1 Modeling binary views

1.1 The binary model

In many settings, we encounter binary data as a view (e.g. gene mutations). In order to incorporate this type of data, we extended the Group Factor Analysis approach to non-Gaussian likelihoods. In the case of binary data an appropriate model is the Bernoulli likelihood, where we model the data Y as

$$Y|Z,W \sim Ber(\sigma(ZW)). \tag{1}$$

Here, $\sigma(a) = (1+e^{-a})^{-1}$ is the logistic link function and Z and W are the latent factors and weights in our model, respectively.

In order to make the inference explicit as in the Gaussian case, we aim to approximate the Bernoulli likelihood by a Gaussian likelihood as it has been proposed in [3]. This allows to recycle all the updates from the model with Gaussian views. To make the approximation justifiable we need to introduce variational parameters that are adjusted alongside the updates to improve the fit. While [3,] assumes a homoscedastic approximation with a spherical Gaussian, we adopt an approach following [2], which allows for heteroscedaticity and provides a tighter bound on the Bernoulli likelihood.

Denoting $x_{ij} = (ZW)_{ij}$ the Jaakolo upper bound [2] on the negative log-likelihood is given by

$$\begin{aligned} -\log\left(p(y_{ij}|x_{ij})\right) &= -\log\left(\sigma\left((2y_{ij}-1)x_{ij}\right)\right) && (2)\\ &\leq -\log(\zeta_{ij}) - \frac{(2y_{ij}-1)x_{ij}-\zeta_{ij})}{2} + \lambda(\zeta_{ij})\left(x_{ij}^2-\zeta_{ij}^2\right) && (3)\\ &=: b_J(\zeta_{ij}, x_{ij}, y_{ij}) && (4)\end{aligned}$$

with λ given by $\lambda(\zeta) = \frac{1}{4\zeta}\tanh\left(\frac{\zeta}{2}\right)$.

This can be derived easily from a first-order Taylor expansion on the function $f(x) = -\log(e^{\frac{x}{2}}+e^{-\frac{x}{2}}) = \frac{x}{2} - \log(\sigma(x))$ in x^2 and by the convexity of f in x^2 this bound is global as discussed in [2,].

In order to make use of this bound but still be able to handle the updates as in the Gaussian case we re-formulate the bound as a Gaussian likelihood on pseudo-data $\tilde{Y}$.

In the variational framework we want to update the variational distribution q(Z,W) in order to minimize

$$\mathbb{E}_q\left(-\log p(Y|Z,W)\right) \tag{5}$$

Using the upper bound in (4)

$$\min_{q(Z,W)} \mathbb{E}_q\left(-\log p(Y|Z,W)\right) \leq \sum_{i,j} \min_{q(x_{ij}),\zeta_{ij}} \mathbb{E}_q b_J(\zeta_{ij}, x_{ij}, y_{ij}) \tag{6}$$

This is minimized iteratively in the variational parameter ζ_{ij} and the variational distribution of Z,W.

Minimizing in the variational parameter ζ this leads to the updates given by

$$\zeta_{ij}^2 = \mathbb{E}x_{ij}^2 \tag{7}$$

as described in [2,], [1,].

For the variational distribution $q(Z, W)$ we observe that the Jaakkola bound can be re-written as

$$b_J(\zeta_{ij}, x_{ij}, y_{ij}) = -\log\left(\varphi\left(\tilde{y}_{ij}; x_{ij}, \frac{1}{2\lambda(\zeta_{ij})}\right)\right) + c(\zeta_{ij}), \tag{8}$$

where $\varphi(x; \mu, \sigma^2)$ denotes the density function of a normal distribution wih mean μ and variance σ^2 and c is a term only depending on ζ. This allows us to re-use the updates for Z and W from a setting with Gaussian likelihood by considering the Gaussian pseudo-data

$$\tilde{y}_{ij} = \frac{2y_{ij} - 1}{4\lambda(\zeta_{ij})} \tag{9}$$

updating the data precision as $\tau_{ij} = 2\lambda(\zeta_{ij})$ using updates that allow for sample- and feature-wise precision parameters on the data as described in 2.1.

2 Appendix

2.1 Extending the gaussian model to sample-wise covariances

We have a data set $\mathbf{Y}$ of M input views with dimensionality $\mathbf{Y}^m \in \mathbb{R}^{N \times D_m}$. The sparse group factor analysis model is defined as follows:

$$y_{nd} \approx \mathcal{N}\left(y^m_{nd} \,|\, \mathbf{w}^m_{d,:}\mathbf{z}_{n,:}, 1/\tau^m_{dn}\right) \tag{10}$$

$$w^m_{dk} \approx \theta\mathcal{N}\left(w_{dk} \,|\, 0, 1/\alpha^m_k\right) + (1-\theta)\delta_0(w^m_{dk}) \tag{11}$$

$$z_{nk} \approx \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \tag{12}$$

The precision of observed data is allowed to vary for samples and features to incorporate a pseudo-data view on the Jaakkola bound. This parameter is updated along with the variational parameter of the pseudo-data and is not treated as a variational node itself.

2.1.1 Changes to standard spike and slab

The updates affected by this change are the one of $w = s\hat{w}$ and z. In the ELB the Bernoulli likelihood of the truly observed data is used. For all other updates and terms refer to spike_slab.tex.

2.1.2 Joint probability density function:

$$\begin{aligned} p(\mathbf{Y}, \hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}) = &\prod_{m=1}^{M}\prod_{n=1}^{N}\prod_{d=1}^{D_m}\mathcal{N}\left(y^m_{nd} \,|\, \sum_{k=1}^{K} s^m_{dk}\hat{w}^m_{dk}z_{nk}, 1/\tau_{nd}\right) \times \\ &\prod_{d=1}^{D_m}\prod_{k=1}^{K}\mathcal{N}\left(\hat{w}^m_{dk} \,|\, 0, 1/\alpha^m_k\right)\theta^{s^m_{dk}}(1-\theta)^{1-s_{dk}} \times \\ &\prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \end{aligned}$$

Full variational distribution:

$$q(\hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} q(\hat{w}_{dk}^m, s_{dk}^m) \prod_{k=1}^{K} \prod_{n=1}^{N} q(z_{n,k})$$

2.1.3 Derivation of update equations

The optimal distribution $\hat{q}_i$ for each variable $\mathbf{x}_i$, is the following:

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const.} \tag{13}$$

where $\mathbb{E}_{i \neq j}$ denotes an expectation with respect to the q distributions over all variables $\mathbf{x}_j$ except for $\mathbf{x}_i$. The additive constant is set by normalising the distribution $\hat{q}_i(\mathbf{z}_i)$:

$$\hat{q}_i(\mathbf{x}_i) = \frac{\exp(\mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})])}{\int \exp(\mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})]) d\mathbf{X}}$$

Latent variables

Term from the likelihood $p(\mathbf{Y}|\hat{\mathbf{W}}, \mathbf{Z}, \boldsymbol{\tau}, \mathbf{S})$:

$$\sum_{m=1}^{M} \sum_{d=1}^{D=m} \langle \tau_{nd}^m \rangle \langle s_{dk}^m \hat{w}_{dk}^m \rangle y_{nd}^m z_{nk} - \frac{1}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau_{nd}^m \rangle \langle (s_{dk}^m \hat{w}_{dk}^m)^2 \rangle z_{nk}^2$$
$$- \frac{1}{2} \sum_{m=1}^{M} \frac{1}{2} \sum_{d=1}^{D} \langle \tau_{nd}^m \rangle \sum_{j \neq k} (\langle s_{dj}^m \hat{w}_{dj}^m \rangle \langle z_{nj} \rangle) \langle \hat{w}_{dk}^m s_{dk}^m \rangle \langle z_{nk} \rangle + \text{const.}$$

Term from the prior $p(z_{nk})$:

$$-\frac{1}{2} z_{nk}^2 + \text{const.}$$

Variational distribution:

$$q(\mathbf{Z}) = \prod_{k=1}^{K} \prod_{n=1}^{N} q(z_{nk}) = \prod_{k=1}^{K} \prod_{n=1}^{N} \mathcal{N}(z_{nk} \,|\, \mu_{z_{nk}}, \sigma_{z_{nk}})$$

where

$$\sigma_{z_{nk}}^2 = \Big(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \tau_{nd}^m \langle (s_{dk}^m \hat{w}_{dk}^m)^2 \rangle + 1 \Big)^{-1}$$

$$\mu_{z_{nk}} = \sigma_{z_{nk}}^2 \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau_{nd}^m \rangle \langle s_{dk}^m \hat{w}_{dk}^m \rangle \Big(y_{nd}^m - \sum_{j \neq k} \langle s_{dj}^m \hat{w}_{dj}^m \rangle \langle z_{nj} \rangle \Big)$$

Spike and Slab Weights

Variational distribution:

$$q(\hat{\mathbf{W}}, \mathbf{S}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} q(\hat{w}_{dk}^m, s_{dk}^m) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} q(\hat{w}_{dk}^m | s_{dk}^m) q(s_{dk}^m)$$

Update for $q(s^m_{dk})$:

$$\gamma_{dk} = q(s_{dk} = 1) = \frac{1}{1 + \exp(-\lambda_{dk})} = logit^{-1}(\lambda_{dk})$$

where

$$\lambda^m_{dk} = \langle \log \frac{\theta}{1-\theta} \rangle + 0.5 \frac{\left(\sum_{n=1}^N y^m_{nd} \langle z_{nk} \rangle \tau^m_{nd} - \sum_{j \neq k} \langle s^m_{dj} \hat{w}^m_{dj} \rangle \sum_{n=1}^N \langle z_{nk} \rangle \langle z_{nj} \rangle \tau^m_{nd} \right)^2}{\sum_{n=1}^N \tau^m_{nd} \langle z^2_{nk} \rangle + \langle \alpha^m_k \rangle}$$
$$+ 0.5 \log(\langle \alpha^m_k \rangle) - 0.5 \log(\sum_{n=1}^N \tau^m_{nd} \langle z^2_{nk} \rangle + \langle \alpha^m_k \rangle)$$

Update for $q(\hat{w}^m_{dk})$:

$$q(\hat{w}^m_{dk} | s^m_{dk} = 0) = \mathcal{N}\left(\hat{w}^m_{dk} \,|\, 0, 1/\alpha^m_k\right)$$
$$q(\hat{w}^m_{dk} | s^m_{dk} = 1) = \mathcal{N}\left(\hat{w}^m_{dk} \,|\, \mu_{w^m_{dk}}, \sigma^2_{w^m_{dk}}\right)$$

where

$$\mu_{w^m_{dk}} = \sigma^2_{w^m_{dk}} \left(\sum_{n=1}^N \tau^m_{nd} y^m_{nd} \langle z_{nk} \rangle - \sum_{j \neq k} \langle s^m_{dj} \hat{w}^m_{dj} \rangle \sum_{n=1}^N \tau^m_{nd} \langle z_{nk} \rangle \langle z_{nj} \rangle \right)$$

$$\sigma^2_{w^m_{dk}} = \frac{1}{\sum_{n=1}^N \tau^m_{nd} \langle z^2_{nk} \rangle + \langle \alpha^m_k \rangle}$$

Taken together this means that we can update $q(\hat{w}^m_{dk}, s^m_{dk})$ using:

$$q(\hat{w}^m_{dk} | s^m_{dk}) \times q(s^m_{dk}) = \mathcal{N}\left(\hat{w}^m_{dk} \,|\, s^m_{dk} \mu_{w^m_{dk}}, s^m_{dk} \sigma^2_{w^m_{dk}} + (1 - s^m_{dk})/\alpha^m_k\right) \quad \times \quad (\lambda^m_{dk})^{s^m_{dk}} (1 - \lambda^m_{dk})^{1 - s_{dk}}$$

2.1.4 Lower bound

Likelihood term

$$\log p(y^m_{nd} | Z, W) = y^m_{nd} \sum_{k=1}^K \langle s^m_{dk} \hat{w}^m_{dk} \rangle \langle z_{nk} \rangle - \log\left(1 + \exp\left(\sum_{k=1}^K \langle s^m_{dk} \hat{w}^m_{dk} \rangle \langle z_{nk} \rangle\right)\right)$$

References

[1] C. M. Bishop. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

[2] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[3] M. Seeger and G. Bouchard. Fast variational bayesian inference for non-conjugate matrix factorization models. In *Artificial Intelligence and Statistics*, pages 1012–1018, 2012.