{X}$ forms a $D \times N$ matrix, whose nth column $\mathbf{x}_n$ is the nth observation, i.e. a realization of a D-dimensional random vector:

$$\mathbf{X}_{D\times N} = [\mathbf{x_1}, ..., \mathbf{x_N}] = \begin{bmatrix} x_{11} \dots x_{1N} \\ \vdots \\ x_{D1} \dots x_{DN} \end{bmatrix} \tag{1}$$

Let each $\mathbf{x}_n$ follow an Exponential family distribution with natural parameter a D-dimensional $\boldsymbol{\theta}_n$:

$$\boldsymbol{\Theta}_{D\times N} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N] = \begin{bmatrix} \theta_{11} \dots \theta_{1N} \\ \vdots \\ \theta_{D1} \dots \theta_{DN} \end{bmatrix} \tag{2}$$

We assume the matrix of natural parameters is factorized into a $K \times N$ matrix $\mathbf{Z}$ of latent factors or latent variables, where K is the dimensionality of the latent representation with $K < D$, and a $D \times K$ matrix $\mathbf{W}$ of parameters, often referred to as factor loadings: $\boldsymbol{\Theta} = \mathbf{W}_{D\times K}\mathbf{Z}_{K\times N}$.[2] So for every observation $\mathbf{x}_n \in \mathbb{R}^D$ there is a corresponding latent variable $\mathbf{z}_n \in \mathbb{R}^K$:

$$\mathbf{Z}_{K\times N} = [\mathbf{z}_1, ..., \mathbf{z}_N] = \begin{bmatrix} z_{11} \dots z_{1N} \\ \vdots \\ z_{K1} \dots z_{KN} \end{bmatrix}, \quad \mathbf{W}_{D\times K} = [\mathbf{w}_1, ..., \mathbf{w}_K] = \begin{bmatrix} w_{11} \dots w_{1K} \\ \vdots \\ w_{D1} \dots w_{DK} \end{bmatrix} \tag{3}$$

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} \dots \theta_{1N} \\ \vdots \\ \theta_{D1} \dots \theta_{DN} \end{bmatrix} = \begin{bmatrix} w_{11} \dots w_{1K} \\ \vdots \\ w_{D1} \dots w_{DK} \end{bmatrix} \begin{bmatrix} z_{11} \dots z_{1N} \\ \vdots \\ z_{K1} \dots z_{KN} \end{bmatrix}, \tag{4}$$

$$\begin{aligned} \boldsymbol{\theta}_n = \begin{pmatrix} \theta_{1n} \\ \vdots \\ \theta_{Dn} \end{pmatrix} &= \begin{pmatrix} w_{11}z_{1n} + ... + w_{1K}z_{Kn} \\ \vdots \\ w_{D1}z_{1n} + ... + w_{DK}z_{Kn} \end{pmatrix} = \begin{pmatrix} w_{11} \\ \vdots \\ w_{D1} \end{pmatrix} z_{1n} + ... + \begin{pmatrix} w_{1K} \\ \vdots \\ w_{DK} \end{pmatrix} z_{Kn} \\ &= \mathbf{w}_1 z_{1n} + ... + \mathbf{w}_K z_{Kn} = \sum_{k=1}^{K} \mathbf{w}_k z_{kn} = \mathbf{W}_{D\times K}\mathbf{z}_{n_{K\times 1}}. \end{aligned} \tag{5}$$

Given a "score vector" $\mathbf{z}_n$ and loadings $\mathbf{W}$, we model observation $\mathbf{x}_n$ by an Exponential family distribution with natural parameter $\boldsymbol{\theta}_n$:

$$\mathbf{x}_n \mid \mathbf{z}_n, \mathbf{W} \sim \text{Expon}(\boldsymbol{\theta}_n) = \text{Expon}\left(\sum_{k=1}^{K} \mathbf{w}_k z_{kn}\right) \tag{6}$$

[2] In Collins et al. [2001] each *row* of $\mathrm{X}^{\mathrm{T}}_{N\times D}$ is made up of D independent random variables, and $\boldsymbol{\Theta}^{\mathrm{T}} = \mathbf{Z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}$.

Subject to this (low rank) factorization of $\mathbf{\Theta}$, the aim of probabilistic PCA is to maximize the marginal likelihood of the observations, $\mathbf{x}_n$'s, assuming a prior over the latent variables and integrating them out.

2 Everything Normal

Recall that in general [3]

$$\mathbb{E}(\mathbf{x}_i) = \mathbb{E}_{\mathbf{z}}\mathbb{E}(\mathbf{x}_i \mid \mathbf{z}_i) = \mathbb{E}_{\mathbf{z}}(\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i) = \boldsymbol{\mu} + \mathbf{W}\boldsymbol{\mu}_0$$

$$\begin{aligned} \mathrm{var}(\mathbf{x}_i) &= \mathrm{var}(\mathbb{E}(\mathbf{x}_i \mid \mathbf{z}_i)) + \mathbb{E}_{\mathbf{z}}(\mathrm{var}(\mathbf{x}_i \mid \mathbf{z}_i)) \\ &= \mathrm{var}(\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i) + \mathbb{E}_{\mathbf{z}}(\sigma^2\mathbf{I}) = \mathbf{W}\mathbf{\Sigma}_0\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I} \end{aligned} \tag{7}$$

Consider now $\mathbf{x}_i = \boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i + \boldsymbol{\epsilon}_i$ with $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ where:

$$\begin{aligned} p(\mathbf{z}_i) &= \mathcal{N}(\mathbf{z}_i \mid \boldsymbol{\mu}_0, \mathbf{\Sigma}_{0_{K\times K}}) \\ p(\mathbf{x}_i \mid \mathbf{z}_i, \mathbf{W}_{D\times K}, \sigma^2) &= \mathcal{N}(\boldsymbol{\mu} + \mathbf{W}_{D\times K}\mathbf{z}_{i_{K\times 1}}, \sigma^2\mathbf{I}) \end{aligned} \tag{8}$$

The induced marginal is:

$$\begin{aligned} p(\mathbf{x}_i \mid \mathbf{W}, \sigma^2) &= \int \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{z}_i \mid \boldsymbol{\mu}_0, \mathbf{\Sigma}_0)d\mathbf{z}_i \\ &= \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu} + \mathbf{W}\boldsymbol{\mu}_0, \mathbf{W}_{D\times K}\mathbf{\Sigma}_{0_{K\times K}}\mathbf{W}^{\mathrm{T}}_{K\times D} + \sigma^2\mathbf{I}) \end{aligned} \tag{9}$$

Call the $D \times D$ covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{\Sigma}_0\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$. Without loss of generality we can set $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\mathbf{\Sigma}_0 = \mathbf{I}$. [4] Notice that $\mathbf{W}$ then appears only in the covariance of the marginal. If the data is centered we also have $\boldsymbol{\mu} = \mathbf{0}$.

Assuming $p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(\mathbf{x}_i \mid \mathbf{z}_i) = \mathcal{N}(\boldsymbol{\mu} + \mathbf{W}\boldsymbol{z}_i, \sigma^2\mathbf{I})$, we have an analytic expression for the posterior of $\mathbf{z}_i$ which is also Normal:

$$p(\mathbf{z}_i \mid \mathbf{x}_i) = \mathcal{N}(\mathbf{z}_i \mid \mathbf{m}_i, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \left(\frac{\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W}}{\sigma^2}\right)^{-1} = \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} = \sigma^2\mathbf{M}^{-1}_{K\times K} \tag{10}$$

$$\mathbf{m}_i = \mathbf{\Sigma}\left(\mathbf{W}^{\mathrm{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right) = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_i - \boldsymbol{\mu})^5$$

[3] McCulloch et al. [2001], pp. 10-12, and pp. 23-36; 4.7: Conditional Expected Value - Statistics LibreTexts, eqn.s 4.7.24-4.7.27.

[4] Cf. Murphy [2012], ch.12.

where $\mathbf{M} = \sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W}$. Notice only $\mathbf{m}_i$ depends on $\mathbf{x}_i$.

Tipping and Bishop [1999] derive MLEs for $\mathbf{W}$ and σ^2 from $p(\mathbf{x}_i \mid \mathbf{W}, \sigma^2)$ and prove that in the limiting case where $\sigma^2 \to 0$ the posterior mean reduces to $\mathbf{m}_i = (\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_i - \boldsymbol{\mu})$ because $\mathbf{M} \to \mathbf{W}^{\mathrm{T}}\mathbf{W}$.[6] The columns of $\mathbf{W}_{ML}$ define the principal subspace of classical PCA, and the posterior means $\mathbf{m}_i$ represent the orthogonal projections of the data $\mathbf{x}_i$ onto the latent space.

2.1 Classical PCA

Classical PCA is usually stated from a complementary perspective to that of the generative model in (7) and (8), as the projection of D-dimensional data to a lower K-dimensional linear subspace ($K << D$). When $\mathbf{Z}_{K\times N} = \mathbf{W}^{\mathrm{T}}_{K\times D}\mathbf{X}_{D\times N}$, the scores of each obsvervation along the kth principal component, given by the kth row of $\mathbf{Z}_{K\times N}$, are each written as a linear combination of the D features with the kth principal component loadings as coefficients:[7]

$$[z_{k1}, ..., z_{kN}] = [w_{1k}x_{11} + ... + w_{Dk}x_{D1}, ..., w_{1k}x_{1N} + ... + w_{Dk}x_{DN}] \tag{11}$$

The ith column of $\mathbf{Z}_{K\times N}$ is the ith latent variable, corresponding to the ith observation.

We briefly mention the intuitive **maximum variance** interpretation of classical PCA.[8] Assuming that $\mathbf{X}_{D\times N}$ is centered, i.e. each column has zero mean, to compute the 1st principal component, we look for the linear combination of the "feature values" with the form[9]

$$z_{1i} = \begin{pmatrix} w_{11}, ..., w_{D1} \end{pmatrix} \begin{pmatrix} x_{1i} \\ \vdots \\ x_{Di} \end{pmatrix} = \begin{pmatrix} w_{11}x_{1i} + ... + w_{D1}x_{Di} \end{pmatrix} \tag{12}$$

that has the largest variance subject to $w_{11}^2 + ... + w_{D1}^2 = 1$, i.e.

$$\max_{w_{11},...,w_{D1}} \frac{1}{N}\sum_{i=1}^{N} z_{1i}^2 = \max \sum_{i=1}^{N} (w_{11}x_{1i} + ... + w_{D1}x_{Di})^2 \tag{13}$$

subject to $\sum_{d=1}^{D} w_{d1}^2 = 1$, and so on for subsequent, orthogonal principal components. This problem, like the minimum error interpretation, to which it is equivalent, is similarly solved by singular value decomposition of the sample covariance matrix (see e.g. Murphy [2012] §12.2.3).

[5] Using eq. 2.111 from Bishop [2006].

[6] Cf. Bishop [2006] §12.2.1; note there is a typo in the var of 12.42.

[7] There are N latent variables corresponding to the N observations. Confusingly, what is sometimes called "factors" are the K principal components of the sample covariance matrix.

[8] Cf. James et al. [2013], p.509.

[9] Compare with §12.2 of James et al. [2013].

Following Murphy [2012] §12.2.1, assuming for notational simplicity that data are centered, i.e. $\mathbf{x}_i$ have zero mean, we want an orthogonal set of K linear basis vectors $\mathbf{w}_k \in \mathbb{R}^D$ and scores $\mathbf{z}_i \in \mathbb{R}^K$ that have **minimum average reconstruction error**:

$$J = \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{x}_i - \hat{\mathbf{x}}_i \|^2 \tag{14}$$

subject to $\mathbf{W}$ being orthonormal, where $\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i$. The optimal solution $\hat{\mathbf{W}}$ is the matrix with columns the K principal eigenvectors of the sample covariance matrix $\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} = \frac{1}{N} \mathbf{X}\mathbf{X}^{\mathrm{T}}$. The optimal low dimensional encoding of the data is given by $\hat{\mathbf{z}}_i = \hat{\mathbf{W}}^{\mathrm{T}} \mathbf{x}_i (= \hat{\mathbf{W}}^{-1} \mathbf{x}_i)$, which is the orthogonal projection of the data onto the latent subspace defined by $\hat{\mathbf{W}}^{\mathrm{T}}$,[10] and corresponds to the MAP estimates $\hat{\mathbf{m}}_i$ in the limiting case when $\sigma^2 \to 0$ (and the data is centered to have zero mean, and we constrain the loading matrix to be orthonormal).

3 Learning alternatives

Instead of maximizing the marginal likelihood $p(\mathbf{x}_i \mid \mathbf{W}, \sigma^2)$ Collins et al. [2001] maximize the likelihood of the conditional density $p(\mathbf{x}_i \mid \mathbf{z}_i)$, dropping the assumption that it follows a Normal distribution, alternately w.r.t. $\mathbf{W}^{\mathrm{T}}$ and $\mathbf{Z}^{\mathrm{T}}$, treating the latent variables as fixed unknown scores. Alternatively, Li and Tao [2010] also maximize the likelihood of the conditional density $p(\mathbf{x}_i \mid \mathbf{z}_i)$ by iteratively taking MAP estimates from the latent posterior $p(\mathbf{z}_i \mid \mathbf{x}_i)$, plugging in the conditional, and maximizing w.r.t. $\mathbf{W}$. Although closer to probabilistic PCA even this latter approach is subject to Welling et al. [2008]'s Bayesian criticism that using a point estimate of the latent posterior is inadequate.[11]

3.1 Maximum likelihood & Bregman divergence

Assuming that the data matrix $\mathbf{X}_{D \times N}$ is centered, the K first principal D-dimensional eigenvectors of the sample covariance $\hat{\Sigma}_{D \times D} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}$ comprise the optimal MLE solution $\hat{\mathbf{W}}_{D \times K}$ that along with the scores $\hat{\mathbf{z}}_n = \hat{\mathbf{W}}^{\mathrm{T}} \mathbf{x}_n$ minimize the square loss $J = \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{x}_i - \hat{\mathbf{x}}_i \|^2$, where $\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i$. These optimal values of $\mathbf{W}$ and $\mathbf{Z}$ are the maximum likelihood estimates of the Normal *conditional* distribution given in (7) in the simplified case where $\boldsymbol{\mu} = \mathbf{0}$ and $\sigma^2 = 1$:

$$p(\mathbf{x}_n \mid \mathbf{z}_n, \mathbf{W}) = \mathcal{N}(\mathbf{W}\mathbf{z}_n, \mathbf{I}) = \mathcal{N}\left(\sum_{k=1}^{K} \mathbf{w}_k z_{kn}, \mathbf{I} \right) \quad \text{i.e.} \quad x_{dn} \mid \mathbf{Z}, \mathbf{W} \sim \mathcal{N}(\theta_{dn}, 1) \tag{15}$$

[10] See Murphy [2012] §12.2.2 for his proof.

[11] Slightly misleadingly Mohamed et al. [2008] suggests that Welling et al. [2008]'s criticism applies to Collins et al. [2001]'s approach.

with mean parameter $\theta_{dn} = w_{d1}z_{1n} + ... + w_{dK}z_{Kn}$ (as in (5) above). The negative log likelihood is given by

$$\begin{aligned}\mathrm{NLL}(\mathbf{W},\mathbf{Z}) &= -\sum_{n=1}^{N}\log p(\mathbf{x}_n \mid \boldsymbol{\theta}_n) = -\sum_{n=1}^{N}\sum_{d=1}^{D}\log p(x_{dn} \mid \theta_{dn}) \\ &= \text{const.} + \sum_{n=1}^{N}\sum_{d=1}^{D}(x_{dn}-\theta_{dn})^2.\end{aligned} \tag{16}$$

Consider the more general case where the observed D-dimensional column vector $\mathbf{x}_n$ (conditionally on $\mathbf{W},\mathbf{Z}$) follows an Exponential family distribution with natural parameter $\boldsymbol{\theta}_n$:

$$\mathbf{x}_n \mid \mathbf{Z},\mathbf{W} \sim \mathrm{Expon}(\boldsymbol{\theta}_n) \tag{17}$$

where $\boldsymbol{\theta}_n = \sum_{k=1}^{K}\mathbf{w}_k z_{kn} = \mathbf{W}_{D\times K}\mathbf{z}_{n_{K\times 1}}$. The negative log likelihood takes the form

$$\begin{aligned}\mathrm{NLL}(\mathbf{W},\mathbf{Z}) &= -\sum_{n=1}^{N}\log p(\mathbf{x}_n \mid \boldsymbol{\theta}_n) \\ &= -\sum_{n=1}^{N}(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\theta}_n - \Omega^*(\boldsymbol{\theta}_n)) \\ &= \sum_{n=1}^{N}\mathrm{Bregman}_{\Omega}(\mathbf{x}_n, \boldsymbol{\mu}_n(\boldsymbol{\theta}_n))\end{aligned} \tag{18}$$

where $\boldsymbol{\mu}_n = \boldsymbol{\mu}_n(\boldsymbol{\theta}_n) = \mathbb{E}(\mathbf{x}_n \mid \mathbf{W},\mathbf{Z}) = \nabla\Omega^*(\boldsymbol{\theta}_n)$, $\Omega^*(\boldsymbol{\theta}_n)$ is the cumulant or log partition, and $\Omega(\boldsymbol{\mu}_n)$ is the conjugate dual of Ω^*, and in this case equal to Shannon negentropy.[12]

Assuming that the elements of the observation column vector are independent, we can also write $\mathrm{NLL}(\mathbf{W},\mathbf{Z}) = -\sum_{n=1}^{N}\sum_{d=1}^{D}\log p(x_{dn} \mid \theta_{dn})$. In fact, Collins et al. [2001] assume that column vector elements are independent and $x_{nd} \mid \mathbf{W},\mathbf{Z} \sim \mathrm{Expon}(\theta_{nd})$ where $\theta_{dn} = w_{d1}z_{1n} + ... + w_{dK}z_{Kn}$.

3.2 Exponential family & Bregman divergence

The Bregman divergence corresponding to a strictly convex function $\Omega : S \to \mathbb{R}$ with $S = \mathrm{dom}(\Omega) \subseteq \mathbb{R}^D$ a convex set, and Ω differentiable on $\mathrm{ri}(S)$ is defined as

$$\mathrm{B}_{\Omega}(\mathbf{x},\boldsymbol{y}) = \Omega(\mathbf{x}) - (\Omega(\boldsymbol{y}) + \langle \mathbf{x}-\boldsymbol{y}, \nabla\Omega(\boldsymbol{y})\rangle) \tag{19}$$

where $\Omega(\boldsymbol{y}) + \langle \mathbf{x}-\boldsymbol{y}, \nabla\Omega(\boldsymbol{y})\rangle$ is the tangent of Ω at $\mathbf{y}$, or the Taylor expansion of Ω at $\mathbf{y}$.

[12] See Mohamed [2011] § 2.16.

For example, the KL divergence can be expressed as a Bregman divergence using as a strictly convex function the negative Shannon entropy $\Omega(\mathbf{p}) = \sum_{d=1}^{D} p_d \log p_d$ of a discrete probability distribution where $\sum_{d=1}^{D} p_d = 1$. The corresponding Bregman divergence is

$$\begin{aligned}
\mathrm{B}_\Omega(\mathbf{p}, \mathbf{q}) &= \Omega(\mathbf{p}) - \Omega(\mathbf{q}) - \langle \mathbf{p} - \mathbf{q}, \nabla\Omega(\mathbf{q}) \rangle \\
&= \sum_{d=1}^{D} p_d \log p_d - \sum_{d=1}^{D} q_d \log q_d - \langle \mathbf{p} - \mathbf{q}, \nabla\Omega(\mathbf{q}) \rangle \\
&= \sum_{d=1}^{D} p_d \log p_d - \sum_{d=1}^{D} q_d \log q_d - \sum_{d=1}^{D} (p_d - q_d)(\log q_d + 1) \\
&= \sum_{d=1}^{D} p_d \log p_d - \sum_{d=1}^{D} p_d \log q_d - \sum_{d=1}^{D} p_d + \sum_{d=1}^{D} q_d \\
&= \sum_{d=1}^{D} p_d \log \frac{p_d}{q_d} = \mathrm{KL}(\mathbf{p}, \mathbf{q}).
\end{aligned} \tag{20}$$

Consider now a regular exponential family

$$p_{\Omega^*, \boldsymbol{\theta}}(\mathbf{x}) = \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \Omega^*(\boldsymbol{\theta})) p_0(\mathbf{x}) \tag{21}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}_{p_{\Omega^*, \boldsymbol{\theta}}} \mathbf{x}$, the minimal sufficient statistic is $\mathbf{x} \in \mathbb{R}^D$, and $\Omega^*(\boldsymbol{\theta})$ is the cumulant or log partition. The conjugate dual of Ω^* is

$$\Omega(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}} \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \Omega^*(\boldsymbol{\theta}). \tag{22}$$

We can obtain the unique $\boldsymbol{\theta}^*$ that corresponds to the sup above in (22):

$$\nabla \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \Omega^*(\boldsymbol{\theta}) \,|_{\boldsymbol{\theta}^*} = 0 \implies \boldsymbol{\mu} = \nabla\Omega^*(\boldsymbol{\theta}^*). \tag{23}$$

Remember also that $\boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta})$ and $\boldsymbol{\theta}(\boldsymbol{\mu}) = \nabla\Omega(\boldsymbol{\mu})$.[13]

By the definition of the Bregman divergence $\mathrm{B}_\Omega(\mathbf{x}, \boldsymbol{\mu}) = \Omega(\mathbf{x}) - \Omega(\boldsymbol{\mu}) - \langle \mathbf{x} - \boldsymbol{\mu}, \nabla\Omega(\boldsymbol{\mu}) \rangle \implies \Omega(\boldsymbol{\mu}) + \langle \mathbf{x} - \boldsymbol{\mu}, \nabla\Omega(\boldsymbol{\mu}) \rangle = \Omega(\mathbf{x}) - \mathrm{B}_\Omega(\mathbf{x}, \boldsymbol{\mu})$. Thus we can write the arguments inside the exponent of the regular Exponential family in terms of B_Ω:

$$\begin{aligned}
\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \Omega^*(\boldsymbol{\theta}) &= \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \Omega^*(\boldsymbol{\theta}) + \langle \mathbf{x} - \boldsymbol{\mu}, \boldsymbol{\theta} \rangle \\
&= \Omega(\boldsymbol{\mu}) + \langle \mathbf{x} - \boldsymbol{\mu}, \nabla\Omega(\boldsymbol{\mu}) \rangle \\
&= -\mathrm{B}_\Omega(\mathbf{x}, \boldsymbol{\mu}) + \Omega(\mathbf{x}),
\end{aligned} \tag{24}$$

and the log likelihood can be expressed in terms of the Bregman divergence as

$$\log p_{\psi, \boldsymbol{\theta}}(\mathbf{x}) = -\mathrm{B}_\Omega(\mathbf{x}, \boldsymbol{\mu}) + \Omega(\mathbf{x}) + \log(p_0(\mathbf{x})) \tag{25}$$

where the last two terms do not depend on the parameters.

[13] See Wainwright et al. [2008], ch.3.

3.3 Classical Multinomial PCA

In this section we review Banerjee et al. [2005]'s (pp. 1726-1727) derivation of an example of a Multinomial likelihood expressed in terms of the Bregman divergence with the Shannon entropy.

Suppose a sample $\mathbf{x}_i = \begin{pmatrix} x_{1i} & ... & x_{Di} \end{pmatrix}^{\mathrm{T}}$ follows a Multinomial distribution with index $M_i = \sum_{d=1}^{D} x_{di}$ and vector vector of probabilities $\mathbf{p}_i = \begin{pmatrix} p_{1i} & ... & p_{Di} \end{pmatrix}^{\mathrm{T}}$ where $\sum_{d=1}^{D} p_{di1} = 1$. Let the corresponding natural parameter be $\boldsymbol{\theta}_{i_{D\times 1}} = \mathbf{W}_{D\times K}\mathbf{z}_{i_{K\times 1}}$. First, we will treat $\mathbf{W}$ and $\mathbf{z}_i$ as fixed unknwon parameters to be estimated for $i = 1, ..., N$.

Following Collins et al. [2001] we consider the negative log likelihood of the Mutlinomial in Exponential family form:

$$\begin{aligned}
p(\mathbf{x}_i \mid \mathbf{z}_i; \boldsymbol{\Theta}) &= \frac{M_i!}{x_{1i}!...x_{Di}!} p_{1i}^{x_{1i}} ... p_{D-1i}^{x_{D-1i}} p_{Di}^{M_i - \sum_{d=1}^{D-1} x_{di}} \\
&= \frac{M_i!}{x_{1i}!...x_{Di}!} \exp\left(x_{1i} \log p_{1i} + ... + x_{D-1i} \log p_{D-1i} + \left(M_i - \sum_{d=1}^{D-1} x_{di} \right) \log p_{Di} \right) \\
&= \frac{M_i!}{x_{1i}!...x_{Di}!} \exp\left(x_{1i} \log \frac{p_{1i}}{p_{Di}} + ... + x_{D-1i} \log \frac{p_{D-1i}}{p_{Di}} - M_i \log \frac{1}{p_{Di}} \right) \\
&= \frac{M_i!}{x_{1i}!...x_{Di}!} \exp\left(x_{1i}\theta_{1i} + ... + x_{D-1i}\theta_{D-1i} - M_i \log\left(1 + \sum_{d=1}^{D-1} e^{\theta_{di}} \right) \right)^{14} \\
&= \frac{M_i!}{x_{1i}!...x_{Di}!} \exp\left(\sum_{d=1}^{D} x_{di}\theta_{di} - M_i \log\left(\sum_{d=1}^{D} e^{\theta_{di}} \right) \right).^{15}
\end{aligned} \tag{26}$$

So,

$$\begin{aligned}
\mathrm{NLL} &= -\sum_{i=1}^{N} \log p(\mathbf{x}_i \mid \mathbf{W}, \mathbf{z}_i) \\
&= -(\log p_0(\mathbf{x}_i) + \langle \mathbf{x}_i, \boldsymbol{\theta}_i \rangle - \Omega^*(\boldsymbol{\theta}_i))
\end{aligned} \tag{27}$$

where $\{x_{di}\}_{d=1}^{D-1}$ are the sufficient statistics, $\boldsymbol{\theta}_i = \{\log \frac{p_{di}}{p_{Di}}\}_{d=1}^{D-1}$ are the natural parameters, $\Omega^*(\boldsymbol{\theta}_i) = M_i \log \frac{1}{p_{Di}} = M_i \log\left(1 + \sum_{d=1}^{D-1} e^{\theta_{di}} \right)$ is the cumulant, and the expectation parameter is $\boldsymbol{\mu}_i = \nabla\Omega^*(\boldsymbol{\theta}_i) =$

[15] Add by parts $p_1 = e^{\theta_1} p_D, ..., p_{D-1} = e^{\theta_{D-1}} p_D$ to get $\sum_{d=1}^{D-1} p_d = 1 - p_D = p_D \sum_{d=1}^{D-1} e^{\theta_d}$ and solve for $p_D = \frac{1}{1 + \sum_{d=1}^{D-1} e^{\theta_d}}$.

[15] By convention $\theta_D = 0$.

$M_i \left[\frac{e^{\theta_{di}}}{1+\sum_{d=1}^{D-1} e^{\theta_{di}}} \right]_{d=1}^{D-1} = [M_i p_{di}]_{d=1}^{D-1}$. The Legendre dual of Ω^* is

$$\begin{aligned}
\Omega(\boldsymbol{\mu}_i) &= \langle \boldsymbol{\mu}_i, \boldsymbol{\theta}_i \rangle - \Omega^*(\boldsymbol{\theta}_i) \\
&= M_i \sum_{d=1}^{D-1} p_{di} \log \frac{p_{di}}{p_{Di}} + M_i \log p_{Di} \\
&= M_i \sum_{d=1}^{D-1} p_{di} \log p_{di} - M_i \log p_{Di} \sum_{d=1}^{D-1} p_{di} + M_i \log p_{Di} \\
&= M_i \sum_{d=1}^{D} p_{di} \log p_{di} = M_i \sum_{d=1}^{D} \frac{\mu_{di}}{M_i} \log \frac{\mu_{di}}{M_i}.
\end{aligned} \tag{28}$$

The Bregman divergence we want is

$$\begin{aligned}
\mathrm{B}_\Omega(\mathbf{x}_i, \boldsymbol{\mu}_i) &= \Omega(\mathbf{x}_i) - \Omega(\boldsymbol{\mu}_i) - \langle \mathbf{x}_i - \boldsymbol{\mu}_i, \nabla\Omega(\boldsymbol{\mu}_i) \rangle \\
&= M_i \sum_{d=1}^{D} \frac{x_{id}}{M_i} \log \frac{x_{di}/M_i}{\mu_{di}/M_i},
\end{aligned} \tag{29}$$

and for any Exponential family we showed above that

$$\log p(\mathbf{x}_i \mid \mathbf{W}, \mathbf{z}_i) = -\mathrm{B}_\Omega(\mathbf{x}_i, \boldsymbol{\mu}_i) + \log b(\mathbf{x}_i) \tag{30}$$

where $\log b(\mathbf{x}_i) = \Omega(\mathbf{x}_i) + \log p_0(\mathbf{x}_i) = M_i \sum_{d=1}^{D} \frac{x_{di}}{M_i} \log \frac{x_{di}}{M_i} + \log \frac{M_i!}{x_{1i}! \ldots x_{Di}!}$ does not involve parameters. So, the loss becomes

$$\mathrm{NLL} = -\sum_{i=1}^{N} \log p(\mathbf{x}_i \mid \mathbf{W}, \mathbf{z}_i) = \sum_{i=1}^{N} \mathrm{B}_\Omega(\mathbf{x}_i, \boldsymbol{\mu}_i) = \sum_{i=1}^{N} \sum_{d=1}^{D} x_{di} \log \frac{x_{di}}{\mu_{di}} \tag{31}$$

where $\mu_{di} = M_i \frac{e^{\theta_{di}}}{\sum_{d=1}^{D} e^{\theta_{di}}}$ for $d = 1, \ldots, D-1$, $\mu_{Di} = M_i \frac{1}{\sum_{d=1}^{D} e^{\theta_{di}}}$, $\theta_{di} = w_{d0} + w_{d1} z_{1i} + \ldots + w_{dK} z_{Ki}$, for simplicity letting the intercept terms $w_{d0} = 0$, and setting $w_{D1} = \ldots = w_{DK} = 0$ to ensure that $\theta_{Di} = 0, \forall i =, \ldots, N$.

3.4 Probabilistic Multinomial PCA & Variational EM

3.4.1 Model

Let $\mathbf{x}_i \sim \mathrm{Multinomial}\left(M_i, \mathbf{p}_i = \begin{pmatrix} p_{1i} & \ldots & p_{Di} \end{pmatrix}^{\mathrm{T}}\right)$ where $\sum_{d=1}^{D} p_{di1} = 1$. Supposing that the natural parameter $\boldsymbol{\theta}_i$ follows a Multivariate Normal prior implies that the mean parameter $\mathbf{p}_i$ follows a

Logistic Normal (Atchison and Shen [1980]). Putting everything together:

$$\begin{aligned} p(\mathbf{x}_i|\boldsymbol{\theta}_i) &\sim \text{Multinomial} \\ \pi(\boldsymbol{\theta}_i) &\sim \mathcal{N} \equiv \pi(\mathbf{p}_i) \sim \text{Logistic Normal} \\ p(\boldsymbol{\theta}_i \mid \mathbf{x}_i) &= \frac{p(\mathbf{x}_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)}{\int p(\mathbf{x}_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i} \\ p(\mathbf{x}_i) &= \int p(\mathbf{x}_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \sim \text{ Logistic Normal Multinomial.} \end{aligned} \tag{32}$$

Since the marginal $p(\mathbf{x}_i)$ depends on the hyperparameters of the prior $\pi(\boldsymbol{\theta}_i)$, it can be used in Empirical Bayes approaches to estimate those hyperparameters.

Let $\mathbf{z}_i$ be a latent variable of dimension $K < D$, $\mathbf{z}_i = \begin{pmatrix} z_{1i} & \dots & z_{Ki} \end{pmatrix}^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We want to factorize $\theta_{di} = w_{d0} + \mathbf{w}_{d_{1\times K}}\mathbf{z}_{i_{K\times 1}}$ where $\mathbf{w}_d = \begin{pmatrix} w_{d1} \cdots w_{dK} \end{pmatrix}$ for $d = 1, \cdots, D$. The complete data log likelihood is

$$\sum_{i=1}^{N} \log p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\Theta}) = \sum_{i=1}^{N} \log p(\mathbf{z}_i) + \log p(\mathbf{x}_i \mid \mathbf{z}_i; \boldsymbol{\Theta}). \tag{33}$$

The Normal prior $\mathbf{z}_i \sim \mathcal{N}_K(\mathbf{0}, \mathbf{I})$ has density

$$p(\mathbf{z}_i) = (2\pi)^{-K/2} \exp -\frac{1}{2}\mathbf{z}_i^{\mathrm{T}}\mathbf{z}_i, \tag{34}$$

i.e. $z_{ki} \sim \mathcal{N}(0, 1)$ for $k = 1, \cdots, K$, and $p(\mathbf{z}_i) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_{ki}^2}$. The Multinomial $p(\mathbf{x}_i \mid \mathbf{z}_i; \boldsymbol{\Theta})$ has the density in (26) above.

So, we can write the complete log likelihood explicitly as

$$\sum_{i=1}^{N} \log p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\Theta}) = -\frac{1}{2}\sum_{i=1}^{N} \mathbf{z}_i^{\mathrm{T}}\mathbf{z}_i + \sum_{i=1}^{N}\sum_{d=1}^{D} x_{di}\theta_{di} - \sum_{i=1}^{N} M_i \log \sum_{d=1}^{D} e^{\theta_{di}}, \tag{35}$$

and plug in $\theta_{di} = w_{d0} + \mathbf{w}_{d_{1\times K}}\mathbf{z}_{i_{K\times 1}}$ for $d = 1 \dots D$, and $w_{D1} = \dots = w_{DK} = 0$.

3.4.2 ELBO

To maximize the joint likelihood $p(\mathbf{x}_i, \mathbf{z}_i)$ with classical EM we would need to compute an expectation w.r.t. $p(\mathbf{z}_i \mid \mathbf{x}_i)$, which is intractable. Alternatively, we can use variational EM (as in Li and Tao [2010], Chiquet et al. [2018], (Zeng et al. [2022][16]) and maximize a lower bound

[16] Li and Tao [2010] first propose the variational EM for Exponential family probabilistic PCA. Logistic Normal Multinomial PCA (LNM PCA) is first proposed by Xia et al. [2013] who use Monte Carlo EM. LNMA PCA is adapted by Zeng et al. [2022] who use variational EM and 0-inflation to circumvent sparse observations. Also worth mentioning is Fang and Subedi [2023], who use variational EM to fit a Logistic Normal Multinomial model without the low rank factorization of PCA, as well as Morton et al. [2021] who describe a neural parametrization of LNM PCA and .

on the log marginal likelihood using a Normal variational distribution $q(\mathbf{z}_i) \sim \mathcal{N}(\mathbf{m}_{i_{K\times 1}}, \mathbf{\Sigma}_{i_{K\times K}})$ where $\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_K^2 \end{bmatrix}_i$:

$$\begin{aligned} q(\mathbf{z}_i) &= (2\pi)^{k/2} \mid \mathbf{\Sigma}_i \mid^{-1/2} e^{-\frac{1}{2}(\mathbf{z}_i - \mathbf{m}_i)^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} (\mathbf{z}_i - \mathbf{m}_i)} \\ q(z_{ki}) &= \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} e^{-\frac{(z_{ki} - m_{ki})^2}{\sigma_{ki}^2}}. \end{aligned} \tag{36}$$

The lower bound we want to derive is

$$\mathrm{ELBO}_i = \mathbb{E}_{q(\mathbf{z}_i)} \log p(\mathbf{x}_i, \mathbf{z}_i) - \mathbb{E}_{q(\mathbf{z}_i)} \log q(\mathbf{z}_i) \tag{37}$$

The second term is the Shannon entropy of a multivariate Normal, so $\mathbb{E}_{q(\mathbf{z}_i)} \log q(\mathbf{z}_i) = \frac{K}{2} + \frac{K}{2}\log(2\pi) + \frac{1}{2}\log \mid \mathbf{\Sigma}_i \mid$. We derive the first term using the densities from (26) and (34)

$$\mathbb{E}_{q(\mathbf{z}_i)} \log(\mathbf{x}_i, \mathbf{z}_i) = \text{const.} - \frac{1}{2}\mathbb{E}_{q(\mathbf{z}_i)} \mathbf{z}_i^{\mathrm{T}} \mathbf{z}_i + \mathbb{E}_{q(\mathbf{z}_i)} \sum_{d=1}^{D} x_{id}\theta_{id} - M_i \log\left(\sum_{d=1}^{D} e^{\theta_{di}}\right) \tag{38}$$

From the first non-constant term of (38) we have

$$\mathbb{E}_{q(\mathbf{z}_i)} \mathbf{z}_i^{\mathrm{T}} \mathbf{z}_i = \mathbb{E}_{q(\mathbf{z}_i)} \sum_{k=1}^{K} z_{ki}^2 = \sum_{k=1}^{K} \sigma_{ki}^2 + m_{ki}^2 = tr(\mathbf{m}_i \mathbf{m}_i^{\mathrm{T}} + \mathbf{\Sigma}_i), \tag{39}$$

and from the second term

$$\mathbb{E}_{q(\mathbf{z}_i)} \sum_{d=1}^{D} x_{di}\theta_{id} = \mathbb{E}_{q(\mathbf{z}_i)} \sum_{d=1}^{D} x_{di}(w_{d0} + \mathbf{w}_d \mathbf{z}_i) = \sum_{d=1}^{D} x_{di}(w_{d0} + \mathbf{w}_d \mathbf{m}_i). \tag{40}$$

For the third term

$$\mathbb{E}_{q(\mathbf{z}_i)} \log\left(\sum_{d=1}^{D} e^{w_{0d} + \mathbf{w}_d \mathbf{z}_i}\right) \leq \log \mathbb{E}_{q(\mathbf{z}_i)} \sum_{d=1}^{D} e^{w_{d0} + w_{d1} z_{1i} + \dots + w_{dK} z_{Ki}} \text{17} \tag{41}$$

we need to compute terms like $\mathbb{E}_{q(\mathbf{z}_i)} e^{w_{dk} z_{ki}}$ where each $z_{ki} \sim \mathcal{N}(m_{ki}, \sigma_{ki}^2)$ since $\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_K^2 \end{bmatrix}_i$. So, $w_{dk} z_{ki} \sim \mathcal{N}(w_{dk} m_{ki}, w_{dk}^2 \sigma_{ki}^2)$, and the expectation $\mathbb{E} e^{w_{dk} z_{ki}}$ is the mean of a Log-Normal distribution $\mathbb{E} e^{w_{dk} z_{ki}} = e^{w_{dk} m_{ki} + \frac{w_{dk}^2 \sigma_{ki}^2}{2}}$.[18] Thus

$$\begin{aligned} \log \mathbb{E}_{q(\mathbf{z}_i)} \sum_{d=1}^{D} e^{w_{d0} + w_{d1} z_{1i} + \dots + w_{dK} z_{Ki}} &= \log \sum_{d=1}^{D} e^{w_{d0} + w_{d1} m_{1i} + \frac{w_{d1}^2 \sigma_{1i}^2}{2} + \dots + w_{dK} m_{Ki} + \frac{w_{dK}^2 \sigma_{Ki}^2}{2}} \\ &= \log \sum_{d=1}^{D} e^{w_{d0} + \mathbf{w}_d \mathbf{m}_i + \frac{1}{2}(w_{d1} \cdots w_{dK}) \mathbf{\Sigma}_i (w_{d1} \cdots w_{dK})^{\mathrm{T}}}. \end{aligned} \tag{42}$$

[17] By Jensen's inequality for any convex function g, such as $-\log$, $\mathbb{E} g(X) \geq g(\mathbb{E} X)$, so $\mathbb{E}(-\log X) \geq -\log \mathbb{E} X$, i.e. $\mathbb{E}(\log X) \leq \log \mathbb{E} X$.

[18] When $X \sim \mathcal{N}(\mu, \sigma^2)$ then $Y = e^X$ (i.e. $X = \log Y$) follows a Log-Normal and $\mathbb{E} Y = \mathbb{E} e^X = e^{\mu + \sigma^2/2}$.

Finally, we can put all terms back together:

$$\begin{aligned}
\text{ELBO}_i &= \mathbb{E}_q \log p(\mathbf{z}_i) + \mathbb{E}_q \log p(\mathbf{x}_i \mid \mathbf{z}_i) - \mathbb{E}_q \log q(\mathbf{z}_i) \\
&= \text{const.} - \frac{1}{2}\mathbb{E}_q \mathbf{z}_i^{\mathrm{T}} \mathbf{z}_i + \mathbb{E}_q \sum_{d=1}^{D} x_{di}\theta_{di} - M_i \mathbb{E}_q \log\left(\sum_{d=1}^{D} e^{\theta_{di}}\right) - \frac{1}{2}\log \mid \mathbf{\Sigma}_i \mid \\
&\geq \text{const.} - \frac{1}{2} tr(\mathbf{m}_i^{\mathrm{T}}\mathbf{m}_i + \mathbf{\Sigma}_i) - \frac{1}{2}\log \mid \mathbf{\Sigma}_i \mid + \sum_{d=1}^{D} x_{id}(w_{d0} + \mathbf{w}_d \mathbf{m}_i) - M_i \log \sum_{d=1}^{D} e^{w_{d0} + \mathbf{w}_d \mathbf{m}_i + \frac{1}{2}\mathbf{w}_d \mathbf{\Sigma}_i \mathbf{w}_d^{\mathrm{T}}}.
\end{aligned} \tag{43}$$

4 Sparse PCA via Fenchel-Young losses

Often $\mathbf{X}$ is made up of very sparse count data (e.g. gene expression, microbiome data, skipgram-style word embeddings, word phoneme feature combinations [CITE]). To model such cases we can derive a model analogous to Collins et al. [2001]'s frequentist treatment of the latent variables as fixed unknown quantities to be estimated, subject to a low rank factorization of the natural parameter matrix. We optimize the conditional density through a Fenchel-Young loss $L_\Omega(\boldsymbol{\theta}, \mathbf{x})$, which is a generalization of the Bregman divergence (Blondel et al. [2020] § 3.2), subject to a low rank factorization of the corresponding parameter matrix. This conditional density is a sparse deformed Exponential family density. Taking Ω to be the Tsallis entropy and $a = 2$ yields a Sparsemax loss (Martins and Astudillo [2016]).

4.1 Simple Fenchel-Young PCA: Toy experiments with Sparsemax

In Figure 1 we compare classical Exponential family PCA with a Bregman divergence objective, fitted to 100 samples, each of which is made up of D independent Bernoullis, and classical FY PCA fitted to data generated in the same way, by minimizing the sum of D independent binary sparsemax projections for each sample.

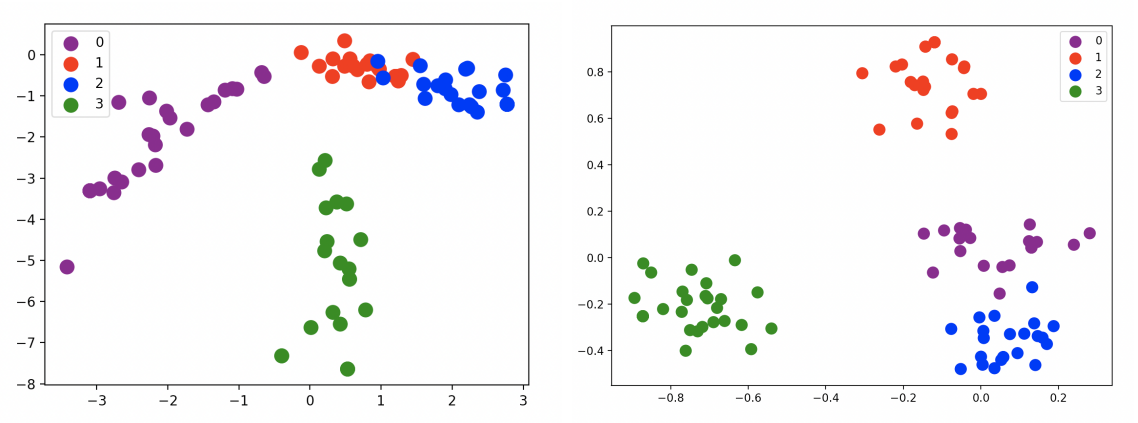

Figure 1: Left: 40 epochs Bernoulli Bregman (Colab). Right: 40 epochs 2-class Sparsemax loss (Colab). There is no generative model to match with the synthetic data generation. 100 samples from 4 templates (colors) of $D = 30$ independent Bernoullis with $p = 0.5$ plus some Bernoulli noise $p = 0.05$ are clustered with 3-dimensional $\mathbf{z}_i$'s, and correspondingly 30×3-dim $\mathbf{W}$.

In Table 1 we compare Logistic Normal Multinomial (LNM) probabilistic PCA (fitted with the ELBO objective in (34)) with classical Multinomial PCA (fitted with Bregman divergence objective), and with classical Sparsemax PCA on low dimensional data reconstruction using as metric $\frac{\sqrt{\sum(p-\hat{p})^2}}{\sqrt{\sum p^2}}$:

$p(x_{ij} = 0)$	LNM PPCA	Multinomial (Bregman)	Sparsemax (FY loss)
0.8	0.1606	0.0930	**0.0665**
0.5	0.1360	0.0859	**0.0616**
0	0.1692	**0.0672**	0.1197

Table 1: All models are highly sensitive to random initialization. We report best values over 3 runs. In all cases there are 100 synthetic samples with 50 observed features and 2 latent dimensions. The PCA factorizations assume 2 latent factors.

The Bregman and FY loss objectives do not assume a probabilistic generative process as both scores and loadings are treated as fixed, unknown quantities. The LNM PPCA model assumes a generative story that is mathched by the experiments in the last row of Table 1, where we do not "induce" sparsity. To generate the synthetic data for the experiments in Table 1 "true" loadings are drawn from a continuous $\text{Uniform}(-3.5, 3.5)$ and the "true" scores are drawn from 4 standard 2-dimensional Gaussians. The exponentiated product of scores and loadings is optionally multiplied with independent Bernoullis to induce sparsity corresponding to the Bernoulli parameter. This step is the only deviation from the generative process of the LNM model. Normalizing the (sparse) product of scores and loadings leads to "true" Multinomial probabilities that we use to sample an observed synthetic data matrix (each row has total counts drawn from a $\text{Uniform}(200, 400)$).

We see the expected effect in Table 1 that Sparsemax PCA outperforms the Mutlinomial PCA and LNM PPCA when the observed data is sparse, since of these models only Sparsemax PCA can produce reconstructions of the 'true" Multinomial probabilities containing 0s, in the other two models the best approximations are instead really small positive values, since the product of estimated scores and loadings is passed through a softmax to produce the reconstructed probabilities. Nevertheless, this task is too easy for all of the models, as they all separate the 4 true underlying clusters corresponding to the generative process.

4.2 Fenchel-Young ELBO

4.2.1 Preliminary definitions

Shannon entropy:

$$-\int p(t)\log p(t)dt = -\mathbb{E}_{p(t)}\log p(t) = \mathbb{E}_{p(t)}\log\frac{1}{p(t)}. \tag{44}$$

Shannon negetropy:

$$\Omega_S(p) = \mathbb{E}_{p(t)}\log p(t) = -\mathbb{E}_{p(t)}\log\frac{1}{p(t)}. \tag{45}$$

We do not want to use the notation q-log to reserve q for something else, so we call the deformed logarithm a-log:

$$\log_a(p) = \begin{cases} \frac{p^{1-a}-1}{1-a} & a \neq 1 \\ \log p & a = 1.\end{cases}^{19} \tag{46}$$

The inverse function is

$$\exp_a(p) = \begin{cases} (1+(1-a)p)^{\frac{1}{1-a}} & a \neq 1 \\ e^p & a = 1. \end{cases} \tag{48}$$

One important property of $\log$ that is not preserved by $\log_a$ is the identity $-\log p = \log\frac{1}{p}$, instead $-\log_a\frac{1}{p} = \log_{2-q} p$.[20][21] Generalize the Shannon negetropy to Tsallis a-negetropy:

$$\Omega_a(p) = -\mathbb{E}_{p(t)}\log_a\frac{1}{p(t)} = \mathbb{E}_{p(t)}\log_{2-a} p(t)^{22} \tag{49}$$

[19] In Martins et al. [2022] q-log is called β-log. For $b = 2a-1$, $a = \frac{b+1}{2}$, $1-a = \frac{1-b}{2}$ we get

$$\log_a(p) = \begin{cases} \frac{2}{1-b}(p^{\frac{1-b}{2}}-1) & b \neq 1 \\ \log p & b = 1 \end{cases} \tag{47}$$

which is closely related to the a-representation of p, $h_a(p) = p^{\frac{1-a}{2}}$ (Amari [2016], 4.35, 4.18).

[20] $\log_a\frac{1}{p} = \frac{1}{1-a}\left(\frac{1}{p^{1-a}}-1\right) = \frac{1}{1-a}(p^{1-2+a}-1) = \frac{1}{1-a}(p^{1-(2-a)}-1) = -\log_{2-a} p$

[21] Sears et al. [2008], p. 60-63.

4.2.2 Probabilistic Fenchel-Young PCA model

Let $p(\mathbf{x} \mid \mathbf{z})$ be the Ω_a-regularized prediction map

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{z}) &= \hat{p}_{\Omega_a} f_{\boldsymbol{\theta}(\mathbf{z})}(\mathbf{x}) \\ &= \underset{p \in M_+^1(S)}{\operatorname{argmax}} \mathbb{E}_p f_{\boldsymbol{\theta}(\mathbf{z})}(\mathbf{x}) - \Omega_a(\mathbf{p}) \end{aligned} \tag{51}$$

and assume a linear parametrization of the scoring function $f_{\boldsymbol{\theta}(\mathbf{z})}(\mathbf{x}) = \boldsymbol{\theta}(\mathbf{z}) t(\mathbf{x})$ where $t(\mathbf{x})$ is the sufficient statistic and the canonical parameter $\boldsymbol{\theta}(\mathbf{z}) = \mathbf{w}\mathbf{z}$ is comprised of a global parameter $\mathbf{w}$ and a latent variable $\mathbf{z}$ that corresponds to $\mathbf{x}$. The set $M_+^1(S)$ consists of all densities $p : S \to \mathbb{R}_+$ such that $\int_S p(\mathbf{x}) d\mathbf{x} = 1$.

For the Shannon negetropy Ω_s the solution of the optimization problem is an exponential family with canonical parameter $\boldsymbol{\theta}(\mathbf{z})$, sufficient statistic $t(\mathbf{x})$ and cumulant function $\Omega_s^*(\boldsymbol{\theta}(\mathbf{z}))$, i.e. the ith observation $\mathbf{x}_i \mid \mathbf{z}_i \sim \operatorname{Expon}(\mathbf{w}\mathbf{z}_i, t(\mathbf{x}_i))$. The parameter of the observation model (or **decoder**) is $\mathbf{w}$ and

$$\begin{aligned} p_{\mathbf{w}}(\mathbf{x} \mid \mathbf{z}) &= \hat{p}_{\Omega_s} f_{\theta(\mathbf{z})}(\mathbf{x}) \\ &= \exp(\mathbf{w}\mathbf{z} t(\mathbf{x}) - \Omega^*(\mathbf{w}, \mathbf{z})) \end{aligned} \tag{52}$$

If, additionally, the set of densities $M_+^1(S)$ is the simplex $\Delta^{|S|}$ then this exponential family is the Categorical. An analytic form of $p(\mathbf{x} \mid \mathbf{z})$ in the general case of $M_+^1(S)$ and $a \neq 1$ is given by eqn. 10 of Martins et al. [2022]:

$$\begin{aligned} p_{\mathbf{w}}(\mathbf{x} \mid \mathbf{z}) &\equiv \hat{p}_{\Omega_a} f_{\boldsymbol{\theta}(\mathbf{z})}(\mathbf{x}) \\ &= \exp_{2-a}(\boldsymbol{\theta}(\mathbf{z}) t(\mathbf{x}) - \Omega_a^*(\boldsymbol{\theta}(\mathbf{z}))) \end{aligned} \tag{53}$$

For the simplex and $a = 2$, we get the sparsemax density in eqn. 11 of Martins et al. [2022].

The variational $q(\mathbf{z}|\mathbf{x})$ distribution (or **encoder**) can also be a Ψ_a-regularized map. Let $f_{\boldsymbol{\eta}}(\mathbf{z})$ be its scoring function. Assuming it has a linear form $f_{\boldsymbol{\eta}}(\mathbf{z}) = \boldsymbol{\eta} t(\mathbf{z})$, we can write it also in the form of eqn. 10 of Martins et al. [2022]. Assume $a = 1$ and $\mathbf{z}$ is continuous, so $q(\mathbf{z} \mid \mathbf{x})$ is a continuous exponential family or simply a Normal[23]

$$q_{\boldsymbol{\eta}}(\mathbf{z} \mid \mathbf{y}) = \exp(\boldsymbol{\eta} t(\mathbf{z}) - \Psi^*(\boldsymbol{\eta})) \tag{55}$$

[22] Double-check:

$$\log_{2-a} p(t) = \begin{cases} \frac{p(t)^{1-(2-a)}-1}{1-(2-a)} = \frac{p(t)^{-1+a}-1}{-1+a} = \frac{-1}{1-a}\left(\frac{1}{p(t)^{1-a}} - 1\right) & 2-a \neq 1 \implies a \neq 1 \\ \log p(t) & a = 1. \end{cases} \tag{50}$$

[23] The univariate Gaussian in exponential form:

$$p(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(z-\mu)^2) = \frac{1}{\sqrt{2\pi}} \exp(\frac{\mu}{\sigma^2} z - \frac{1}{2\sigma^2} z - \frac{1}{2\sigma^2}\mu^2 - \log\sigma) \tag{54}$$

In § 3.4.2. above, we have non-amortized variational parameters (m_{ik}, σ^2_{ik}), so the number of parameters scales with additional observations.[24] Classically in VAEs, we employ an amortized mean-field variational distribution: $q_{\boldsymbol{\eta}}(\mathbf{z} \mid \mathbf{x}) = \prod_{i=1}^N q_\eta(\mathbf{z}_i \mid \mathbf{x}_i)$ where

$$q_\eta(\mathbf{z}_i \mid \mathbf{x}_i) = \mathcal{N}(\mathbf{z}_i \mid \mu(\mathbf{x}_i), \sigma^2(\mathbf{x}_i)\mathbf{I}_{K\times K}) \tag{56}$$

and $\mu(\mathbf{x}_i), \sigma^2(\mathbf{x}_i)$ are functions of the data, i.e. Neural Nets. So, $\boldsymbol{\eta} = \eta(\phi, \mathbf{x})$, where ϕ is a parameter that summarizes the corresponding NN parameters, i.e. a global, variational parameter. We can write $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathrm{Expon}(\eta(\phi, \mathbf{x}), t(\mathbf{z}))$, otherwise as PPCA is described above $q_{\boldsymbol{\eta}}(\mathbf{z}) = \mathrm{Expon}(\boldsymbol{\eta}, t(\mathbf{z}))$ the variational parameters $\boldsymbol{\eta}$ are estimated along with the parameter $\mathbf{w}$ of the observation model.

4.2.3 Bregman divergence and FY losses

We defined above the Bregman divergence corresponding to a strictly convex function, for example the negetropy $\Psi : dom(\Psi) \to \mathbb{R}$. The arguments of the Bregman divergence are elements of $dom(\Psi)$. In the exponential family, the elements of $dom(\Psi)$ are mean parameters, so if we consider two elements $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, $\mathrm{B}_\Psi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \Psi(\boldsymbol{\mu}_1) - \Psi(\boldsymbol{\mu}_2) - \langle \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \nabla\Psi(\boldsymbol{\mu}_2)\rangle$.[25] Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ be corresponding natural parameters. The Fenchel-Young loss is

$$\begin{aligned} L_\Psi(\boldsymbol{\theta}_1, \boldsymbol{\mu}_2) &= \Psi(\boldsymbol{\mu}_2) + \Psi^*(\boldsymbol{\theta}_1) - \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\mu}_2 \\ &= \Psi(\boldsymbol{\mu}_2) - \Psi(\boldsymbol{\mu}_1) - \nabla\Psi(\boldsymbol{\mu}_1)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathrm{B}_\Psi(\boldsymbol{\mu}_2, \boldsymbol{\mu}_1), \end{aligned} \tag{57}$$

since $\Psi^*(\boldsymbol{\theta}_1) = \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\mu}_1 - \Psi(\boldsymbol{\mu}_1) = \nabla\Psi(\boldsymbol{\mu}_1)\boldsymbol{\mu}_1 - \Psi(\boldsymbol{\mu}_1)$ and $\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\mu}_2 = \nabla\Psi(\boldsymbol{\mu}_1)\boldsymbol{\mu}_2$.[26]

This is used to express the Bregman divergence between the variational density $q(\mathbf{z} \mid \mathbf{y})$ and the prior $p(\mathbf{z})$ in terms of a FY loss:[27]

$$\mathrm{B}_\Psi(q(\mathbf{z}|\mathbf{x}), p(\mathbf{z})) = L_\Psi(\boldsymbol{\eta}; q(\mathbf{z}|\mathbf{x})) = \Psi(q(\mathbf{z}|\mathbf{x})) + \Psi^*(\boldsymbol{\eta}) - \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\eta(\mathbf{z})]. \tag{58}$$

Note the last term $\mathbb{E}_{q(\mathbf{z}|\mathbf{y})}\eta(\mathbf{z})$ is a more general expression that holds for non-exponential densities. If we say that $p(\mathbf{z}) = \exp(\boldsymbol{\eta}\mathbf{z} - \Psi^*(\boldsymbol{\eta}))$ then $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\boldsymbol{\eta}\mathbf{z} = \boldsymbol{\eta}\mathbb{E}_{q(\mathbf{z}|\mathbf{y})}\mathbf{z}$ (i.e. the product of the natural parameter of $p(\mathbf{z})$ and mean of $q(\mathbf{z} \mid \mathbf{y})$). But if $p(\mathbf{z}) = \exp(\boldsymbol{\eta}t(\mathbf{z}) - \Psi^*(\boldsymbol{\eta}))$, for this more general term, the equality in eqn. (57) may not hold.[28]

---

the natural parameter $\eta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$, suff. stat. $t(z) = \begin{pmatrix} z \\ z^2 \end{pmatrix}$, $\Psi^*(\eta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$, $h(z) = \frac{1}{\sqrt{2\pi}}$, $p(z : \mu, \sigma^2) = h(z)\exp(\eta^{\mathrm{T}}(z) - \Psi^*(\eta))$.

[24] Lucas et al. [2019] and Morton et al. [2021] use NNs to amortize variational parameters in similar settings.

[25] Like Taylor series $\Psi(\boldsymbol{\mu}_1) \approx \Psi(\boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\nabla\Psi(\boldsymbol{\mu}_2) + ...$

[26] Take the definition $\Psi^*(\boldsymbol{\theta}_1) = \sup_{\boldsymbol{\mu}}\langle\boldsymbol{\theta}_1, \boldsymbol{\mu}\rangle - \Psi(\boldsymbol{\mu})$. In exponential families for specific argument $\boldsymbol{\theta}_1$, there is a dual such that $\boldsymbol{\theta}_1 = \nabla\Psi(\boldsymbol{\mu}_1)$, which is the specific value of $\boldsymbol{\mu}$ for which the $\sup_{\boldsymbol{\mu}}\langle\boldsymbol{\theta}_1, \boldsymbol{\mu}\rangle - \Psi(\boldsymbol{\mu})$ is attained, and equals $\Psi^*(\boldsymbol{\theta}_1)$. Hence we can plug in for $\boldsymbol{\theta}_1$ the equal quantity $\langle\boldsymbol{\theta}_1, \boldsymbol{\mu}_1\rangle - \Psi(\boldsymbol{\mu}_1)$ or $\nabla\Psi(\boldsymbol{\mu}_1)\boldsymbol{\mu}_1 - \Psi(\boldsymbol{\mu}_1)$.

[27] Eqn. above (7) in A. Martins' "FY ELBO" note.

[28] $\eta(\mathbf{z})$? The reason for trying to write this Bregman divergence as a FY loss is to get a nice expression for the FY posterior in eqn. (8) of "FY ELBO" and for the FY marginal likelihood of $\mathbf{x}$ or evidence in (9).

Consider eqn. (7) of "FY ELBO"

$$\begin{aligned} F_{\Omega,\Psi}(q;\mathbf{y}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[L_\Omega(\theta(\mathbf{z}),\mathbf{x})] + \Psi(q(\mathbf{z}\mid\mathbf{x})) + \Psi^*(\boldsymbol{\eta}) - \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\eta(\mathbf{z})] \\ &= \mathbb{E}_{q(z|\mathbf{x})}[L_\Omega(\boldsymbol{\theta}(\mathbf{z}),\mathbf{x}) - \eta(\mathbf{z})] + \Psi(q(\mathbf{z}\mid\mathbf{x})) + \Psi^*(\boldsymbol{\eta}). \end{aligned} \tag{59}$$

If the Bregman divergence between $q(\mathbf{z}\mid\mathbf{x})$ and $p(\mathbf{z})$ is a KL divergence we can derive its analytic form, to get derivatives w.r.t. $\mathbf{w}, \phi$, and apply reparametrization trick.[29]

4.2.4 Gradients

We derive the gradients of (59) w.r.t the observation model and the inference network parametrs, **W** and ϕ respectively, for the simplest case when

$$\mathrm{KL}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathrm{B}_\Psi(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2) = L_\Psi(\boldsymbol{\theta}_2,\boldsymbol{\mu}_1) = \Psi(\boldsymbol{\mu}_1) + \Psi^*(\boldsymbol{\theta}_2) - \boldsymbol{\mu}_1\boldsymbol{\theta}_2. \tag{60}$$

Let the prior of the latent variable $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K\times K})$, and let the variational approximate posterior be a multivariate Normal with diagonal covariance:

$$q_\phi(\mathbf{z}\mid\mathbf{x}_i) = \mathcal{N}(\mathbf{m}_i, \mathrm{diag}\{\sigma_{1i}^2, ..., \sigma_{Ki}^2\}) \tag{61}$$

where $m_i(\mathbf{x}_i)$ and $\log\sigma_{ki}(\mathbf{x}_i)$ are non-linear functions of the data point $\mathbf{x}_i$, i.e.

$$(m_{i1}, ..., m_{iK}, \log\sigma_{i1}, ..., \log\sigma_{Ki}) = NN_\phi(\mathbf{x}_i).$$

The parameter ϕ summarizes the corresponding neural network parameters and it is a global parameter (the same for all observations $\mathbf{x}_i$).[30]

Consider the empirical data distribution associated with the ith observation, i.e. the proportion $\frac{1}{M_i}\mathbf{x}_i = \frac{1}{M_i}\begin{pmatrix} x_{1i} & ... & x_{Di}\end{pmatrix}^{\mathrm{T}}$. We wish to obtain scoring parameters $\boldsymbol{\theta}_i \in \mathbb{R}^D$ that factor as $\boldsymbol{\theta}_i = \mathbf{W}_{D\times K}\mathbf{z}_{i_{K\times 1}}$ such that $\mathrm{sparsemax}(\boldsymbol{\theta}_i) = \hat{p}_\Omega[\boldsymbol{\theta}_i] \in \Delta^D$ approximates the observed $\frac{\mathbf{x}_i}{M_i}$. That is, the $\boldsymbol{\theta}_i$ that results from minimizing the FY loss

$$L_\Omega(\boldsymbol{\theta}_i;\mathbf{x}_i/M_i) = L_\Omega(\mathbf{W}\mathbf{z}_i;\mathbf{x}_i/M_i) \quad M_i = x_{1i} + ... + x_{Di} \tag{62}$$

Consider the contribution of the ith observation $\mathbf{x}_i$ to the "variational free energy" of (eqn. 2 of FY-ELBO)

$$F_{\Omega,\Psi}(q;\mathbf{x}_i) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}[L_\Omega(\theta(\mathbf{W}\mathbf{z}_i;\mathbf{x}_i/M_i))] + \mathrm{KL}(q_\phi(\mathbf{z}\mid\mathbf{x}_i)\,\|\,p(\mathbf{z})) \tag{63}$$

The KL divergence between the diagonal Gaussian $q_\phi(\mathbf{z}\mid\mathbf{x}_i)$ and the standard Gaussian has an analytic form

$$\mathrm{KL}(q_\phi(\mathbf{z}\mid\mathbf{x}_i)\,\|\,p(\mathbf{z})) = \frac{1}{2}\sum_{k=1}^{K}(\sigma_{ik}^2 + m_{ik}^2 - 1 - \log\sigma_{ik}^2) \tag{64}$$

[29] Eqn.s (8) and (9) of "FY ELBO" are derived from (7) by FY duality.

[30] Alternatively if we do not model $\mathbf{m}_i$ and $\boldsymbol{\Sigma}_i$ as functions of $\mathbf{x}_i$, we denote all the parameters of $q_{\phi_i}(\mathbf{z})$ by $\boldsymbol{\phi}_i = (m_{i1}, ..., m_{iK}, \sigma_{i1}^2, ..., \sigma_{iK}^2)$.

Given a set of observations $\mathbf{x}_i \in \mathbb{R}^D, i = 1, ..., N$, we estimate $\mathbf{W}_{D\times K}$ and ϕ by minimizing the objective

$$R(\mathbf{W}, \phi) = \sum_{i=1}^{N} F_{\Omega,\Psi}(q, \mathbf{x}_i) \tag{65}$$

The gradient w.r.t. $\mathbf{W}$ is

$$\nabla R(\mathbf{W}, \phi) = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \nabla_{\mathbf{W}} L_\Omega(\mathbf{W}\mathbf{z}_i; \mathbf{x}_i/M_i) \tag{66}$$

From Proposition 2 of Blondel et al. [2020] the gradient of $L_\Omega(\boldsymbol{\theta}_i; \mathbf{x}_i/M_i)$ w.r.t. $\boldsymbol{\theta}_i$ is the residual vector (see p. 36)

$$\nabla_{\boldsymbol{\theta}_i} L_\Omega(\boldsymbol{\theta}_i; \mathbf{x}_i/M_i) = \hat{p}_\Omega(\boldsymbol{\theta}_i) - \mathbf{x}_i/M_i \in \mathbb{R}^D \tag{67}$$

Using the chain rule and setting $\boldsymbol{\theta}_i = \mathbf{W}_{D\times K}\mathbf{z}_{i_{K\times 1}}$ we can calculate the gradient of scalar loss L_Ω w.r.t. $\mathbf{W}$:[31]

$$\frac{\partial L_\Omega}{\partial \mathbf{W}}(\boldsymbol{\theta}_i; \mathbf{x}_i/M_i) = \left(\mathbf{z}_i \left(\frac{\partial L_\Omega}{\partial \boldsymbol{\theta}_i}\right)^{\mathrm{T}}\right)^{\mathrm{T}} = \frac{\partial L_\Omega}{\partial \boldsymbol{\theta}_i}\mathbf{z}_i^{\mathrm{T}} = [\hat{p}_\Omega(\boldsymbol{\theta}_i) - \mathbf{x}_i/M_i]\mathbf{z}_i^{\mathrm{T}} \tag{68}$$

So,

$$\nabla_{\mathbf{W}} R(\mathbf{W}, \phi) = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}[\hat{p}_\Omega(\boldsymbol{\theta}_i) - \mathbf{x}_i/M_i]\mathbf{z}_i^{\mathrm{T}} \approx \sum_{i=1}^{N}[\hat{p}_\Omega(\mathbf{W}\boldsymbol{m}_i) - \mathbf{x}_i/M_i]\mathbf{m}_i^{\mathrm{T}} \tag{69}$$

plugging in the posterior mean to get the last expression. And

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \eta \nabla_{\mathbf{W}} R(\mathbf{W}, \phi) \tag{70}$$

where the specific ϕ defines $\mathbf{m}$. Alternatively, instead of the MAP we can estimate the $\mathbf{W}$ update with MC samples

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \eta \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \nabla_{\mathbf{W}} L_\Omega(\mathbf{W}\mathbf{z}_i; \mathbf{x}_i/M_i) \tag{71}$$

The gradient w.r.t. ϕ is

$$\nabla_\phi F_{\Omega,\Psi}(q; \mathbf{x}_i) = \nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}[L_\Omega(\boldsymbol{\theta}(\mathbf{W}\mathbf{z}_i; \mathbf{x}_i/M_i))] + \nabla_\phi \mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}_i) \,\|\, p(\mathbf{z})) \tag{72}$$

The second term is easy to calculate. For the first term, the variational Gaussian distribution is reparametrizable in the sense that we can obtain a sample from the variational posterior by sampling from a base noise distribution and applying a transformation

$$\epsilon_{ij} \sim \mathcal{N}(0, 1) \quad z_{ij} = m_{ij} + \sigma_{ij}\epsilon_{ij} \tag{73}$$

In vector notation (Kingma et al. [2019], p.25):

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K\times K}) \quad \mathbf{z}_i = \mathbf{m}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon} \tag{74}$$

[31] And using the fact that $\frac{\partial}{\partial \mathbf{W}^{\mathrm{T}}} L_\Omega(\boldsymbol{\theta}_i^{\mathrm{T}}; \mathbf{x}_i/M_i) = (\mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}} \frac{\partial}{\partial \boldsymbol{\theta}_i^{\mathrm{T}}} L_\Omega(\boldsymbol{\theta}_i^{\mathrm{T}}; \mathbf{x}_i/M_i) = \mathbf{z}_i(\frac{\partial L_\Omega}{\partial \theta_{i1}}, ..., \frac{\partial L_\Omega}{\partial \theta_{iD}}) = \mathbf{z}_i \left(\frac{\partial L_\Omega}{\partial \boldsymbol{\theta}_i}\right)^{\mathrm{T}}$.

where $\mathbf{m}_i = \begin{pmatrix} m_{1i} \\ \vdots \\ m_{Ki} \end{pmatrix}$, $\boldsymbol{\sigma}_i = \begin{pmatrix} \sigma_{1i} \\ \vdots \\ \sigma_{Ki} \end{pmatrix}$, and $\odot$ is the element-wise product. We may now express the gradient w.r.t. ϕ as (Kim et al. [2018])

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}[L_\Omega(\mathbf{W}\mathbf{z}_i; \mathbf{x}_i/M_i)] &= \nabla_\phi \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} L_\Omega(\mathbf{W}(\mathbf{m}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}); \mathbf{x}_i/M_i) \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \nabla_\phi L_\Omega(\mathbf{W}(\mathbf{m}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}); \mathbf{x}_i/M_i) \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \nabla_{\boldsymbol{\theta}_i} L_\Omega(\boldsymbol{\theta}_i; \mathbf{x}_i/M_i) \frac{\partial \mathbf{W}(\mathbf{m}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon})}{\partial \phi}{}_{32} \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}[\hat{p}(\mathbf{W}(\mathbf{m}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}) - \mathbf{x}_i/M_i] \frac{\partial \mathbf{W}(\mathbf{m}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon})}{\partial \phi}
\end{aligned} \tag{75}$$

We can approximate the expectation in the gradient with a single sample $\boldsymbol{\epsilon}$ (Kim et al. [2018], p.24).

$$\phi^{(t+1)} = \phi^{(t)} + \eta \nabla_\phi F_\Omega(q, \mathbf{x}_i) \tag{76}$$

If we don't express the variational posterior parameters as NNs then $\phi = [m_1, ..., m_N, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_N]$ and for observation i we update $m_i, \boldsymbol{\Sigma}_i$.

References

S.-i. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

J. Atchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

M. Blondel, A. F. Martins, and V. Niculae. Learning with fenchel-young losses. *The Journal of Machine Learning Research*, 21(1):1314–1382, 2020.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems*, 14, 2001.

R. Cotterell, A. Poliak, B. Van Durme, and J. Eisner. Explaining and generalizing skip-gram through exponential family principal component analysis. In *Proceedings of the 15th conference*

[32] using again Proposition 2 of Blondel et al. [2020]

of the European chapter of the association for computational linguistics: volume 2, short papers, pages 175–181, 2017.

Y. Fang and S. Subedi. Clustering microbiome data using mixtures of logistic normal multinomial models. *Scientific Reports*, 13(1):14758, 2023.

Y. Guan and J. Dy. Sparse probabilistic principal component analysis. In *Artificial Intelligence and Statistics*, pages 185–192. PMLR, 2009.

G. James, D. Witten, T. Hastie, R. Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 366–373. JMLR Workshop and Conference Proceedings, 2010.

Y. Kim, S. Wiseman, and A. M. Rush. A tutorial on deep latent variable models of natural language. *arXiv preprint arXiv:1812.06834*, 2018.

D. P. Kingma, M. Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

J. Li and D. Tao. Simple exponential family pca. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 453–460. JMLR Workshop and Conference Proceedings, 2010.

M. Lu, J. Z. Huang, and X. Qian. Sparse exponential family principal component analysis. *Pattern recognition*, 60:681–691, 2016.

J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Martins and R. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016.

A. F. Martins, M. Treviso, A. Farinhas, P. M. Aguiar, M. A. Figueiredo, M. Blondel, and V. Niculae. Sparse continuous distributions and fenchel-young losses. *The Journal of Machine Learning Research*, 23(1):11728–11801, 2022.

C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, linear, and mixed models*, volume 325. Wiley Online Library, 2001.

S. Mohamed. *Generalised Bayesian matrix factorisation models*. PhD thesis, University of Cambridge, 2011.

S. Mohamed, Z. Ghahramani, and K. A. Heller. Bayesian exponential family pca. *Advances in neural information processing systems*, 21, 2008.

J. T. Morton, J. Silverman, G. Tikhonov, H. Lähdesmäki, and R. Bonneau. Scalable estimation of microbial co-occurrence networks with variational autoencoders. *bioRxiv*, pages 2021–11, 2021.

K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

T. Sears et al. Generalized maximum entropy, convexity and machine learning. 2008.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

M. Welling, C. Chemudugunta, and N. Sutter. Deterministic latent variable models and their pitfalls. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 196–207. SIAM, 2008.

F. Xia, J. Chen, W. K. Fung, and H. Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.

Y. Zeng, D. Pang, H. Zhao, and T. Wang. A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, pages 1–14, 2022.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.