0.1 Hard sparsity prior: spike and Slab

0.1.1 The model

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:

$$y_{nd} \approx \mathcal{N}\left(y_{nd}^m \,|\, \mathbf{w}_{d,:}^m \mathbf{z}_{n,:}, 1/\tau_d^m\right) \tag{1}$$

$$w_{dk}^m \approx \theta \mathcal{N}\left(w_{dk} \,|\, 0, 1/\alpha_k^m\right) + (1-\theta)\delta_0(w_{dk}^m) \tag{2}$$

$$z_{nk} \approx \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \tag{3}$$

0.1.2 Efficient variational inference

The presence of the Dirac delta mass function makes the application of variational inference troublesome. However, there exists a simple reparameterization of the spike and slab prior that is more amenable to approximate inference.
Assume a Gaussian random variable $w \approx \mathcal{N}\left(\hat{w} \,|\, 0, \sigma^2\right)$ and a Bernoulli random variable $s \approx \pi^s(1-\pi)^{1-s}$. The product $s\hat{w}$ forms a new random variable distributed according to the original spike and slab prior. Thus, we can reparametrise $w = s\hat{w}$ and assign the prior distributions on s and $\hat{w}$. Thus, the reparametrised spike and slab takes the form:

$$p(\hat{w}, s) = \mathcal{N}\left(\hat{w} \,|\, 0, \sigma^2\right) \pi^s (1-\pi)^{1-s}$$

A simple approach for variational inference is to consider the mean-field approximation:

$$q(\hat{w}, s) = q(\hat{w})q(s)$$

which leads to easy updates. However, each pair of variables $\{s\hat{w}\}$ is strongly correlated since their product is the underlying variable that interacts with the data. Therefore, considering a unimodal distribution with the variables $\hat{w}$ and s being independent leads toa very inefficient inference. Instead, we introduce a paired mean field approximation $q(w, s)$ which approximates better the factorial nature of the true posterior.

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

Joint probability density function:

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

Full variational distribution:

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

0.1.3 Derivation of update equations

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

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{i\neq j}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const.} \tag{4}$$

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

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

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

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

Variational distribution:

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

where

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

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

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 $q(s_{dk}^m)$:

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

where

$$\lambda_{dk}^m = \langle \log \frac{\theta}{1-\theta} \rangle + 0.5 \log \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle} - 0.5 \log \Big(\sum_{n=1}^N \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle} \Big)$$
$$+ \frac{\langle \tau_d^m \rangle}{2} \frac{\Big(\sum_{n=1}^N y_{nd}^m \langle z_{nk} \rangle - \sum_{j \neq k} \langle s_{dj}^m \hat{w}_{dj}^m \rangle \sum_{n=1}^N \langle z_{nk} \rangle \langle z_{nj} \rangle \Big)^2}{\sum_{n=1}^N \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle}}$$

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

$$q(\hat{w}_{dk}^m | s_{dk}^m = 0) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right)$$
$$q(\hat{w}_{dk}^m | s_{dk}^m = 1) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, \mu_{w_{dk}^m}, \sigma_{w_{dk}^m}^2\right)$$

where

$$\mu_{w_{dk}^m} = \frac{\sum_{n=1}^N y_{nd}^m \langle z_{nk} \rangle - \sum_{j \neq k} \langle s_{dj}^m \hat{w}_{dj}^m \rangle \sum_{n=1}^N \langle z_{nk} \rangle \langle z_{nj} \rangle}{\sum_{n=1}^N \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle}}$$
$$\sigma_{w_{dk}^m} = \frac{\langle \tau_d^m \rangle^{-1}}{\sum_{n=1}^N \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle}}$$

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

$$q(\hat{w}_{dk}^m | s_{dk}^m) \times q(s_{dk}^m) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, s_{dk}^m \mu_{w_{dk}^m}, s_{dk}^m \sigma_{w_{dk}^m}^2 + (1 - s_{dk}^m)/\alpha_k^m\right) \quad \times \quad (\lambda_{dk}^m)^{s_{dk}^m} (1 - \lambda_{dk}^m)^{1 - s_{dk}}$$

ARD precision (alpha)

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const.} \tag{5}$$

Term from the prior $\log p(\alpha_k^m)$:

$$(a_0^\alpha - 1) \log(\alpha_k^m) - b_0^\alpha \alpha_k^m + \text{const.}$$

Term from the prior $\log p(\mathbf{w}_{:k}^m) = \sum_{d=1}^{D_m} \log p(s_{dk}^m, \hat{w}_{dk}^m)$:

$$\frac{D_m}{2} \log(\alpha_k^m) - \frac{\alpha_k^m}{2} \sum_{d=1}^{D} \langle \hat{w}_{dk}^2 \rangle + \sum_{d=1}^{D_m} \{ \langle s_{dk}^m \rangle \log \theta_0 + (1 - \langle s_{dk}^m \rangle) \log(1 - \theta_0) \} + \text{const.}$$

Writing everything together:

$$\Big(a_0^\alpha + \frac{D_m}{2} - 1 \Big) \log \alpha_k^m - \Big(b_0^\alpha + \frac{1}{2} \sum_{d=1}^{D_m} \langle (\hat{w}_{d,k}^m)^2 \rangle \Big) \alpha_k^m + \text{const.}$$

Variational distribution:

$$q(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}(\alpha_k^m | \hat{a}_{mk}^\alpha, \hat{b}_{mk}^\alpha)$$

where

$$\hat{a}_{mk}^\alpha = a_0^\alpha + \frac{D_m}{2}$$
$$\hat{b}_{mk}^\alpha = b_0^\alpha + \frac{\sum_{d=1}^{D_m} \langle (\hat{w}_{d,k}^m)^2 \rangle}{2}$$

Noise precision (tau)

Term from the prior $p(\tau_d^m)$:

$$(a_0^\tau - 1)\log\tau_d^m - b_0^\tau\tau_d^m + \text{const.}$$

Term from the likelihood $\mathcal{N}\left(y_{n,d}^m \mid \sum_{k=1}^K \hat{w}_{dk}^m s_{dk}^m z_{nk}, \tau_d^m\right)$:

$$\frac{N}{2}\log\tau_d^m - \frac{\tau_d^m}{2}\sum_{n=1}^N \langle (y_{nd}^m - \sum_k^K w_{dk}^m s_{dk}^m z_{nk})^2 \rangle + \text{const.}$$

Rewriting everything together:

$$\left(a_0^\tau - 1 + \frac{N}{2}\right)\log\tau_d^m - \left(b_0 + \frac{1}{2}\sum_{n=1}^N \langle (y_{nd}^m - \sum_k^K \hat{w}_{dk}^m s_{dk}^m z_{nk})^2 \rangle\right)\tau_d^m$$

Variational distribution:

$$q(\boldsymbol{\tau}) = \prod_{m=1}^M \prod_{d=1}^{D_m} q(\tau_d^m) = \prod_{m=1}^M \prod_{d=1}^{D_m} \mathcal{G}(\tau_d^m | \hat{a}_{md}^\tau, \hat{b}_{md}^\tau)$$

where

$$\hat{a}_{md}^\tau = a_0^\tau + \frac{N}{2}$$

$$\hat{b}_{md}^\tau = b_0^\tau + \frac{1}{2}\sum_{n=1}^N \langle (y_{nd}^m - \sum_k^K \hat{w}_{dk}^m s_{dk}^m z_{n,k})^2 \rangle$$

Spike and Slab sparsity parameter θ

Unless a given factor is specifically annotated in a given view, the sparisty parameter θ_k^m of the Spike and Slab prior on $w_{k,d}^m, \forall d$ is given a Beta prior: $P(\theta_k^m) = \text{Beta}(a_0, b_0)$. The postrerior $q(\theta_k^m)$ is Beta distributed and the update of its parameters a_k^m and b_k^m are given below:

$$a_k^m = \sum_d \langle S_{k,d}^m \rangle + a_0$$

$$b_k^m = b_0 - \sum_d \langle S_{k,d}^m \rangle + D_m$$

0.1.4 Lower bound

Likelihood term

Vector form:

$$-\sum_{m=1}^M \frac{ND_m}{2}\log(2\pi) + \frac{N}{2}\sum_{m=1}^M \sum_{d=1}^{D_m} \log(\tau_d^m) - \frac{1}{2}\sum_{m=1}^M \sum_{d=1}^{D_m} \left(\mathbf{y}_d^m - \sum_{k=1}^K \langle s_{dk}^m \hat{w}_{dk}^m \rangle \langle \mathbf{z}_k \rangle\right)^T \left(\tau_d^m \mathbf{I}\right)\left(\mathbf{y}_d^m - \sum_{k=1}^K \langle s_{dk}^m \hat{w}_{dk}^m \rangle \langle \mathbf{z}_k \rangle\right)$$

Scalar form:

$$-\sum_{m=1}^M \frac{ND_m}{2}\log(2\pi) + \frac{N}{2}\sum_{m=1}^M \sum_{d=1}^{D_m} \log(\langle\tau_d^m\rangle) - \sum_{m=1}^M \sum_{d=1}^{D_m} \frac{\langle\tau_d^m\rangle}{2}\sum_{n=1}^N \left(y_{nd}^m - \sum_{k=1}^K \langle s_{dk}^m \hat{w}_{dk}^m \rangle \langle z_{nk} \rangle\right)^2$$

Extending terms and rearranging:

W and S terms

$p(\hat{\mathbf{W}}, \mathbf{S})$:

$$-\sum_{m=1}^{M} \frac{KD_m}{2} \log(2\pi) + \sum_{m=1}^{M} \frac{D_m}{2} \sum_{k=1}^{K} \log(\alpha_k^m) - \sum_{m=1}^{M} \frac{\alpha_k^m}{2} \sum_{d=1}^{D_m} \sum_{k=1}^{K} \langle (\hat{w}_{dk}^m)^2 \rangle$$
$$+ \langle \log(\theta) \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \sum_{k=1}^{K} \langle s_{dk}^m \rangle + \langle \log(1-\theta) \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \sum_{k=1}^{K} (1 - \langle s_{dk}^m \rangle)$$

$q(\hat{\mathbf{W}}, \mathbf{S})$:

$$-\sum_{m=1}^{M} \frac{KD_m}{2} \log(2\pi) + \frac{1}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \sum_{k=1}^{K} \log(\langle s_{dk}^m \rangle \sigma_{w_{dk}^m}^2 + (1 - \langle s_{dk}^m \rangle)/\alpha_k^m)$$
$$+ \sum_{m=1}^{M} \sum_{d=1}^{D_m} \sum_{k=1}^{K} (1 - \langle s_{dk}^m \rangle) \log(1 - \langle s_{dk}^m \rangle) - \langle s_{dk}^m \rangle \log \langle s_{dk}^m \rangle$$

Z term

$$\mathbb{E}[\log P(\mathbf{Z})] = -\frac{NK}{2} \log(2\pi) - \frac{1}{2} \sum_{n=1}^{N} \langle z_{nk}^2 \rangle$$
$$\mathbb{E}[\log q(\mathbf{Z})] = -\frac{NK}{2} (1 + \log(2\pi)) - \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \log(\sigma_{z_{nk}}^2)$$

alpha term

$$\mathbb{E}[\log p(\boldsymbol{\alpha})] = \sum_{m=1}^{M} \sum_{k=1}^{K} \Big(a_0^\alpha \log b_0^\alpha + (a_0^\alpha - 1) \langle \log \alpha_k \rangle - b_0^\alpha \langle \alpha_k \rangle - \log \Gamma(a_0^\alpha) \Big)$$
$$\mathbb{E}[\log q(\boldsymbol{\alpha})] = \sum_{m=1}^{M} \sum_{k=1}^{K} \Big(\hat{a}_k^\alpha \log \hat{b}_k^\alpha + (\hat{a}_k^\alpha - 1) \langle \log \alpha_k \rangle - \hat{b}_k^\alpha \langle \alpha_k \rangle - \log \Gamma(\hat{a}_k^\alpha) \Big)$$

tau

$$\mathbb{E}[\log P(\tau)] = \sum_{m=1}^{M} D_m a_0^\tau \log b_0^\tau + \sum_{m=1}^{M} \sum_{d=1}^{Dm} (a_0^\tau - 1) \langle \log \tau_d^m \rangle - \sum_{m=1}^{M} \sum_{d=1}^{Dm} b_0^\tau \langle \tau_d^m \rangle - \sum_{m=1}^{M} D_m \Gamma(a_0^\tau)$$

$$\mathbb{E}[\log Q(\boldsymbol{\tau})] = \sum_{m=1}^{M} \sum_{d=1}^{D_m} \Big(\hat{a}_{dm}^\tau \log \hat{b}_{dm}^\tau + (\hat{a}_{dm}^\tau - 1) \langle \log \tau_d^m \rangle - \hat{b}_{dm}^\tau \langle \tau_d^m \rangle - \log \Gamma(\hat{a}_{dm}^\tau) \Big)$$

Theta

$$\mathbb{E}\left[\log P(\theta)\right] = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{d=1}^{D_m} \big((a_0 - 1) \times \langle \log(\pi_{d,k}^m) \rangle + (b_0 - 1) \langle \log(1 - \pi_{d,k}^m) \rangle - \log(\mathrm{B}(a_0, b_0)) \big)$$
$$\mathbb{E}\left[\log Q(\theta)\right] = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{d=1}^{D_m} \big((a_{k,d}^m - 1) \times \langle \log(\pi_{d,k}^m) \rangle + (b_{k,d}^m - 1) \langle \log(1 - \pi_{d,k}^m) \rangle - \log(\mathrm{B}(a_{k,d}^m, b_{k,d}^m)) \big)$$

where the expectations are calculated as follows:

Expectations

The expectations are calculated as follows:

$$\langle s_{dk}^m \hat{w}_{dk}^m \rangle = \lambda_{dk}^m \mu_{w_{dk}^m}$$

$$\langle s_{dk}^m \hat{w}_{dk}^{m2} \rangle = \lambda_{dk}^m (\mu_{w_{dk}^m}^2 + \sigma_{w_{dk}^m}^2)$$

$$\langle \hat{w}_{dk}^{m2} \rangle = \lambda_{dk}^m (\mu_{w_{dk}^m}^2 + \sigma_{w_{dk}^m}^2) + (1 - \lambda_{dk}^m)/\alpha_k^m$$

$$\langle z_{nk} \rangle = \mu_{z_{nk}}$$

$$\langle z_{nk}^2 \rangle = \mu_{z_{nk}}^2 + \sigma_{z_{nk}}^2$$

$$\langle \tau_d^m \rangle = \tilde{a}_{md}^\tau / \tilde{b}_{md}^\tau$$

$$\langle \log \tau_d^m \rangle = \psi(\tilde{a}_{md}^\tau) - \log \tilde{b}_{md}^\tau$$

$$\langle (y_{nd}^m - \sum_k^K w_{dk}^m s_{dk}^m z_{nk})^2 \rangle =$$

$$(y_{nd}^m)^2 - 2y_{nd}^m \sum_k^K \langle w_{dk}^m s_{dk}^m \rangle \langle z_{nk} \rangle + \sum_k^K \sum_j^K \langle w_{dk}^m s_{dk}^m \rangle \langle z_{nk} \rangle \langle w_{dj}^m s_{dj}^m \rangle \langle z_{nj} \rangle =$$

$$(y_{nd}^m)^2 - 2y_{nd}^m \sum_k^K \langle w_{dk}^m s_{dk}^m \rangle \langle z_{nk} \rangle + \sum_k^K \langle (w_{dk}^m s_{dk}^m)^2 \rangle \langle z_{nk}^2 \rangle + 2 \sum_{j>k}^K \langle w_{dk}^m s_{dk}^m \rangle \langle w_{dj}^m s_{dj}^m \rangle \langle z_{nk} \rangle \langle z_{nj} \rangle$$

0.2 Add cluster specific prior on the latent variables

The samples are divided into C clusters. The index c refers to one of these clusters, $c \in [[1, C]]$. When a sample n belongs to cluster c we write $n \in c$ (meaning that the notation c is used both to designate the cluster and its index)

0.2.1 Prior architecture on Z

A cluster specific prior is used for every latent variable:

$$z_{n,k} \sim \mathcal{N}(\mu_{k,c}, \sigma_{k,c}), \forall n \in c, \tag{6}$$

where $\mu_{k,c}$ is given a normal prior:

$$\mu_{k,c} \sim \mathcal{N}(\mu_0, \sigma_0), \forall k, c \tag{7}$$

0.2.2 New update for Z

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

$$\mu_{z_{nk}} | n \in c = \sigma_{z_{nk}}^2 \left[\frac{\langle \mu_{k,c} \rangle}{1} + \sum_{m=1}^M \sum_{d=1}^{D_m} \langle \tau_d^m \rangle \langle s_{dk}^m \hat{w}_{dk}^m \rangle \Big(y_{nd}^m - \sum_{j \neq k} \langle s_{dj}^m \hat{w}_{dj}^m \rangle \langle z_{nj} \rangle \Big) \right]$$

0.2.3 Update for $\mu_{k,c}$

$$\sigma^2_{\mu_{k,c}} = \left(\sum_{n \in c} \frac{1}{\sigma^2_{z_{n,k}}} + \frac{1}{\sigma_0^2} \right)^{-1}$$

$$\mu_{\mu_{k,c}} = \sigma^2_{\mu_{k,c}} \times \left(\sum_{n \in c} \frac{\langle z_{n,k} \rangle}{\sigma^2_{z_{n,k}}} + \frac{\mu_0}{\sigma_0^2} \right)$$

0.2.4 ELBO tern for μ

Cross entropy term

$$\mathbb{E}\left[\log(P(\mu_{k,c}))\right] = \sum_c \sum_k \log\left(\frac{1}{\sigma_0 (2\pi)^{1/2}} \right) - \frac{1}{2\sigma_0^2} \left(\langle \mu_{k,c}^2 \rangle - 2\langle \mu_{k,c} \rangle \mu_0 + \mu_0^2 \right)$$

Entropy of Q:

$$-\mathbb{E}\left[\log(Q(\mu_{k,c}))\right] = \sum_c \sum_k 1/2 \log(2\sigma^2_{\mu_{k,c}} \pi e)$$