Deep Learning with Hierarchical Convolutional Factor Analysis

[1]Bo Chen, [2]Gungor Polatkan, [3]Guillermo Sapiro, [2]David Blei,
[4]David Dunson and [1]Lawrence Carin

[1]Electrical & Computer Engineering Department, [4]Statistical Sciences Department
Duke University, Durham, NC, USA

[2]Computer Science Department
Princeton University, Princeton, NJ

[3]Electrical & Computer Engineering Department
University of Minnesota, Minneapolis, MN, USA

SUPPLEMENTARY MATERIAL: GIBBS UPDATE EQUATIONS

For each MCMC iteration, the samples are drawn from the following conditional distributions:

- Sampling b_{nk} and $\mathbf{W}_{nk}$ ($i.e.$, $\{w_{nki}\}_{i\in\mathcal{S}}$): we have

$$p(b_{nk}=1|-) = \widetilde{\pi}_{nk} \tag{1}$$

$$p(w_{nki}, i\in\mathcal{S}|-) = (1-b_{nk})\mathcal{N}(0,\alpha_{nki}^{-1}) + b_{nk}\mathcal{N}(\mu_{nki},\Sigma_{nki}) \tag{2}$$

where

$$\Sigma_{nki} = (\boldsymbol{d}_{ki}^T\boldsymbol{d}_{ki}\gamma_n + \alpha_{nki})^{-1} \tag{3}$$

$$\mu_{nki} = \Sigma_{nki}\gamma_n\mathbf{X}_{nki}^T\boldsymbol{d}_{ki} \tag{4}$$

$$\frac{\widetilde{\pi}_{nk}}{1-\widetilde{\pi}_{nk}} = \frac{\pi_k}{1-\pi_k}\cdot\frac{\mathcal{N}(\mathbf{X}_{nk}|\mathbf{W}_{nk}*\boldsymbol{d}_k,\gamma_n^{-1}\mathbf{I}_P)}{\mathcal{N}(\mathbf{X}_{nk}|\mathbf{0},\gamma_n^{-1}\mathbf{I}_P)} \tag{5}$$

Here $\mathbf{X}_{nk}=\mathbf{X}_{-n}+\mathbf{W}_{nk}*\boldsymbol{d}_k$, $\mathbf{X}_{-n}=\mathbf{X}_n-\sum_{k=1}^K b_{nk}\mathbf{W}*\boldsymbol{d}_k$ and b_{nk} is the most recent sample. Taking advantage of the convolution property, we simultaneously update the posterior mean and covariance of the coefficients for all the shifted versions of one dictionary element. Consequently,

$$\boldsymbol{\Sigma}_{nk} = \mathbf{1}\oslash(\gamma_n\|\boldsymbol{d}_k\|_2^2 b_{nk}+\boldsymbol{\alpha}_{nk}) \tag{6}$$

$$\boldsymbol{\mu}_{nk} = b_{nk}\gamma_n\boldsymbol{\Sigma}_{nk}\odot(\mathbf{X}_{-n}*\boldsymbol{d}_k+\|\boldsymbol{d}_k\|_2^2\mathbf{W}_{nk}) \tag{7}$$

where both of $\boldsymbol{\Sigma}_{nk}$ and $\boldsymbol{\mu}_{nk}$ have the same size with $\mathbf{W}_{nk}$. The symbol $\odot$ is the element-wise product operator and $\oslash$ the element-wise division operator.

After sampling $\mathbf{W}_{nk}$, we perform max-pooling on it at each Gibbs iteration (within each max-pooling region, we set all coefficients to zero, other than the maximum-amplitude coefficient). We do this for two reasons: (i) we have found that max-pooling can speed up the model inference, in that it explicitly imposes sparseness; and (ii) as stated in [**?**], without sparsifying the hidden units the parameter learning may be unstable (unless one employs a very large number of Gibbs samples, which is undesirable computationally). We now continue to the rest of the Gibbs draws.

- Sampling $\boldsymbol{d}_k$:

$$P(d_{kj}|-) = \mathcal{N}(\xi_{kj}, \Lambda_{kj}) \tag{8}$$

  where ξ_{kj} and Λ_{kj} is the j-th element of the following vectors ξ_k and Λ_k respectively

$$\Lambda_k = \mathbf{1} \oslash (\sum_{n=1}^{N} \boldsymbol{\gamma}_n b_{nk} \|\mathbf{W}_{nk}\|_2^2 + \beta_k), \tag{9}$$

$$\boldsymbol{\xi}_k = \boldsymbol{\Lambda}_k \odot (\sum_{n=1}^{N} b_{nk}\gamma_n(\mathbf{X}_{-n} * \mathbf{W}_{nk} + \boldsymbol{d}_k \|\mathbf{W}_{nk}\|_2^2)). \tag{10}$$

  Here $\mathbf{X}_{nki(i)} \in \mathbb{R}^J$ represents the corresponding values of $\mathbf{X}_{nki}$ at the same support locations of the i-th shifted version.

- Sampling π_k:

$$P(\pi_k|-) = \text{Beta}(\widetilde{a}, \widetilde{b}), \tag{11}$$

  where $\widetilde{a} = \sum_{n=1}^{N} b_{nk} + \frac{1}{K}, \qquad \widetilde{b} = N + b - \sum_{n=1}^{N} b_{nk}.$

- Sampling γ_n:

$$P(\gamma_n|-) = \text{Gamma}(\widetilde{c}, \widetilde{d}), \tag{12}$$

  where $\widetilde{c} = c + \frac{1}{2}, \qquad \widetilde{d} = d + \frac{1}{2}\|\mathbf{X}_n\|_2^2.$

- Sampling α_{nki}:

$$P(\alpha_{nki}|-) = \text{Gamma}(\widetilde{e}, \widetilde{f}), \tag{13}$$

  where $\widetilde{e} = e + \frac{1}{2}, \qquad \widetilde{f} = f + \frac{1}{2}w_{nki}^2.$

- Sampling β_{kj}:

$$P(\beta_{kj}|-) = \text{Gamma}(\widetilde{g}, \widetilde{h}), \tag{14}$$

  where $\widetilde{g} = g + \frac{1}{2}, \qquad \widetilde{h} = h + \frac{1}{2}d_{kj}^2.$

During the inference, we perform efficient block Gibbs sampling by sampling each dictionary element and corresponding hidden units given other parameters in each layer. Throughout the inference procedure, there is no matrix inversion since the dictionary and hidden units (factor scores) have been factorized. We explicitly utilize the dictionary element matrix $\boldsymbol{d}_k \in \mathbb{R}^{n'_x \times n'_y}$ instead of the zero-padded dictionary matrix, which is far smaller than the size of images. Moreover, thanks to the sparseness of the binary matrix b_{nk} and the max-pooling operation, the sparse matrix $\mathbf{W}_n$ only requires modest computer memory. In practice we typically do not need a large number of Gibbs iterations to achieve useful results (although, from a pure statistics standpoint, rigorous convergence for the full posterior distribution has certainly not been achieved).

SUPPLEMENTARY MATERIAL: BATCH VB UPDATE EQUATIONS

- For b^l_{nk} we have $q(b^l_{nk}) = \text{Bernoulli}(b^l_{nk}; \rho^l_{nk})$ where ρ^l_{nk} is the possibility of $b^l_{nk} = 1$. We consider the following two conditions:

$$\begin{aligned} \ln[q(b^l_{nk}=1)] &\propto \zeta_1 = \langle \ln \pi^l_k \rangle - \frac{\langle \gamma^l_n \rangle}{2}(-2\langle \| \boldsymbol{X}^l_{nk} \odot (\boldsymbol{d}^l_k * \boldsymbol{W}^l_{nk}) \|_2^2 \rangle + \langle \| \boldsymbol{W}^l_{nk} \|_2^2 \rangle \langle \| \boldsymbol{d}^l_k \|_2^2 \rangle) \\ \ln[q(b^l_{nk}=0)] &\propto \zeta_2 = \langle \ln(1-\pi^l_k) \rangle \end{aligned} \tag{15}$$

  where $\mathbf{X}^l_{nk} = \mathbf{X}^l_{-n} + \mathbf{W}^l_{nk} * \boldsymbol{d}^l_k$, $\mathbf{X}^l_{-n} = \mathbf{X}^l_n - \sum_{k=1}^{K^l} b^l_{nk} \mathbf{W}^l * \boldsymbol{d}^l_k$ and b^l_{nk} is the most recent estimation, $\langle \ln \pi^l_k \rangle = \Psi(\frac{1}{K^l} + \sum_{n=1}^N \langle b^l_{nk} \rangle) - \Psi(\frac{1}{K^l} + a + N)$, $\langle \ln(1-\pi^l_k) \rangle = \Psi(a + N - \sum_{n=1}^N \langle b^l_{nk} \rangle) - \Psi(\frac{1}{K^l} + a + N)$, $\Psi(x) = \frac{\partial}{\partial x} \ln \Gamma(x)$ and $\Gamma(x) = \int_0^\infty d\tau \tau^{x-1} e^{-\tau}$. Therefore, we can calculate $\rho^l_{nk} = \frac{1}{1+\exp(\zeta_0 - \zeta_1))}$.
- For π^l_k we have $q(\pi^l_k) = \text{Beta}(\pi^l_k; \tau^l_{k1}, \tau^l_{k2})$ where

$$\tau^l_{k1} = \sum_{n=1}^N \langle b^l_{nk} \rangle + \frac{1}{K^l}, \quad \tau^l_{k2} = N + b - \sum_{n=1}^N \langle b^l_{nk} \rangle \tag{16}$$

- For $\boldsymbol{d}^l_k$ we have $q(\boldsymbol{d}^l_k) = \mathcal{N}(\boldsymbol{d}^l_k; \boldsymbol{\xi}^l_k, \boldsymbol{\Lambda}^l_k)$, with

$$\begin{aligned} \boldsymbol{\Lambda}^l_k &= \mathbf{1} \oslash (\sum_{n=1}^N \langle \gamma^l_n \rangle \langle b^l_{nk} \rangle \langle \| \mathbf{W}^l_{nk} \|_2^2 \rangle + \langle \beta^l_k \rangle) \\ \boldsymbol{\xi}^l_k &= \boldsymbol{\Lambda}^l_k \odot (\sum_{n=1}^N \langle b^l_{nk} \rangle \langle \gamma^l_n \rangle (\mathbf{X}^l_{-n} * \langle \mathbf{W}^l_{nk} \rangle + \langle \boldsymbol{d}^l_k \rangle \langle \| \mathbf{W}^l_{nk} \|_2^2 \rangle)). \end{aligned} \tag{17}$$

- For $\mathbf{W}^l_{nk}$ we have $q(w^l_{nki}) = \mathcal{N}(w^l_{nki}; \mu^l_{nki}, \Sigma^l_{nki})$, with

$$\begin{aligned} \boldsymbol{\Sigma}^l_{nk} &= \mathbf{1} \oslash (\langle \gamma^l_n \rangle \langle \| \boldsymbol{d}^l_k \|_2^2 \rangle \langle b^l_{nk} \rangle + \langle \boldsymbol{\alpha}^l_{nk} \rangle) \\ \boldsymbol{\mu}^l_{nk} &= \langle b^l_{nk} \rangle \langle \gamma^l_n \rangle \boldsymbol{\Sigma}^l_{nk} \odot (\mathbf{X}^l_{-n} * \langle \boldsymbol{d}^l_k \rangle + \langle \| \boldsymbol{d}^l_k \|_2^2 \rangle \langle \mathbf{W}^l_{nk} \rangle), \end{aligned} \tag{18}$$

where both of $\mathbf{\Sigma}_{nk}^{l}$ and $\boldsymbol{\mu}_{nk}^{l}$ have the same size with $\mathbf{W}_{nk}^{l}$.

- For γ_n^l we have $q(\gamma_n^l) = \text{Gamma}(\gamma_n^l; \lambda_{n1}^l, \lambda_{n2}^l)$, where

$$\lambda_{n1}^l = c + \frac{p^l}{2}, \quad \lambda_{n2}^l = d + \frac{1}{2}\langle \|\mathbf{X}_n^l - \sum_{k=1}^{K^l} b_{nk}^l \boldsymbol{d}_k^l * \mathbf{W}_{nk}^l\|_2^2 \rangle \tag{19}$$

- For β_{kj}^l we have $q(\beta_{kj}^l) = \text{Gamma}(\beta_{kj}^l; \upsilon_{kj1}^l, \upsilon_{kj2}^l)$, with

$$\upsilon_{kj1}^l = g + 1/2, \quad \upsilon_{kj2}^l = h + \frac{1}{2}\langle (d_{kj}^l)^2 \rangle \tag{20}$$

- For α_{nki}^l we have $q(\alpha_{nki}^l) = \text{Gamma}(\alpha_{nki}^l; \nu_{nki1}^l, \nu_{nki2}^l)$, with

$$\nu_{nki1}^l = e + 1/2, \quad \nu_{nki2}^l = f + \frac{1}{2}\langle (w_{nki}^l)^2 \rangle \tag{21}$$