Mark Goldstein and Narendra Mukherjee: 6.882 project model

1 Introduction

Let $\mathbf{X}$ be signal data segmented into N time windows of ℓ samples each. We are interested in inferring the frequency components present in each time window. To do this, we will use a non-parametric latent feature model with a Beta Process prior. In the generative story, each time window of the signal exhibits a blend of several latent features. Each of these latent features is defined by a sin at a particular frequency. There are an infinite amount of such sins. Each time window exhibits a finite amount of these sins.

Let $\mathbf{X} \in \mathbb{R}^{D=1\times N}$ be a matrix of the time windows. We set the dimension D of each time window to $D = 1$ and consider each time window as an indivisible scalar-like entity for the purposes of the following setup, though each window is actually an array of ℓ samples. The only place where we will need to consider the underlying samples is in the likelihood term. Otherwise, we reason only at the window level by assigning blends of frequencies to windows, where each frequency is also ℓ samples.

Let $\mathbf{\Phi} \in \mathbb{R}^{D=1\times K}$ be a matrix of parameters for a set of K signal basis elements to be allocated as latent features to the time windows. Each ϕ_k is the log of the frequency parameter for the k^{th} basis element, $\sin(\exp[\phi_k])$. We misuse notation by considering a matrix product with $\mathbf{\Phi}$ to be a combination of $\sin(\exp[\phi_k])$'s rather than a combination of ϕ_k's. So if $\mathbf{a} \in \mathbb{R}^K$, then we let $\mathbf{\Phi a} = \sum_{k=1}^{K} a_k \sin(\exp[\phi_k])$. Let $\mathbf{Z} \in \mathbb{R}^{K\times N}$ be a binary matrix indicating the presence of the k^{th} latent feature in the i^{th} time window of the signal with $k \in \{1 \dots K\}$ and $i \in \{1 \dots N\}$. Let $\mathbf{W} \in \mathbb{R}^{K\times N}$ be the weights of the k^{th} latent feature in the i^{th} time window. Let $\mathbf{E} \in \mathbb{R}^{D\times N}$ be a noise matrix, such that $X = \mathbf{\Phi}(\mathbf{Z} \circ \mathbf{W}) + \mathbf{E}$ with $x_i = \mathbf{\Phi}(\mathbf{z}_i \circ \mathbf{w}_i) + \epsilon_i$. The generative model is:

$$\begin{aligned}
\epsilon_i &\sim \mathcal{N}(0, \sigma_i^2) \\
w_{ik} &\sim \mathcal{N}(0, \sigma_w^2) \\
\phi_k = &\sim \mathcal{N}(\mu_\phi, \sigma_\phi^2) \\
z_{ik} &\sim \text{Bernoulli}(\pi_k) \\
\pi_k &\sim \text{Beta}(a/K, b(K-1)/K)
\end{aligned}$$

This corresponds to a Beta Process prior on Z. The logs of the frequency parameters ϕ_k are drawn $i.i.d$ from base distribution H_0, in this case the Normal. The π_k's are drawn from $\text{Beta}(a/K, b(K = 1)/K)$. This gives us $H = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \sim BP(a, b, H_0)$ as $K \to \infty$. (Paisley 2009).

Our likelihood model is that each window i of the true signal, now to be thought of as ℓ-samples-dimensional, is normally distributed around the i^{th} time window of the estimated signal:

$$\hat{\mathbf{x}}_i = \sum_{k=1}^{K} z_{ik} w_{ik} \big[\sin\big(\exp[\phi_k] \big) \big]$$

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \mathbf{\Phi}, \{\sigma_i^2\}_{i=1}^{N}) = \sum_{i=1}^{N} \log \mathcal{N}(\mathbf{x}_i | \hat{\mathbf{x}}_i, \sigma_i^2 \mathbf{I})$$

2 Variational Inference Overview

Let $\mathbf{B} = \{\boldsymbol{\pi}, \mathbf{Z}, \boldsymbol{\Phi}, \mathbf{W}\}$ and $\boldsymbol{\theta} = \{\sigma_i^2, \sigma_w^2, \mu_\phi, \sigma_\phi^2, a, b\}$. We will approximate the true posterior $p(\mathbf{B}|\mathbf{X}, \boldsymbol{\theta})$, intractable because of the *evidence* $p(X)$, with a variational distribution $q(\mathbf{B})$ that minimizes the divergence $\mathrm{KL}(q(\mathbf{B})||p(\mathbf{B}|\mathbf{X}, \boldsymbol{\theta})$. Let our variational distribution take the form $q_{\boldsymbol{\eta}}(\mathbf{B}) = q_{\boldsymbol{\tau}}(\boldsymbol{\pi})q_{\boldsymbol{\gamma}}(\boldsymbol{\Phi})q_{\boldsymbol{\lambda}}(\mathbf{W})q_{\boldsymbol{\nu}}(\mathbf{Z})$. We optimize q's parameters $\boldsymbol{\eta} = \{\boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\nu}\}$ to most closely match the true posterior. Omit conditioning on the hyperparameters $\boldsymbol{\theta}$ for brevity. Minimizing

$$\mathrm{KL}(q(\mathbf{B})||p(\mathbf{B}|\mathbf{X})) = \mathbb{E}_q[\log q(\mathbf{B})] - \mathbb{E}_q[\log p(\mathbf{B}, \mathbf{X})] + \log p(\mathbf{X})$$

still depends on the intractable term $p(\mathbf{X})$. We drop $p(\mathbf{X})$ because it does not depend on q. We maximize the negative of the leftover terms

$$\mathrm{ELBO}[q] = \mathbb{E}_q[\log p(\mathbf{B}, \mathbf{X})] - \mathbb{E}_q[\log q(\mathbf{B})]$$

The first term is the joint of the data and the model and the second term is the entropy $H[q]$ of the variational distribution. We expand the first term and get

$$\mathrm{ELBO}[q] = \mathbb{E}_q[\log p(\mathbf{B})] + \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{B})] - H[q]$$

The first term considers the priors, the middle term the likelihood, and the last term the entropy of the variational distribution. This also equals

$$\mathrm{ELBO}[q] = \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{B})] - \mathrm{KL}(q(\mathbf{B})||p(\mathbf{B}))$$

which makes explicit that we pick a q that gives us high likelihood but that stays close to the priors. This optimization is non-convex and we are only guaranteed to find local optima. We cycle through each variational parameter, and perform coordinate ascent to optimize the ELBO. The ELBO has its name because this is equivalent to maximizing a lower bound on the evidence (also called marginal likelihood)

$$\begin{aligned}
q_{\tau_k}(\pi_k) &= \mathrm{Beta}(\pi_k|\tau_k^a, \tau_k^b) \\
q_{\gamma_k}(\phi_k) &= \mathcal{N}(\gamma_k^\mu, \gamma_k^{\sigma^2}) \\
q_{\lambda_{ik}}(w_{ik}) &= \mathcal{N}(\lambda_{ik}^\mu, \lambda_{ik}^{\sigma^2}) \\
q_{\nu_{ik}}(z_{ik}) &= \mathrm{Bernoulli}(z_{ik}|\nu_{ik})
\end{aligned}$$

3 Expanding the ELBO:

$$\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\theta}) &\geq \mathbb{E}_{q(\mathbf{B})}[\log p(\mathbf{X},\mathbf{B}|\boldsymbol{\theta})] + H[q] \\
&= \sum_{k=1}^{K} \mathbb{E}_{q(\pi_k)}[\log p(\pi_k|a,b)] + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{q(\pi_k),q(z_{ik})}[\log p(z_{ik}|\pi_k)] \\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{q(\phi_k)}[\log p(\phi_k|\mu_\phi,\sigma_\phi^2)] + \sum_{k=1}^{K} \mathbb{E}_{q(w_{ik})}[\log p(w_{ik}|\sigma_w^2)] \\
&\quad + \sum_{i=1}^{N} \mathbb{E}_{q(z_i),q(\phi),q(w_i)}[\log p(x_i|\mathbf{z_i},\boldsymbol{\phi},\mathbf{w}_i,\sigma_i^2)] + H[q]
\end{aligned}$$

Next, we must substitute in the actual distributions for each of these terms. We now simplify each term separately before presenting the whole *ELBO* in its expanded form.

3.1 $\mathbb{E}_{q(\pi_k)}[\log p(\pi_k|a,b)]$

The Beta PDF is $\text{Beta}(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$ where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. The Gamma function $\Gamma()$ has the property that $\Gamma(x+1) = (x)\Gamma(x)$. Let $\psi()$ be the digamma function. $\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. Our distributions are:

$$\begin{aligned}
p(\pi_k) &= \text{Beta}(\alpha,\beta) \\
q_{\tau_k}(\pi_k) &= \text{Beta}(\pi_k|\tau_k^a,\tau_k^b)
\end{aligned}$$

where $\alpha = \frac{a}{K}$ and $\beta = b(K-1)/K$, with $a,b \in \boldsymbol{\theta}$ from the original model specification.

$$\begin{aligned}
\mathbb{E}_{q(\pi_k)}[\log p(\pi_k|\alpha,\beta)] &= \mathbb{E}_{q(\pi_k)}[\log\big(\frac{\pi_k^{\alpha-1}(1-\pi_k)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}\big)] \\
&= \mathbb{E}_{q(\pi_k)}[\log\big(\pi_k^{\alpha-1}(1-\pi_k)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\big)] \\
&= \mathbb{E}_{q(\pi_k)}[(\alpha-1)\log(\pi_k) + (\beta-1)\log(1-\pi_k) + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta)] \\
&= (\alpha-1)\mathbb{E}_{q(\pi_k)}[\log(\pi_k)] + (\beta-1)\mathbb{E}_{q(\pi)}[\log(1-\pi)] + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) \\
&= (\alpha-1)(\psi(\tau_k^a) - \psi(\tau_k^a+\tau_k^b)) + (\beta-1)(\psi(\tau_k^b) - \psi(\tau_k^a+\tau_k^b)) \\
&\quad + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta)
\end{aligned}$$

The above uses the property that if $q(\pi_k) = \text{Beta}(\pi_k|\tau_k^a,\tau_k^b)$ then $\mathbb{E}_{q(\pi_k)}[\log\pi_k] = \psi(\tau_k^a) - \psi(\tau_k^a+\tau_k^b)$ and $\mathbb{E}_{q(\pi_k)}[\log(1-\pi_k)] = \psi(\tau_k^b) - \psi(\tau_k^a+\tau_l^b)$. This comes from the property that $\pi_k \sim \text{Beta}(\tau_k^a,\tau_k^b)$ means that $1-\pi_k$ is distributed $\text{Beta}(\tau_k^b,\tau_k^a)$. The constants that do not depend on q will come out in the derivations of the coordinate ascent updates.

3.2 $\mathbb{E}_{q(\pi_k),q(z_{ik})}[\log p(z_{ik}|\pi_k)]$

Recall that $q(z) = \text{Bernoulli}(z_ik|\nu_{ik})$ and $q(\pi_k) = \text{Beta}(\pi_k|\tau_k^a,\tau_k^b)$

$$
\begin{aligned}
\mathbb{E}_{q(\pi_k),q(Z)}[\log p(z_{nk}|\pi_k)] &= \mathbb{E}_{q(\pi_k),q(Z)}\big[\log\big(\pi_k^{z_{ik}}(1-\pi_k)^{1-z_{ik}}\big)\big] \\
&= \mathbb{E}_{q(\pi_k),q(Z)}\big[z_{ik}\log\pi_k + (1-z_{ik})\log(1-\pi_k) big] \\
&= \mathbb{E}_{q(Z)}[z_{ik}]\mathbb{E}_{q(\pi_k)}[\log\pi_k] + (1-\mathbb{E}_{q(Z)}[z_{ik}])\mathbb{E}_{q(\pi_k)}[\log(1-\pi_k)] \\
&= \nu_{ik}\psi(\tau_k^a) + (1-\nu_{ik})\psi(\tau_k^b) - \psi(\tau_k^a+\tau_k^b)
\end{aligned}
$$

3.3 $\mathbb{E}_{q(\phi_k)}[\log p(\phi_k|\mu_\phi,\sigma_\phi^2)]$

Recall that $p(\phi_k) = \mathcal{N}(\mu_\phi,\sigma_\phi^2)$ and $q_{\gamma_k}(\phi_k) = \mathcal{N}(\gamma_k^\mu,\gamma_k^{\sigma^2})$

$$
\begin{aligned}
\mathbb{E}_{q(\phi_k)}[\log p(\phi_k|\mu_\phi,\sigma_\phi^2)] &= \mathbb{E}_{q(\phi_k)}\Big[\log\Big(\frac{1}{\sqrt{2\pi\sigma_\phi^2}}\exp\big[-\frac{(\phi_k-\mu_\phi)}{2\sigma_\phi^2}]\Big)\Big] \\
&= -\frac{1}{2}\log\Big[2\pi\sigma_\phi^2\Big] - \frac{\gamma_k^{\sigma^2} + \big(\gamma_k^\mu - \mu_\phi\big)^2}{2\sigma_\phi^2}
\end{aligned}
$$

3.4 $\mathbb{E}_{q(w_{ik})}[\log p(w_{ik}|\sigma_w^2)]$

Recall that $p(w_{ik}) = \mathcal{N}(0,\sigma_w^2)$ and $q_{\lambda_{ik}}(w_{ik}) = \mathcal{N}(\lambda_{ik}^\mu,\lambda_{ik}^{\sigma^2})$. Using the result just above, we get

$$
\mathbb{E}_{q(w_{ik})}[\log p(w_{ik}|\sigma_w^2)] = -\frac{1}{2}\log\Big[2\pi\sigma_w^2\Big] - \frac{\lambda_{ik}^{\sigma^2} + \big(\lambda_{ik}^\mu - 0\big)^2}{2\sigma_w^2}
$$

3.5 $\mathbb{E}_{q(z_i),q(\phi),q(w_i)}[\log p(x_i|\mathbf{z_i},\boldsymbol{\phi},\mathbf{w}_i,\sigma_i^2)]$ (likelihood term)

...

3.6 The expanded ELBO with our concrete choices of p's and q's

...

4 Coordinate Ascent Updates

...

References

- Bretthorst. Bayesian Spectrum Analysis and Parameter Estimation. 1988.
- Doshi-Velez et al. Variational Inference for the IBP. May (not April) 2009.
- Liang and Hoffman. Beta Process Non-negative Matrix Factorization with Stochastic Structured Mean-Field Variational Inference. 2014.
- Paisley and Carin. Nonparametric Factor Analysis with Beta Process Priors. 2009.

- Paisley, Carin, and Blei. Variational Inference for Stick-Breaking Beta Process Priors. 2011.
- Paisley, Blei, and Jordan. Stick-Breaking Beta Processes and the Poisson Process. 2012.
- Turner and Sahani. Time-frequency analysis as probabilistic inference. 2014.