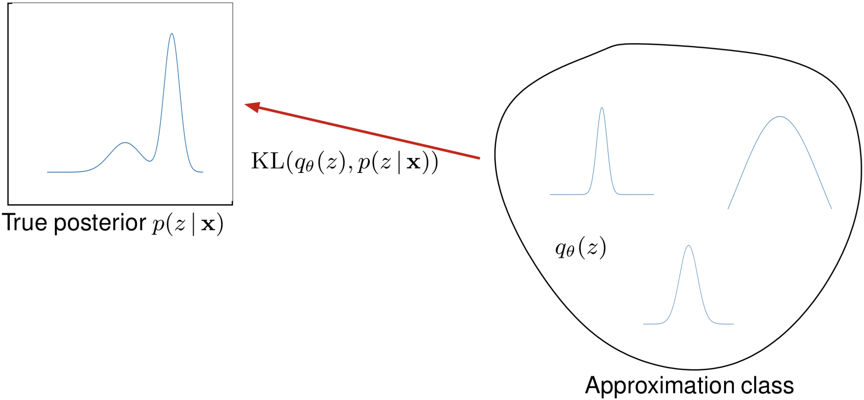
Variational inference

Partly based on material developed together with Helge Langseth

Andrés Masegosa and Thomas Dyhre Nielsen

Introduction

We will approximate the true posterior distribution $p(z \mid \mathbf{x})$ with a variational distribution belonging to a tractable family of distributions.



Task: Fit the variational parameters θ so that the 'distance' $\mathrm{KL}(q_\theta(z), p(z \mid \mathbf{x}))$ is minimized:

$$\hat{q}(z) = \arg\min_\theta \mathrm{KL}(q_\theta(z), p(z \mid \mathbf{x})) = \arg\min_\theta \int_{\mathbf{z}} q(z) \, \log\left(\frac{q(z)}{p(z \mid \mathbf{x})}\right) \mathrm{d}z$$

ELBO: Evidence Lower-BOund

We can rearrange the KL divergence as follows:

$$
\begin{aligned}
\mathrm{KL}\left(q(\mathbf{z}) || p(\mathbf{z} \,|\, \mathbf{x})\right) &= \mathbb{E}_q\left[\log \frac{q(\mathbf{z})}{p(\mathbf{z} \,|\, \mathbf{x})}\right] \\
&= \mathbb{E}_q\left[\log \frac{q(\mathbf{z}) \cdot p(\mathbf{x})}{p(\mathbf{z} \,|\, \mathbf{x}) \cdot p(\mathbf{x})}\right] \\
&= \log p(\mathbf{x}) - -\mathbb{E}_q\left[\log \tfrac{q(\mathbf{z})}{p(\mathbf{z},\mathbf{x})}\right] = \log p(\mathbf{x}) - \mathcal{L}\left(q\right)
\end{aligned}
$$

where $\mathcal{L}\left(q\right) = -\mathbb{E}_q\left[\log \frac{q(\mathbf{z})}{p(\mathbf{z},\mathbf{x})}\right]$ is the so-called Evidence Lower Bound (ELBO)

ELBO: Evidence Lower-BOund

We can rearrange the KL divergence as follows:

$$
\begin{aligned}
\mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})\right) &= \mathbb{E}_q\left[\log\frac{q(\mathbf{z})}{p(\mathbf{z}\,|\,\mathbf{x})}\right] \\
&= \mathbb{E}_q\left[\log\frac{q(\mathbf{z})\cdot p(\mathbf{x})}{p(\mathbf{z}\,|\,\mathbf{x})\cdot p(\mathbf{x})}\right] \\
&= \log p(\mathbf{x}) - -\mathbb{E}_q\left[\log\frac{q(\mathbf{z})}{p(\mathbf{z},\mathbf{x})}\right] = \log p(\mathbf{x}) - \mathcal{L}\left(q\right)
\end{aligned}
$$

where $\mathcal{L}\left(q\right) = -\mathbb{E}_q\left[\log\frac{q(\mathbf{z})}{p(\mathbf{z},\mathbf{x})}\right]$ is the so-called Evidence Lower Bound (ELBO)

VI focuses on the ELBO:

$$
\log p(\mathbf{x}) = \mathcal{L}\left(q\right) + \mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})\right)
$$

Since $\log p(\mathbf{x})$ is constant wrt. q and $\mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})\right) \geq 0$ it follows:

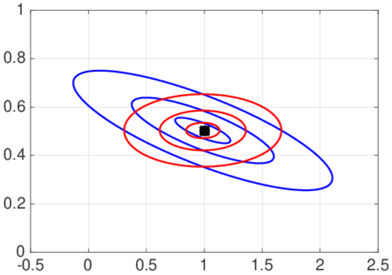
- We can minimize $\mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})\right)$ by maximizing $\mathcal{L}\left(q\right)$
- This is **computationally simpler** because it uses $p(\mathbf{z},\mathbf{x})$ instead of $p(\mathbf{z}\,|\,\mathbf{x})$.
- $\mathcal{L}\left(q\right)$ is a *lower bound* of the marginal data log likelihood $\log p(\mathbf{x})$.

$\rightsquigarrow$ During inference, we will look for $\hat{q}(\mathbf{z}) = \arg\max_{q\in\mathcal{Q}}\mathcal{L}\left(q\right)$.

The mean field assumption

We will often use the mean field assumption, which states that $\mathcal{Q}$ consists of all distributions that *factorizes* according to the equation

$$q(\mathbf{z}) = \prod_i q_i\left(z_i\right)$$



Note! This may seem like a very restricted set. However, we can choose any $q(\mathbf{z}) \in \mathcal{Q}$, and this is how the magic ($\sim$ "absorbing information from $\mathbf{x}$") happens.

Wrapping it all up: The VB algorithm under MF

Algorithm:

- We have observed $\mathbf{X} = \mathbf{x}$, and have access to the full joint $p(\mathbf{z}, \mathbf{x})$.
- We posit a *variational family* of distributions $q_j(\cdot \mid \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.
- The posterior approximation is assumed to factorize according to the mean-field assumption, and we use the $\mathrm{KL}\left(q(\mathbf{z}) || p(\mathbf{z} \mid \mathbf{x})\right)$ as our objective.

Algorithm:

Repeat until negligible improvement in terms of $\mathcal{L}\left(q\right)$:

- For each j:
    - Somehow choose $\boldsymbol{\lambda}_j$ to maximize $\mathcal{L}\left(q\right)$, typically based on $\{\boldsymbol{\lambda}_i\}_{i \neq j}$.
- Calculate the new $\mathcal{L}\left(q\right)$.

Solving the VB optimization

Solving the VB equation

We will maximize $\mathcal{L}(q) = \mathbb{E}_q\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right]$ under the assumption that $q(\cdot)$ factorizes.
Let us pick one j, utilize that $q(\mathbf{z}) = q_j(z_j) \cdot q_{\neg j}(\mathbf{z}_{\neg j})$, and assume $q_{\neg j}(\cdot)$ is kept fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] \\ &= \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right]\end{aligned}$$

Solving the VB equation

We will maximize $\mathcal{L}(q) = \mathbb{E}_q\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right]$ under the assumption that $q(\cdot)$ factorizes.
Let us pick one j, utilize that $q(\mathbf{z}) = q_j(z_j) \cdot q_{\neg j}(\mathbf{z}_{\neg j})$, and assume $q_{\neg j}(\cdot)$ is kept fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] \\ &= \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right]\end{aligned}$$

For the term $\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right]$ we simply define $f_j(z_j)$ so that

$$\log f_j(z_j) = \mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right]$$

Solving the VB equation

We will maximize $\mathcal{L}(q) = \mathbb{E}_q\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right]$ under the assumption that $q(\cdot)$ factorizes.
Let us pick one j, utilize that $q(\mathbf{z}) = q_j(z_j) \cdot q_{\neg j}(\mathbf{z}_{\neg j})$, and assume $q_{\neg j}(\cdot)$ is kept fixed.

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j} \log f_j(z_j) - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right]
\end{aligned}$$

Solving the VB equation

We will maximize $\mathcal{L}(q) = \mathbb{E}_q\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right]$ under the assumption that $q(\cdot)$ factorizes.
Let us pick one j, utilize that $q(\mathbf{z}) = q_j(z_j)\cdot q_{\neg j}(\mathbf{z}_{\neg j})$, and assume $q_{\neg j}(\cdot)$ is kept fixed.

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j}\log f_j(z_j) - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right]
\end{aligned}$$

For the other term, notice that $\log q(\mathbf{z}) = \log q_j(z_j) + \log q_{\neg j}(\mathbf{z}_{\neg j})$ Therefore

$$\begin{aligned}
\mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right] &= \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q_j(z_j) + \log q_{\neg j}(\mathbf{z}_{\neg j})\right] \\
&= \mathbb{E}_{q_j}\left[\log q_j(z_j)\right] + \mathbb{E}_{q_{\neg j}}\left[\log q_{\neg j}(\mathbf{z}_{\neg j})\right] \\
&= \mathbb{E}_{q_j}\left[\log q_j(z_j)\right] + c,
\end{aligned}$$

because $\mathbb{E}_{q_{\neg j}}\left[\log q_{\neg j}(\mathbf{z}_{\neg j})\right]$ is constant wrt. $q_j(\cdot)$.

We will maximize $\mathcal{L}(q) = \mathbb{E}_q\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right]$ under the assumption that $q(\cdot)$ factorizes.
Let us pick one j, utilize that $q(\mathbf{z}) = q_j(z_j) \cdot q_{\neg j}(\mathbf{z}_{\neg j})$, and assume $q_{\neg j}(\cdot)$ is kept fixed.

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j} \log f_j(z_j) - \mathbb{E}_{q_j}\left[\log q_j(z_j)\right] + c \\
&= -\mathrm{KL}\left(q_j(z_j) || f_j(z_j)\right) + c
\end{aligned}$$

We will maximize $\mathcal{L}(q) = \mathbb{E}_q\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right]$ under the assumption that $q(\cdot)$ factorizes.
Let us pick one j, utilize that $q(\mathbf{z}) = q_j(z_j) \cdot q_{\neg j}(\mathbf{z}_{\neg j})$, and assume $q_{\neg j}(\cdot)$ is kept fixed.

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}_{q_j}\mathbb{E}_{q_{\neg j}}\left[\log q(\mathbf{z})\right] \\
&= \mathbb{E}_{q_j} \log f_j(z_j) - \mathbb{E}_{q_j}\left[\log q_j(z_j)\right] + c \\
&= -\mathrm{KL}\left(q_j(z_j) || f_j(z_j)\right) + c
\end{aligned}$$

We get the following result:

The ELBO is maximized wrt. q_j by choosing

$$q_j(z_j) = \frac{1}{Z} \exp\left(\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z},\mathbf{x})\right]\right)$$

. . . and made the following assumptions to get there:

- Mean field: $q(\mathbf{z}) = \prod_i q_i(z_i)$, and specifically $q(\mathbf{z}) = q_j(z_j) \cdot q_{\neg j}(\mathbf{z}_{\neg j})$.
- We optimize wrt. $q_j(\cdot)$, while keeping $q_{\neg j}(\cdot)$ fixed – i.e., we do coordinate ascent in probability distribution space.

Setup

- We have observed $\mathbf{X} = \mathbf{x}$, and can calculate the full joint $p(\mathbf{z}, \mathbf{x})$.
- The posterior approximation is assumed to factorize according to the mean-field assumption, and we use the $\mathrm{KL}\left(q(\mathbf{z}) || p(\mathbf{z} \mid \mathbf{x})\right)$ as our objective.
- We posit a *variational family* of distributions $q_j(z_j \mid \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.
- The optimal $\boldsymbol{\lambda}_j$ **will** depend on $\mathbf{x}$ – in fact $\boldsymbol{\lambda}_j$ encodes all the information about the other variables in the domain that Z_j is “aware of”.

VB w/ MF: algorithm

Setup

- We have observed $\mathbf{X} = \mathbf{x}$, and can calculate the full joint $p(\mathbf{z}, \mathbf{x})$.
- The posterior approximation is assumed to factorize according to the mean-field assumption, and we use the $\mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z} \mid \mathbf{x})\right)$ as our objective.
- We posit a *variational family* of distributions $q_j(z_j \mid \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.
- The optimal $\boldsymbol{\lambda}_j$ **will** depend on $\mathbf{x}$ – in fact $\boldsymbol{\lambda}_j$ encodes all the information about the other variables in the domain that Z_j is "aware of".

Algorithm

Repeat until negligible improvement in terms of $\mathcal{L}\left(q\right)$:

- For each j:
  - Calculate $\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]$ using current estimates for $q_i(\cdot \mid \boldsymbol{\lambda}_i)$, $i \neq j$.
  - Choose $\boldsymbol{\lambda}_j$ so that $q_j(z_j \mid \boldsymbol{\lambda}_j) \propto \exp\left(\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right)$.
- Calculate the new $\mathcal{L}\left(q\right)$.

Algorithm

Repeat until negligible improvement in terms of $\mathcal{L}(q)$:

- For each j:
  - Calculate $\mathbb{E}_{q_{\neg j}}[\log p(\mathbf{z}, \mathbf{x})]$ using current estimates for $q_i(\cdot \mid \boldsymbol{\lambda}_i)$, $i \neq j$.
  - Choose $\boldsymbol{\lambda}_j$ so that $q_j(z_j \mid \boldsymbol{\lambda}_j) \propto \exp\left(\mathbb{E}_{q_{\neg j}}[\log p(\mathbf{z}, \mathbf{x})]\right)$.
- Calculate the new $\mathcal{L}(q)$.

Calculating $q_j(z_j \mid \boldsymbol{\lambda}_j)$

The update-rule can equivalently be expressed as

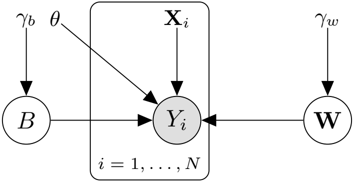
$$\begin{aligned}\log q_j(z_j \mid \boldsymbol{\lambda}_j) &= \mathbb{E}_{q_{\neg j}}[\ln p(\mathbf{z}, \mathbf{x})] + c. \\ &= \sum_{x \in \mathrm{mb}(z_j)} \mathbb{E}_{q_{\neg j}} \log p(x \mid \mathrm{pa}(x)) + \sum_{z \in \mathrm{mb}(z_j)} \mathbb{E}_{q_{\neg j}} \log p(z \mid \mathrm{pa}(z)) + c'.\end{aligned}$$

Note!

- We only need to consider terms that share a factor with z_j – all other terms get absorbed into the constant c.
- $\rightsquigarrow$ need only reason about variables in the Markov blanket of Z_j – just as for Gibbs sampling!

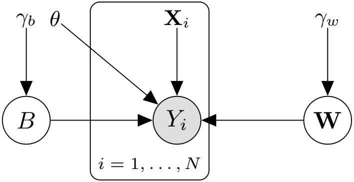
Bayesian linear regression

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i \mid \{\mathbf{w}, \mathbf{x}_i, \theta\} \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1}\mathbf{I}_{M\times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The Bayesian linear regression model

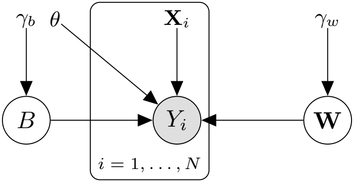


- Num. of data dim: M
- Num. of data inst: N
- $Y_i \mid \{\mathbf{w}, \mathbf{x}_i, \theta\} \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1}\mathbf{I}_{M\times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

$$p(\cdot \mid \mathbf{w}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}, \theta) p(\mathbf{w} \mid \gamma_w) p(b \mid \gamma_b)$$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i \mid \{\mathbf{w}, \mathbf{x}_i, \theta\} \sim \mathcal{N}(\mathbf{w}^\intercal \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

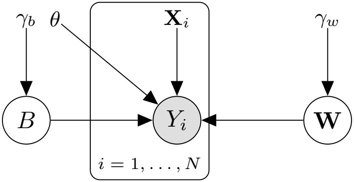
The probability model

$$p(\cdot \mid \mathbf{w}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}, \theta) p(\mathbf{w} \mid \gamma_w) p(b \mid \gamma_b)$$

...after taking the log

$$\log p(\cdot \mid \mathbf{w}, b, \theta, \gamma_w, \gamma_b) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w}, \theta) + \log p(\mathbf{w} \mid \gamma_w) + \log p(b \mid \gamma_b)$$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i \,|\, \{\mathbf{w}, \mathbf{x}_i, \theta\} \sim \mathcal{N}(\mathbf{w}^\intercal \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M\times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

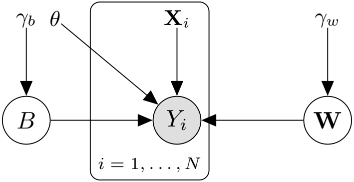
The probability model

$$p(\cdot \,|\, \mathbf{w}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^{N} p(y_i \,|\, \mathbf{x}_i, \mathbf{w}, \theta) p(\mathbf{w} \,|\, \gamma_w) p(b \,|\, \gamma_b)$$

...after taking the log

$$\begin{aligned}\log p(\cdot \,|\, \mathbf{w}, b, \theta, \gamma_w, \gamma_b) &= \sum_{i=1}^{N} \log p(y_i \,|\, \mathbf{x}_i, \mathbf{w}, \theta) + \log p(\mathbf{w} \,|\, \gamma_w) + \log p(b \,|\, \gamma_b) \\ &= \sum_{i=1}^{N} \log p(y_i \,|\, \mathbf{x}_i, \mathbf{w}, \theta) + \sum_{j=1}^{M} \log p(w_j \,|\, \gamma_w) + \log p(b \,|\, \gamma_b)\end{aligned}$$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i \mid \{\mathbf{w}, \mathbf{x}_i, \theta\} \sim \mathcal{N}(\mathbf{w}^\intercal \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

$$p(\cdot \mid \mathbf{w}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}, \theta) p(\mathbf{w} \mid \gamma_w) p(b \mid \gamma_b)$$

The variational model (full mean field)

$$q(\cdot) = q(b \mid \cdot) \prod_{i=1}^{M} q(w_i \mid \cdot)$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q \neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q \neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c$$

Recall

$$\log p(\cdot \mid \mathbf{w}, b, \theta, \gamma) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w}, \theta) + \sum_{j=1}^{M} \log p(w_j \mid \gamma) + \log p(b \mid \gamma_b)$$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q_{\neg w_j}} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c$$

Recall

$$\log p(\cdot \mid \mathbf{w}, b, \theta, \gamma) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w}, \theta) + \sum_{j=1}^{M} \log p(w_j \mid \gamma) + \log p(b \mid \gamma_b)$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

Recall

$$\log p(\cdot \mid \mathbf{w}, b, \theta, \gamma) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w}, \theta) + \sum_{j=1}^{M} \log p(w_j \mid \gamma) + \log p(b \mid \gamma_b)$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

The normal distribution

$$\log p(y_i \mid \mathbf{x}_i, \mathbf{w}, b, \theta) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\theta) - \frac{\theta}{2}(y_i - (\mathbf{w}^\intercal \mathbf{x}_i + b))^2$$

$$\log p(w_j \mid \gamma_w) = \log \mathcal{N}(w_j \mid 0, \gamma_w^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\gamma_w) - \frac{\gamma_w}{2} w_j^2$$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

The normal distribution

$$\log p(y_i \mid \mathbf{x}_i, \mathbf{w}, b, \theta) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\theta) - \frac{\theta}{2}(y_i - (\mathbf{w}^\intercal \mathbf{x}_i + b))^2$$

$$\log p(w_j \mid \gamma_w) = \log \mathcal{N}(w_j \mid 0, \gamma_w^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\gamma_w) - \frac{\gamma_w}{2} w_j^2$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \underset{q\neg w_j}{\mathbb{E}} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

The normal distribution

$$\log p(y_i \mid \mathbf{x}_i, \mathbf{w}, b, \theta) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\theta) - \frac{\theta}{2}(y_i - (\mathbf{w}^\intercal \mathbf{x}_i + b))^2$$

$$\log p(w_j \mid \gamma_w) = \log \mathcal{N}(w_j \mid 0, \gamma_w^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\gamma_w) - \frac{\gamma_w}{2} w_j^2$$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q_{\neg w_j}} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

Expanding the square

$$(y - (\mathbf{w}^\intercal \mathbf{x} + b))^2 = y^2 + \mathbf{x}^\intercal \mathbf{w} \mathbf{w}^\intercal \mathbf{x} + b^2 + 2\mathbf{w}^\intercal \mathbf{x} b - 2y\mathbf{w}^\intercal \mathbf{x} - 2yb$$

$$\mathbf{x}^\intercal \mathbf{w} \mathbf{w}^\intercal \mathbf{x} = x_j^2 w_j^2 + \sum_{h,k \neq j} x_k x_h w_k w_h + 2x_j w_j \sum_{k \neq j} x_k w_k$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

Expanding the square

$$(y - (\mathbf{w}^\intercal \mathbf{x} + b))^2 = y^2 + \mathbf{x}^\intercal \mathbf{w} \mathbf{w}^\intercal \mathbf{x} + b^2 + 2\mathbf{w}^\intercal \mathbf{x} b - 2y\mathbf{w}^\intercal \mathbf{x} - 2yb$$

$$\mathbf{x}^\intercal \mathbf{w} \mathbf{w}^\intercal \mathbf{x} = x_j^2 w_j^2 + \sum_{h,k \neq j} x_k x_h w_k w_h + 2 x_j w_j \sum_{k \neq j} x_k w_k$$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \theta \sum_{i=1}^{N} (\frac{1}{2} x_{ij}^2 w_j^2 + w_j (\sum_{k \neq j} x_{ij} x_{ik} \, \mathbb{E}(W_k) + x_{ij} \, \mathbb{E}(B) - y x_{ij}) + c$$

Expanding the square

$$(y - (\mathbf{w}^\intercal \mathbf{x} + b))^2 = y^2 + \mathbf{x}^\intercal \mathbf{w} \mathbf{w}^\intercal \mathbf{x} + b^2 + 2\mathbf{w}^\intercal \mathbf{x} b - 2y\mathbf{w}^\intercal \mathbf{x} - 2yb$$

$$\mathbf{x}^\intercal \mathbf{w} \mathbf{w}^\intercal \mathbf{x} = x_j^2 w_j^2 + \sum_{h,k \neq j} x_k x_h w_k w_h + 2 x_j w_j \sum_{k \neq j} x_k w_k$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \theta \sum_{i=1}^{N} (\frac{1}{2} x_{ij}^2 w_j^2 + w_j (\sum_{k \neq j} x_{ij} x_{ik}\, \mathbb{E}(W_k) + x_{ij}\, \mathbb{E}(B) - y x_{ij}) + c$$

$$= -\frac{1}{2} (\gamma_w + \theta \sum_{i=1}^{N} (x_{ij}^2)) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij} (y_i - (\sum_{k \neq j} x_{ik}\, \mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{1}{2}(\gamma_w + \theta \sum_{i=1}^{N} (x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{1}{2}(\gamma_w + \theta \sum_{i=1}^{N}(x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij}(y_i - (\sum_{k \neq j} x_{ik}\, \mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

Recall the normal distribution

$$\log p(z \mid \mu, \tau) = \log \mathcal{N}(z \mid \mu, \tau^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{\tau}{2}(z-\mu)^2$$

$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{1}{2}\tau z^2 - \frac{\tau}{2}\mu^2 + z\tau\mu$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q\neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{1}{2}(\gamma_w + \theta \sum_{i=1}^{N}(x_{ij}^2)w_j^2 + w_j\theta \sum_{i=1}^{N} x_{ij}(y_i - (\sum_{k \neq j} x_{ik}\,\mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

Recall the normal distribution

$$\log p(z \mid \mu, \tau) = \log \mathcal{N}(z \mid \mu, \tau^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{\tau}{2}(z-\mu)^2$$

$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{1}{2}\tau z^2 - \frac{\tau}{2}\mu^2 + z\tau\mu$$

We choose the variational distribution so that

$$\log q(w_j) = \mathop{\mathbb{E}}_{q \neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(w_j \mid \gamma_w) + c$$

$$= -\frac{1}{2}(\gamma_w + \theta \sum_{i=1}^{N}(x_{ij}^2))w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij}(y_i - (\sum_{k \neq j} x_{ik}\, \mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

Thus, we see that $q(w_j)$ is normally distributed with

- precision $\tau \leftarrow (\gamma_w + \theta \sum_{i=1}^{N}(x_{ij}^2))$
- mean $\mu \leftarrow \tau^{-1}\theta \sum_{i=1}^{N} x_{ij}(y_i - (\sum_{k \neq j} x_{ik}\, \mathbb{E}(W_k) + \mathbb{E}(B)))$

Recall the normal distribution

$$\log p(z \mid \mu, \tau) = \log \mathcal{N}(z \mid \mu, \tau^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{\tau}{2}(z-\mu)^2$$

$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{1}{2}\tau z^2 - \frac{\tau}{2}\mu^2 + z\tau\mu$$

Exercise: Derive the updating rule for $q(b)$

Now it is your turn!

- Derive the updating rule for $q(b)$.
- Implement the updating rule in the notebook

  `students_lin_reg.ipynb`

It may be useful to recall the definition:

- Gauss: $\log p(x \,|\, \mu, 1/\gamma) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\gamma) - \frac{\gamma}{2}\,(x-\mu)^2$.

VB for Bayesian linear regression: updating $q(b)$

We choose the variational distribution so that

$$
\begin{aligned}
\log q(b) &= \mathop{\mathbb{E}}_{q_{\neg w_j}} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \mid \gamma_b) + c \\
&= \dots \\
&= -\frac{1}{2}(\gamma_b + \theta N) b^2 + b \left(\theta \sum_{i=1}^{N} (y_i - \mathbb{E}(\mathbf{W}^\intercal)\mathbf{x}_i) \right) + c
\end{aligned}
$$

We choose the variational distribution so that

$$\log q(b) = \mathop{\mathbb{E}}_{q \neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \mid \gamma_b) + c$$

$$= \dots$$

$$= -\frac{1}{2}(\gamma_b + \theta N) b^2 + b \left(\theta \sum_{i=1}^{N} (y_i - \mathbb{E}(\mathbf{W}^\intercal) \mathbf{x}_i) \right) + c$$

Recall the normal distribution

$$\log p(z \mid \mu, \tau) = \log \mathcal{N}(z \mid \mu, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2}(z - \mu)^2$$

$$= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \tau z^2 - \frac{\tau}{2} \mu^2 + z \tau \mu$$

VB for Bayesian linear regression: updating $q(b)$

We choose the variational distribution so that

$$\log q(b) = \mathop{\mathbb{E}}_{q \neg w_j} \log p(\cdot \,|\, \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \,|\, \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \,|\, \gamma_b) + c$$

$$= \dots$$

$$= -\frac{1}{2}(\gamma_b + \theta N) b^2 + b \left(\theta \sum_{i=1}^{N} (y_i - \mathbb{E}(\mathbf{W}^\intercal)\mathbf{x}_i) \right) + c$$

Thus, we get that $q(w_j)$ is normally distributed with

- precision $\tau \leftarrow (\gamma_b + \theta N)$
- mean $\mu \leftarrow \tau^{-1} \theta \sum_{i=1}^{N} (y_i - \mathbb{E}(\mathbf{W}^\intercal)\mathbf{x}_i)$

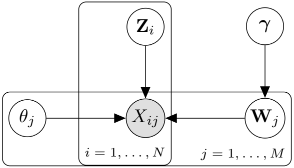
Recall the normal distribution

$$\log p(z \,|\, \mu, \tau) = \log \mathcal{N}(z \,|\, \mu, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2}(z - \mu)^2$$

$$= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \tau z^2 - \frac{\tau}{2} \mu^2 + z \tau \mu$$

Factor analysis

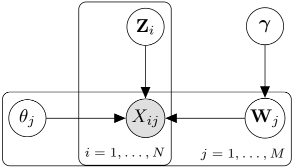
The factor analysis model



- $X_{ij} \mid \{\mathbf{w}_j, \mathbf{z}_i, \theta_j\} \sim \mathcal{N}(\mathbf{w}_j^\intercal \mathbf{z}_i, 1/\theta_j)$
- $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D\times D})$
- $\mathbf{W}_j \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}_{D\times D})$
- $\theta_j \sim$ Gamma$(\theta_\theta, \theta_\theta)$
- $\gamma \sim$ Gamma$(\alpha_\gamma, \beta_\gamma)$

- Num. of latent dim: D
- Num. of data dim: M
- Num. of data inst: N

The factor analysis model



- $X_{ij} \mid \{\mathbf{w}_j, \mathbf{z}_i, \theta_j\} \sim \mathcal{N}(\mathbf{w}_j^{\intercal}\mathbf{z}_i, 1/\theta_j)$
- $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D\times D})$
- $\mathbf{W}_j \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}_{D\times D})$
- $\theta_j \sim$ Gamma$(\theta_\theta, \theta_\theta)$
- $\gamma \sim$ Gamma$(\alpha_\gamma, \beta_\gamma)$

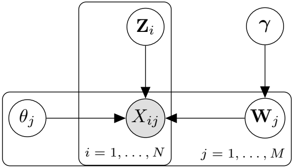
- Num. of latent dim: D
- Num. of data dim: M
- Num. of data inst: N

The probability model

$$p(\cdot) = p(\gamma)\left[\prod_{i=1}^{N} p(\mathbf{z}_i)\right]\left[\prod_{j=1}^{M} p(\mathbf{w}_j \mid \gamma)p(\theta_j)\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{M} p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)\right]$$

VB for the factor analysis model

The factor analysis model



- $X_{ij} \mid \{\mathbf{w}_j, \mathbf{z}_i, \theta_j\} \sim \mathcal{N}(\mathbf{w}_j^\intercal \mathbf{z}_i, 1/\theta_j)$
- $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D\times D})$
- $\mathbf{W}_j \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}_{D\times D})$
- $\theta_j \sim$ Gamma$(\theta_\theta, \theta_\theta)$
- $\gamma \sim$ Gamma$(\alpha_\gamma, \beta_\gamma)$

- Num. of latent dim: D
- Num. of data dim: M
- Num. of data inst: N

The probability model

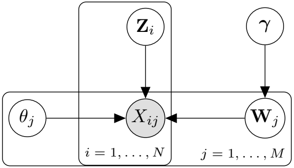
$$p(\cdot) = p(\gamma)\left[\prod_{i=1}^{N} p(\mathbf{z}_i)\right]\left[\prod_{j=1}^{M} p(\mathbf{w}_j \mid \gamma)p(\theta_j)\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{M} p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)\right]$$

…after taking the log

$$\log p(\cdot) = \log p(\gamma) + \sum_{i=1}^{N}\log p(\mathbf{z}_i) + \sum_{j=1}^{M}\left[\log p(\mathbf{w}_j \mid \gamma) + \log p(\theta_j)\right] + \sum_{i=1}^{N}\sum_{j=1}^{M}\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)$$

VB for the factor analysis model

The factor analysis model



- $X_{ij} \mid \{\mathbf{w}_j, \mathbf{z}_i, \theta_j\} \sim \mathcal{N}(\mathbf{w}_j^\intercal \mathbf{z}_i, 1/\theta_j)$
- $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D\times D})$
- $\mathbf{W}_j \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}_{D\times D})$
- $\theta_j \sim \text{Gamma}(\theta_\theta, \theta_\theta)$
- $\gamma \sim \text{Gamma}(\alpha_\gamma, \beta_\gamma)$

- Num. of latent dim: D
- Num. of data dim: M
- Num. of data inst: N

The probability model

$$p(\cdot) = p(\gamma)\left[\prod_{i=1}^{N} p(\mathbf{z}_i)\right]\left[\prod_{j=1}^{M} p(\mathbf{w}_j \mid \gamma)p(\theta_j)\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{M} p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)\right]$$

The variational model

$$q(\cdot) = q(\gamma)\prod_{i=1}^{N} q(\mathbf{z}_i \mid \cdot)\prod_{j=1}^{M} q(\mathbf{w}_j \mid \cdot)q(\theta_j \mid \cdot)$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}} [\log p(\cdot)] + c$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c$$

Recall

$$\log p(\cdot)$$
$$= \log p(\gamma) + \sum_{i=1}^{N} \log p(\mathbf{z}_i) + \sum_{j=1}^{M} \left[\log p(\mathbf{w}_j \mid \gamma) + \log p(\theta_j)\right] + \sum_{i=1}^{N}\sum_{j=1}^{M} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \,|\, \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c$$

Recall

$$\log p(\cdot)$$

$$= \log p(\gamma) + \sum_{i=1}^{N} \log p(\mathbf{z}_i) + \sum_{j=1}^{M} \left[\log p(\mathbf{w}_j \,|\, \gamma) + \log p(\theta_j)\right] + \sum_{i=1}^{N}\sum_{j=1}^{M} \log p(x_{ij} \,|\, \mathbf{w}_j, \mathbf{z}_i, \theta_j)$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}} [\log p(\cdot)] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

Recall

$$\log p(\cdot)$$

$$= \log p(\gamma) + \sum_{i=1}^{N} \log p(\mathbf{z}_i) + \sum_{j=1}^{M} [\log p(\mathbf{w}_j \mid \gamma) + \log p(\theta_j)] + \sum_{i=1}^{N}\sum_{j=1}^{M} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

The gamma and multivariate normal

$$\log p(\gamma \mid \alpha_\gamma, \beta_\gamma) = \alpha_\gamma \log(\beta_\gamma) + (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma - \log(\Gamma(\alpha_\gamma))$$

$$\log p(\mathbf{w}_j \mid \gamma) = \log \mathcal{N}(\mathbf{w}_j \mid \mathbf{0}, \gamma^{-1}\mathbf{I}) = -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{w}_j^T\mathbf{w}_j$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \,|\, \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \,|\, \gamma) + c$$

The gamma and multivariate normal

$$\log p(\gamma \,|\, \alpha_\gamma, \beta_\gamma) = \alpha_\gamma \log(\beta_\gamma) + (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma - \log(\Gamma(\alpha_\gamma))$$

$$\log p(\mathbf{w}_j \,|\, \gamma) = \log \mathcal{N}(\mathbf{w}_j \,|\, \mathbf{0}, \gamma^{-1}\mathbf{I}) = -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{w}_j^T\mathbf{w}_j$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

$$= (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma + \frac{D \cdot M}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{j=1}^{M} \mathbb{E}(\mathbf{W}_j^T \mathbf{W}_j) + c$$

The gamma and multivariate normal

$$\log p(\gamma \mid \alpha_\gamma, \beta_\gamma) = \alpha_\gamma \log(\beta_\gamma) + (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma - \log(\Gamma(\alpha_\gamma))$$

$$\log p(\mathbf{w}_j \mid \gamma) = \log \mathcal{N}(\mathbf{w}_j \mid \mathbf{0}, \gamma^{-1}\mathbf{I}) = -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{w}_j^T \mathbf{w}_j$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

$$= (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma + \frac{D \cdot M}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{j=1}^{M} \mathbb{E}(\mathbf{W}_j^T \mathbf{W}_j) + c$$

$$= \left[\alpha_\gamma - 1 + \frac{D \cdot M}{2}\right]\log(\gamma) - \left[\beta_\gamma + \frac{1}{2}\sum_{j=1}^{M} \mathbb{E}(\mathbf{W}_j^T \mathbf{W}_j)\right] \cdot \gamma + c$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

$$= (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma + \frac{D \cdot M}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j) + c$$

$$= \left[\alpha_\gamma - 1 + \frac{D \cdot M}{2}\right]\log(\gamma) - \left[\beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j)\right]\cdot\gamma + c$$

Recall the (generic) gamma distribution

$$\log p(x \mid \alpha, \beta) = \alpha\log(\beta) + (\alpha - 1)\log(x) - \beta \cdot x - \log(\Gamma(\alpha))$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

$$= (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma + \frac{D \cdot M}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T \mathbf{W}_j) + c$$

$$= \left[\alpha_\gamma - 1 + \frac{D \cdot M}{2}\right]\log(\gamma) - \left[\beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T \mathbf{W}_j)\right] \cdot \gamma + c$$

Recall the (generic) gamma distribution

$$\log p(x \mid \alpha, \beta) = \alpha \log(\beta) + (\alpha - 1)\log(x) - \beta \cdot x - \log(\Gamma(\alpha))$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

$$= (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma + \frac{D \cdot M}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j) + c$$

$$= \left[\alpha_\gamma - 1 + \frac{D \cdot M}{2}\right]\log(\gamma) - \left[\beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j)\right] \cdot \gamma + c$$

Thus we see that $q(\gamma \mid \cdot)$ is gamma distributed with

- shape parameter: $\alpha \leftarrow \alpha_\gamma + \frac{DM}{2}$
- rate parameter: $\beta \leftarrow \beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}_{q(\mathbf{w}_j)}\left[\mathbf{W}_j^T\mathbf{W}_j\right]$

Recall the (generic) gamma distribution

$$\log p(x \mid \alpha, \beta) = \alpha \log(\beta) + (\alpha - 1)\log(x) - \beta \cdot x - \log(\Gamma(\alpha))$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \,|\, \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \,|\, \gamma) + c$$

$$= (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma + \frac{D \cdot M}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j) + c$$

$$= \left[\alpha_\gamma - 1 + \frac{D \cdot M}{2}\right]\log(\gamma) - \left[\beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j)\right]\cdot\gamma + c$$

Thus we see that $q(\gamma \,|\, \cdot)$ is gamma distributed with

- shape parameter: $\alpha \leftarrow \alpha_\gamma + \frac{DM}{2}$
- rate parameter: $\beta \leftarrow \beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}_{q(\mathbf{w}_j)}\left[\mathbf{W}_j^T\mathbf{W}_j\right]$

Calculation of $\mathbb{E}_{q(\mathbf{w}_j)}\left[\mathbf{W}_j^T\mathbf{W}_j\right]$

$$\mathbb{E}_{q(\mathbf{w}_j)}\left[\mathbf{W}_j^T\mathbf{W}_j\right] = \sum_{d=1}^{D}\text{Var}_{q(\mathbf{w}_j)}\left[\mathbf{W}_{jd}\right] + \sum_{d=1}^{D}\left(\mathbb{E}_{q(\mathbf{w}_j)}\left[\mathbf{W}_{j,d}\right]\right)^2$$

VB for the factor analysis model: updating $q(\gamma)$

We choose the variational distribution so that

$$\log q(\gamma \mid \cdot) = \mathbb{E}_{q_{\neg\gamma}}\left[\log p(\cdot)\right] + c = \log p(\gamma) + \sum_{j=1}^{M} \mathbb{E}_{q_{\neg j}} \log p(\mathbf{w}_j \mid \gamma) + c$$

$$= (\alpha_\gamma - 1)\log(\gamma) - \beta_\gamma \cdot \gamma + \frac{D \cdot M}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j) + c$$

$$= \left[\alpha_\gamma - 1 + \frac{D \cdot M}{2}\right]\log(\gamma) - \left[\beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}(\mathbf{W}_j^T\mathbf{W}_j)\right]\cdot\gamma + c$$

Thus we see that $q(\gamma \mid \cdot)$ is gamma distributed with

- shape parameter: $\alpha \leftarrow \alpha_\gamma + \frac{DM}{2}$
- rate parameter: $\beta \leftarrow \beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbb{E}_{q(\mathbf{w}_j)}\left[\mathbf{W}_j^T\mathbf{W}_j\right]$

Compare this to the Gibbs sampler:

- shape parameter: $\alpha \leftarrow \alpha_\gamma + \frac{DM}{2}$
- rate parameter: $\beta \leftarrow \beta_\gamma + \frac{1}{2}\sum_{j=1}^{M}\mathbf{w}_j^{\mathsf{T}}\mathbf{w}_j$

VB uses posterior expectations where Gibbs uses samples!

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q\neg\mathbf{w}_j}[\log p(\cdot)] + c$$

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q \neg \mathbf{w}_j}[\log p(\cdot)] + c$$

Recall

$$\log p(\cdot)$$
$$= \log p(\gamma) + \sum_{i=1}^{N} \log p(\mathbf{z}_i) + \sum_{j=1}^{M} [\log p(\mathbf{w}_j \mid \gamma) + \log p(\theta_j)] + \sum_{i=1}^{N} \sum_{j=1}^{M} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)$$

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c$$

Recall

$$\log p(\cdot)$$

$$= \log p(\gamma) + \sum_{i=1}^{N} \log p(\mathbf{z}_i) + \sum_{j=1}^{M} [\log p(\mathbf{w}_j \mid \gamma) + \log p(\theta_j)] + \sum_{i=1}^{N} \sum_{j=1}^{M} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)$$

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E} \log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

Recall

$$\log p(\cdot)$$

$$= \log p(\gamma) + \sum_{i=1}^{N} \log p(\mathbf{z}_i) + \sum_{j=1}^{M} [\log p(\mathbf{w}_j \mid \gamma) + \log p(\theta_j)] + \sum_{i=1}^{N} \sum_{j=1}^{M} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j)$$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q\neg\mathbf{w}_j}[\log p(\cdot)] + c = \mathbb{E}\log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E}\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

The (multivariate) normal distribution

$$\log p(\mathbf{w}_j \mid \gamma) = \log \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}) = -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{w}_j^T\mathbf{w}_j$$

$$\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) = \log \mathcal{N}(\mathbf{w}_j^\intercal\mathbf{z}_i, \theta_j) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\theta_j) - \frac{\theta_j}{2}(x_{ij} - \mathbf{w}_j^T\mathbf{z}_i)^2$$

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E} \log p(\mathbf{w}_j \,|\, \gamma) + \sum_{i=1}^{N} \mathbb{E} \log p(x_{ij} \,|\, \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

The (multivariate) normal distribution

$$\log p(\mathbf{w}_j \,|\, \gamma) = \log \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}) = -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{w}_j^T \mathbf{w}_j$$

$$\log p(x_{ij} \,|\, \mathbf{w}_j, \mathbf{z}_i, \theta_j) = \log \mathcal{N}(\mathbf{w}_j^\intercal \mathbf{z}_i, \theta_j) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\theta_j) - \frac{\theta_j}{2}(x_{ij} - \mathbf{w}_j^T \mathbf{z}_i)^2$$

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E} \log p(\mathbf{w}_j \,|\, \gamma) + \sum_{i=1}^{N} \mathbb{E} \log p(x_{ij} \,|\, \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{\mathbb{E}(\gamma)}{2} \mathbf{w}_j^T \mathbf{w}_j - \frac{\mathbb{E}(\theta_j)}{2} \sum_{i=1}^{N} \mathbb{E}(x_{ij} - \mathbf{w}_j^T \mathbf{Z}_i)^2 + c$$

The (multivariate) normal distribution

$$\log p(\mathbf{w}_j \,|\, \gamma) = \log \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}) = -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{w}_j^T \mathbf{w}_j$$

$$\log p(x_{ij} \,|\, \mathbf{w}_j, \mathbf{z}_i, \theta_j) = \log \mathcal{N}(\mathbf{w}_j^{\mathsf{T}} \mathbf{z}_i, \theta_j) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\theta_j) - \frac{\theta_j}{2}(x_{ij} - \mathbf{w}_j^T \mathbf{z}_i)^2$$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E}\log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E}\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{\mathbb{E}(\gamma)}{2}\mathbf{w}_j^T\mathbf{w}_j - \frac{\mathbb{E}(\theta_j)}{2}\sum_{i=1}^{N}\mathbb{E}(x_{ij} - \mathbf{w}_j^T\mathbf{Z}_i)^2 + c$$

Expanding the square

$$\begin{aligned}(x_{ij} - \mathbf{w}_j^T\mathbf{z}_i)^2 &= x_{ij}^2 - 2x_{ij}\mathbf{w}_j^\intercal\mathbf{z}_i + \mathbf{w}_j^\intercal\mathbf{z}_i\mathbf{w}_j^\intercal\mathbf{z}_i \\ &= x_{ij}^2 - 2x_{ij}\mathbf{w}_j^\intercal\mathbf{z}_i + \mathbf{w}_j^\intercal\mathbf{z}_i\mathbf{z}_i^\intercal\mathbf{w}_j\end{aligned}$$

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q\neg\mathbf{w}_j}[\log p(\cdot)] + c = \mathbb{E}\log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E}\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{\mathbb{E}(\gamma)}{2}\mathbf{w}_j^T\mathbf{w}_j - \frac{\mathbb{E}(\theta_j)}{2}\sum_{i=1}^{N}\mathbb{E}(x_{ij} - \mathbf{w}_j^T\mathbf{Z}_i)^2 + c$$

$$= -\frac{\mathbb{E}(\gamma)}{2}\mathbf{w}_j^T\mathbf{w}_j - \mathbb{E}(\theta_j)\frac{1}{2}\sum_{i=1}^{N}\left[-2x_{ij}\mathbf{w}_j^\intercal\mathbb{E}(\mathbf{Z}_i) + \mathbf{w}_j^\intercal\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\intercal)\mathbf{w}_j)\right] + c$$

Expanding the square

$$\begin{aligned}(x_{ij} - \mathbf{w}_j^T\mathbf{z}_i)^2 &= x_{ij}^2 - 2x_{ij}\mathbf{w}_j^\intercal\mathbf{z}_i + \mathbf{w}_j^\intercal\mathbf{z}_i\mathbf{w}_j^\intercal\mathbf{z}_i \\ &= x_{ij}^2 - 2x_{ij}\mathbf{w}_j^\intercal\mathbf{z}_i + \mathbf{w}_j^\intercal\mathbf{z}_i\mathbf{z}_i^\intercal\mathbf{w}_j\end{aligned}$$

We choose the variational distribution so that

$$\begin{aligned}
\log q(\mathbf{w}_j) &= \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E}\log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E}\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c \\
&= -\frac{\mathbb{E}(\gamma)}{2}\mathbf{w}_j^T\mathbf{w}_j - \frac{\mathbb{E}(\theta_j)}{2}\sum_{i=1}^{N}\mathbb{E}(x_{ij} - \mathbf{w}_j^T\mathbf{Z}_i)^2 + c \\
&= -\frac{\mathbb{E}(\gamma)}{2}\mathbf{w}_j^T\mathbf{w}_j - \mathbb{E}(\theta_j)\frac{1}{2}\sum_{i=1}^{N}\left[-2x_{ij}\mathbf{w}_j^{\intercal}\mathbb{E}(\mathbf{Z}_i) + \mathbf{w}_j^{\intercal}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^{\intercal})\mathbf{w}_j)\right] + c \\
&= -\frac{1}{2}\mathbf{w}_j^T\left[\mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^{\intercal})\right]\mathbf{w}_j + \mathbf{w}_j^{\intercal}\mathbb{E}(\theta_j)\sum_{i=1}^{N}x_{ij}\mathbb{E}(\mathbf{Z}_i) + c
\end{aligned}$$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E} \log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{1}{2}\mathbf{w}_j^T \left[\mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j) \sum_{i=1}^{N} \mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\intercal)\right] \mathbf{w}_j + \mathbf{w}_j^\intercal \mathbb{E}(\theta_j) \sum_{i=1}^{N} x_{ij} \mathbb{E}(\mathbf{Z}_i) + c$$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E} \log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E} \log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{1}{2}\mathbf{w}_j^T \left[\mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\intercal)\right]\mathbf{w}_j + \mathbf{w}_j^\intercal \mathbb{E}(\theta_j)\sum_{i=1}^{N} x_{ij}\mathbb{E}(\mathbf{Z}_i) + c$$

Recall the (generic) multivariate normal distribution

$$\log p(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{Q}) = \log \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{Q}^{-1}) = -\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{Q}] - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\intercal \mathbf{Q}(\mathbf{y}-\boldsymbol{\mu})^\intercal)$$

$$= -\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{Q}] - \frac{1}{2}\mathbf{y}^\intercal \mathbf{Q}\mathbf{y} + \mathbf{y}^\intercal \mathbf{Q}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\intercal \mathbf{Q}\boldsymbol{\mu})$$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q\neg\mathbf{w}_j}[\log p(\cdot)] + c = \mathbb{E}\log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E}\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{1}{2}\mathbf{w}_j^T \left[\mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\intercal)\right]\mathbf{w}_j + \mathbf{w}_j^\intercal\mathbb{E}(\theta_j)\sum_{i=1}^{N} x_{ij}\mathbb{E}(\mathbf{Z}_i) + c$$

Thus we see that $q(\mathbf{w}_j \mid \cdot)$ is normally distributed with

- precision $\mathbf{Q} \leftarrow \mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\intercal)$
- mean $\boldsymbol{\mu} \leftarrow \mathbf{Q}^{-1}\left[\mathbb{E}(\theta_j)\sum_{i=1}^{N} x_{ij}\mathbb{E}(\mathbf{Z}_i)\right]$

Recall the (generic) multivariate normal distribution

$$\log p(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{Q}) = \log\mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{Q}^{-1}) = -\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{Q}] - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\intercal\mathbf{Q}(\mathbf{y}-\boldsymbol{\mu})^\intercal)$$

$$= -\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{Q}] - \frac{1}{2}\mathbf{y}^\intercal\mathbf{Q}\mathbf{y} + \mathbf{y}^\intercal\mathbf{Q}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\intercal\mathbf{Q}\boldsymbol{\mu})$$

VB for the factor analysis model: updating $q(\mathbf{w}_j)$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q_{\neg \mathbf{w}_j}}[\log p(\cdot)] + c = \mathbb{E}\log p(\mathbf{w}_j \mid \gamma) + \sum_{i=1}^{N} \mathbb{E}\log p(x_{ij} \mid \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{1}{2}\mathbf{w}_j^T \left[\mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^{\mathsf{T}})\right]\mathbf{w}_j + \mathbf{w}_j^{\mathsf{T}}\mathbb{E}(\theta_j)\sum_{i=1}^{N} x_{ij}\mathbb{E}(\mathbf{Z}_i) + c$$

Thus we see that $q(\mathbf{w}_j \mid \cdot)$ is normally distributed with

- precision $\mathbf{Q} \leftarrow \mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^{\mathsf{T}})$
- mean $\boldsymbol{\mu} \leftarrow \mathbf{Q}^{-1}\left[\mathbb{E}(\theta_j)\sum_{i=1}^{N} x_{ij}\mathbb{E}(\mathbf{Z}_i)\right]$

Calculation of $\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^{\mathsf{T}})$

$$\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^{\mathsf{T}}) = \mathrm{Cov}(\mathbf{Z}_i) + \mathbb{E}(\mathbf{Z}_i)\mathbb{E}(\mathbf{Z}_i)^{\mathsf{T}}$$

We choose the variational distribution so that

$$\log q(\mathbf{w}_j) = \mathbb{E}_{q\neg\mathbf{w}_j}[\log p(\cdot)] + c = \mathbb{E}\log p(\mathbf{w}_j \,|\, \gamma) + \sum_{i=1}^{N} \mathbb{E}\log p(x_{ij} \,|\, \mathbf{w}_j, \mathbf{z}_i, \theta_j) + c$$

$$= -\frac{1}{2}\mathbf{w}_j^T \left[\mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\intercal)\right]\mathbf{w}_j + \mathbf{w}_j^\intercal\mathbb{E}(\theta_j)\sum_{i=1}^{N} x_{ij}\mathbb{E}(\mathbf{Z}_i) + c$$

Thus we see that $q(\mathbf{w}_j \,|\, \cdot)$ is normally distributed with

- precision $\mathbf{Q} \leftarrow \mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j)\sum_{i=1}^{N}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\intercal)$
- mean $\boldsymbol{\mu} \leftarrow \mathbf{Q}^{-1}\left[\mathbb{E}(\theta_j)\sum_{i=1}^{N} x_{ij}\mathbb{E}(\mathbf{Z}_i)\right]$

Compare this to the Gibbs sampler:

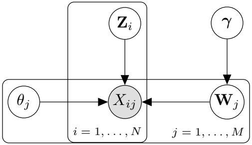
- $\mathbf{Q} \leftarrow \gamma\mathbf{I} + \theta_j\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i^\intercal$
- $\boldsymbol{\mu} \leftarrow \mathbf{Q}^{-1}\left[\theta_j\sum_{i=1}^{N} x_{ij}\mathbf{z}_i\right]$

Once again, the only difference between VB and Gibbs is that where VB uses posterior expectations, Gibbs uses samples.

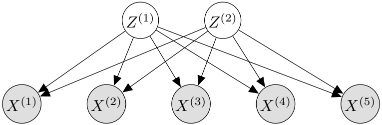
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

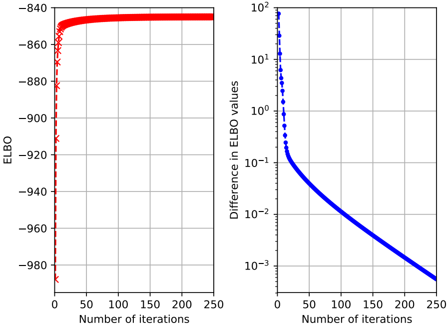
Global model



Local model



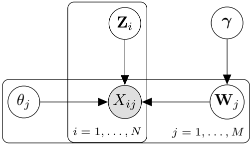
Monitoring convergence



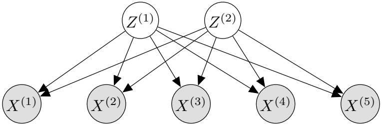
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

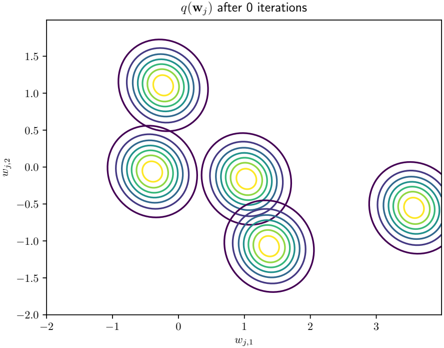
Global model



Local model



Variational posteriors

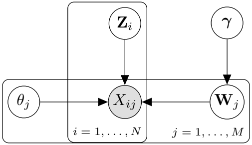


VB for the factor analysis model

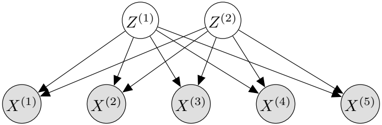
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

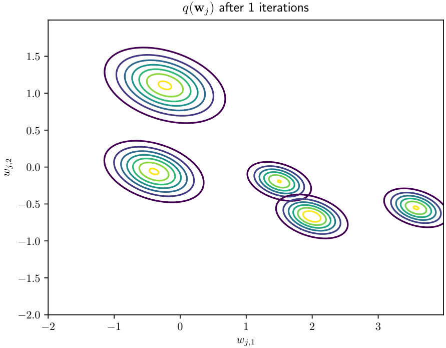
Global model



Local model



Variational posteriors

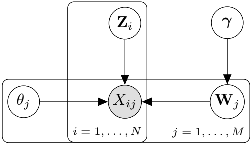


VB for the factor analysis model

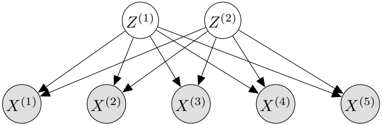
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

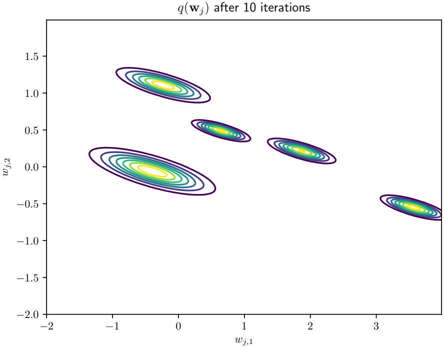
Global model



Local model



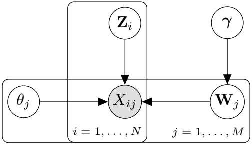
Variational posteriors



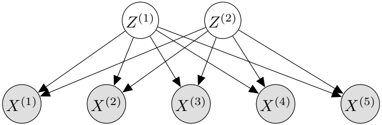
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

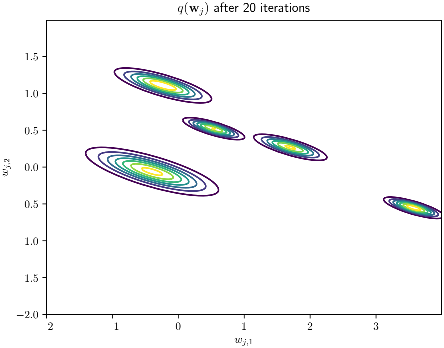
Global model



Local model



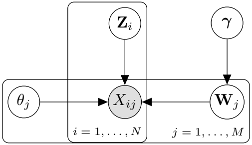
Variational posteriors



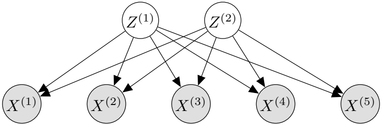
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

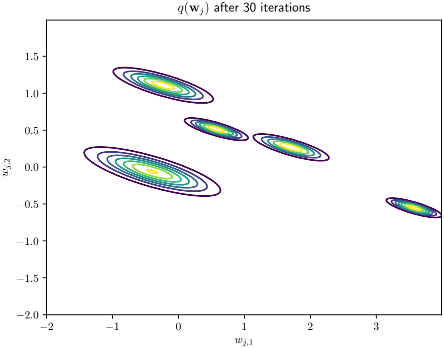
Global model



Local model



Variational posteriors

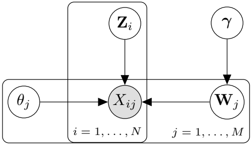


VB for the factor analysis model

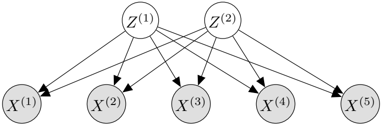
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

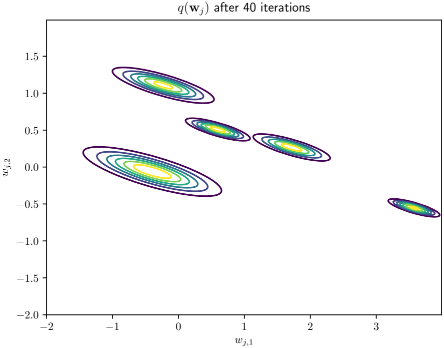
Global model



Local model



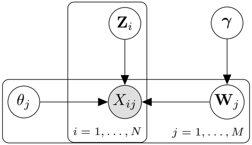
Variational posteriors



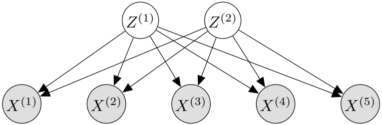
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

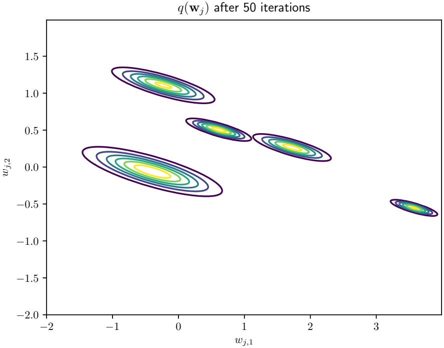
Global model



Local model



Variational posteriors

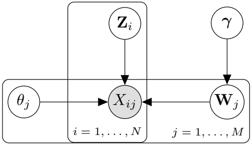


VB for the factor analysis model

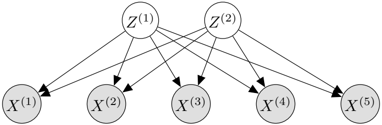
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

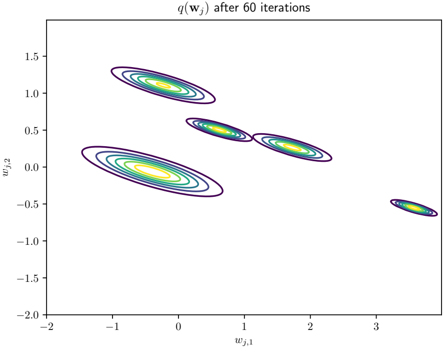
Global model



Local model



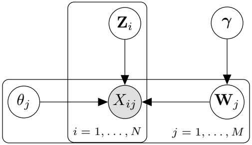
Variational posteriors



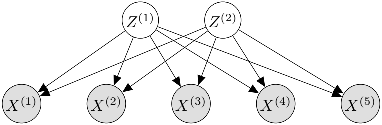
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

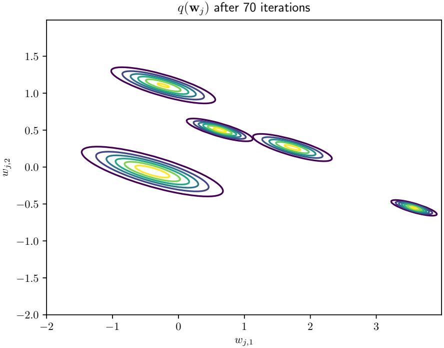
Global model



Local model



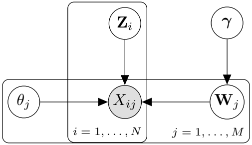
Variational posteriors



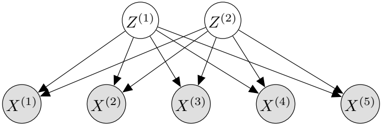
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

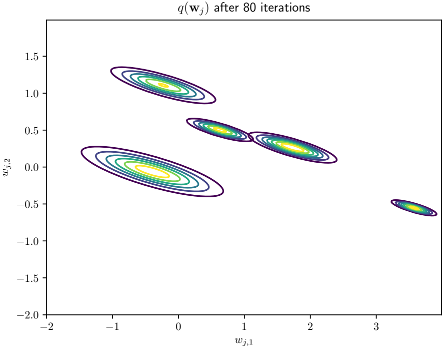
Global model



Local model



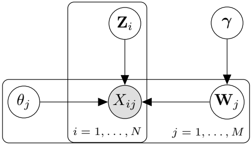
Variational posteriors



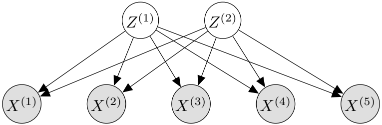
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

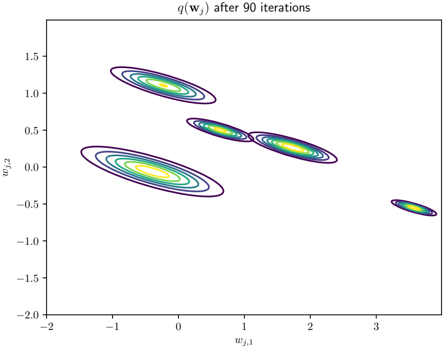
Global model



Local model



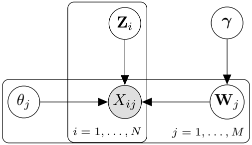
Variational posteriors



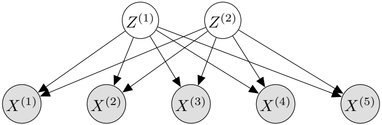
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

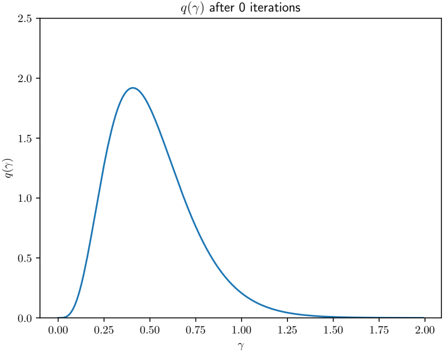
Global model



Local model



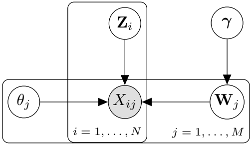
Variational posteriors



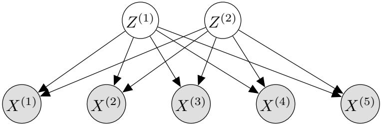
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

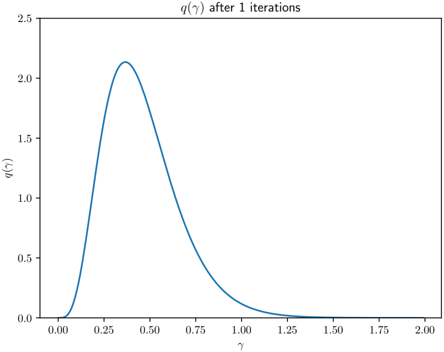
Global model



Local model



Variational posteriors

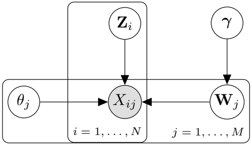


VB for the factor analysis model

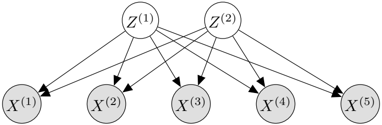
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

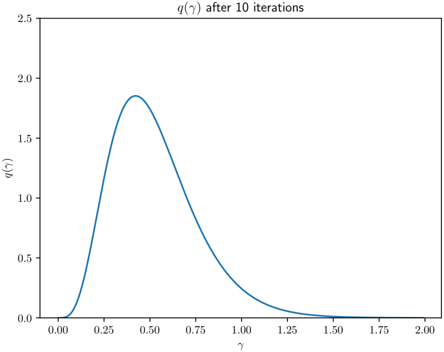
Global model



Local model



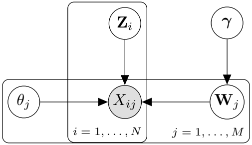
Variational posteriors



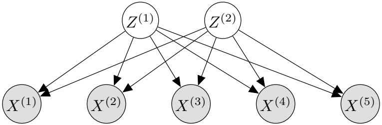
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

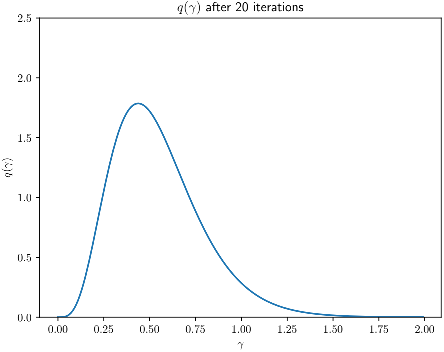
Global model



Local model



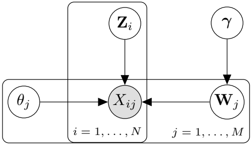
Variational posteriors



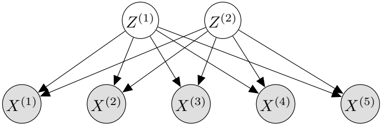
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

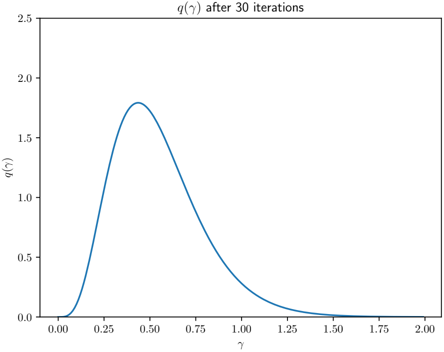
Global model



Local model



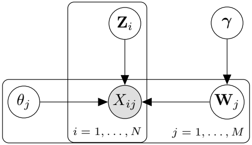
Variational posteriors



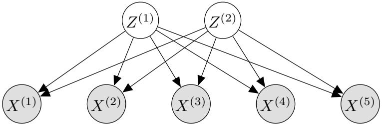
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

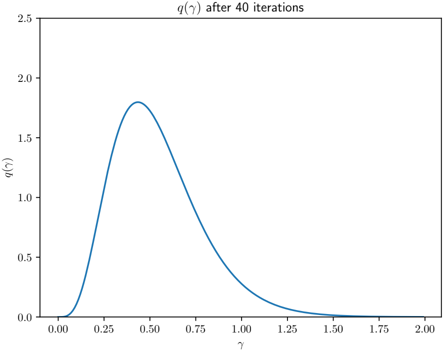
Global model



Local model



Variational posteriors

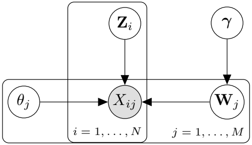


VB for the factor analysis model

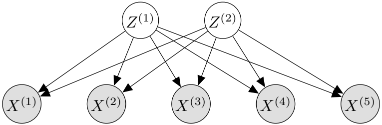
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

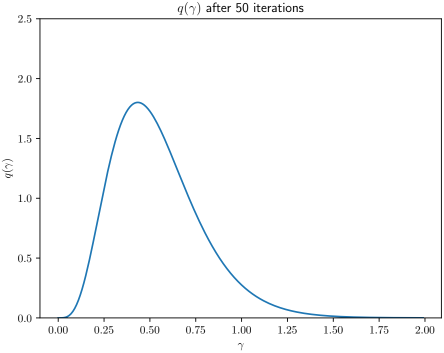
Global model



Local model



Variational posteriors

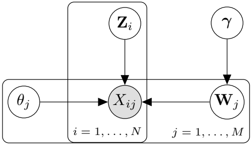


VB for the factor analysis model

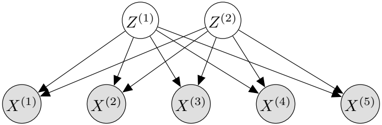
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

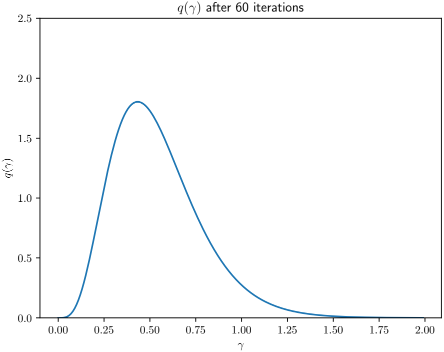
Global model



Local model



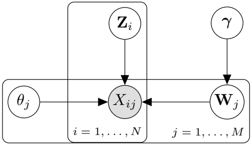
Variational posteriors



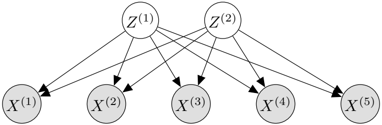
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

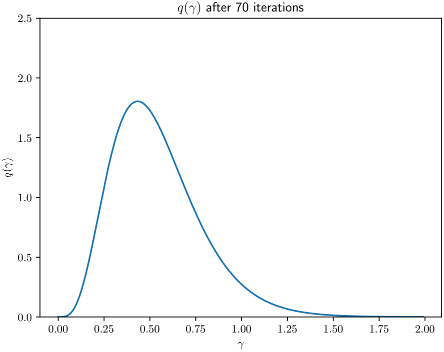
Global model



Local model



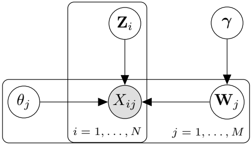
Variational posteriors



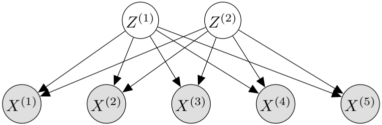
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

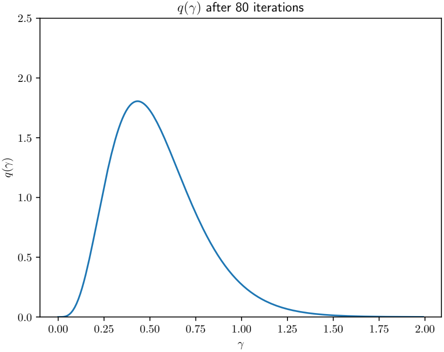
Global model



Local model



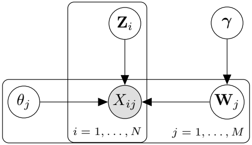
Variational posteriors



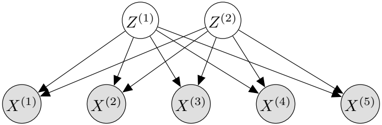
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

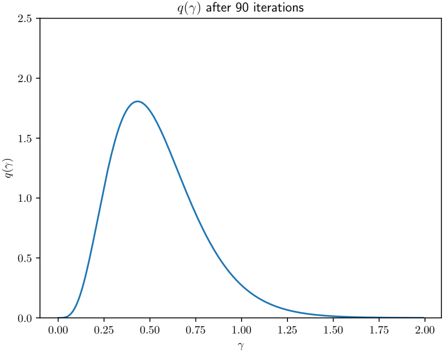
Global model



Local model



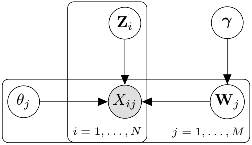
Variational posteriors



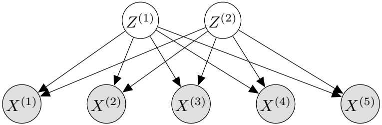
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

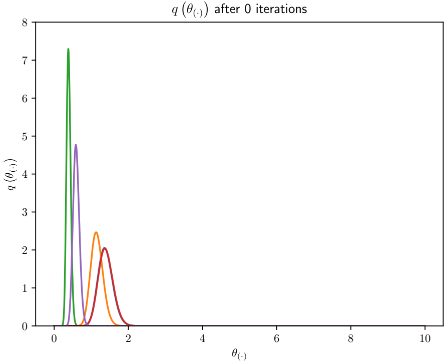
Global model



Local model



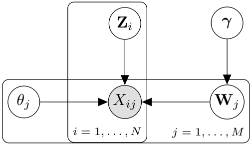
Variational posteriors



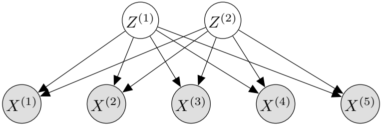
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

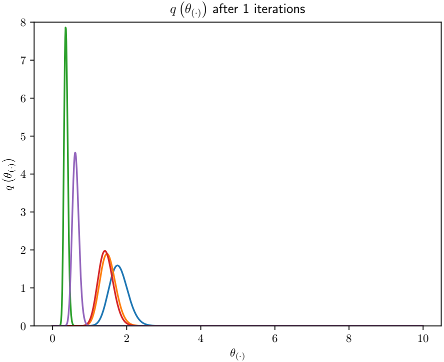
Global model



Local model



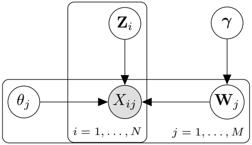
Variational posteriors



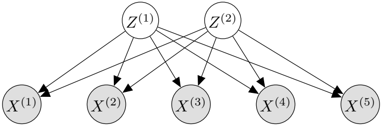
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

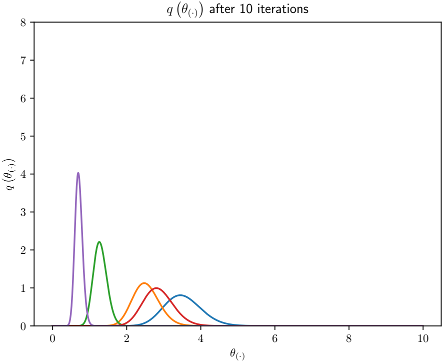
Global model



Local model



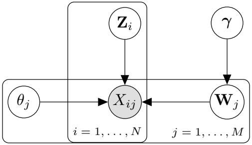
Variational posteriors



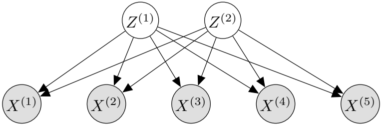
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

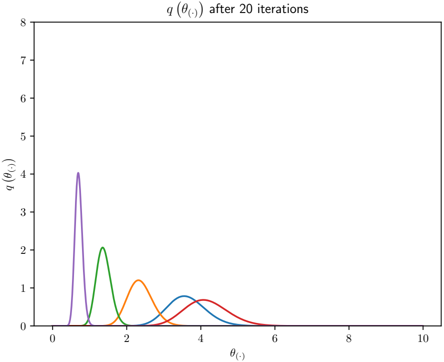
Global model



Local model



Variational posteriors

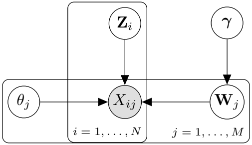


VB for the factor analysis model

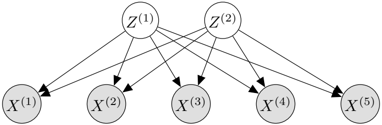
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

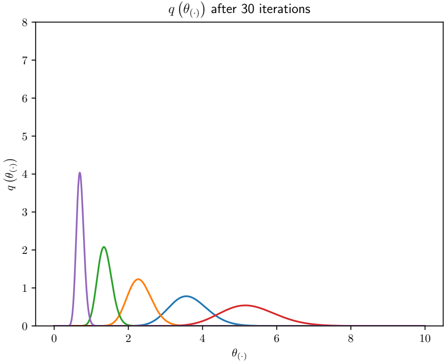
Global model



Local model



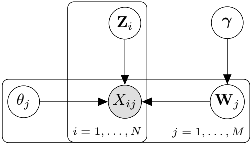
Variational posteriors



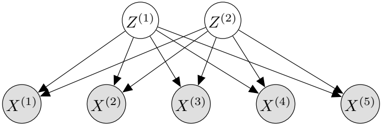
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

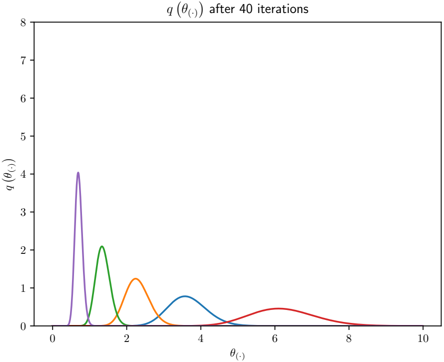
Global model



Local model



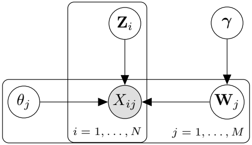
Variational posteriors



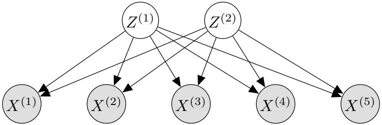
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

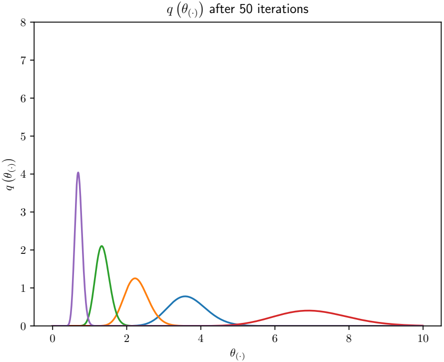
Global model



Local model



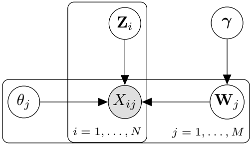
Variational posteriors



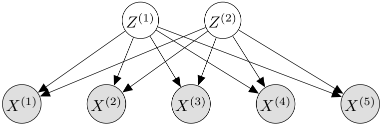
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

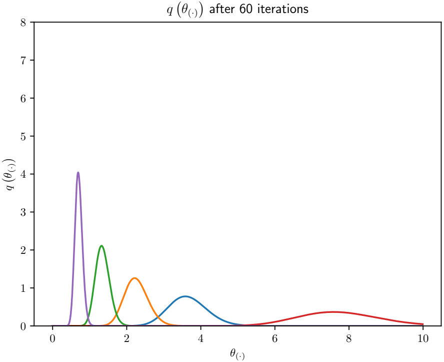
Global model



Local model



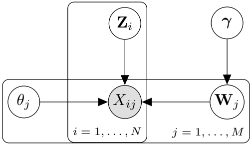
Variational posteriors



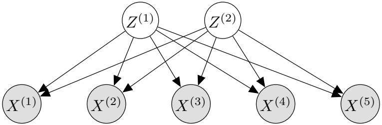
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

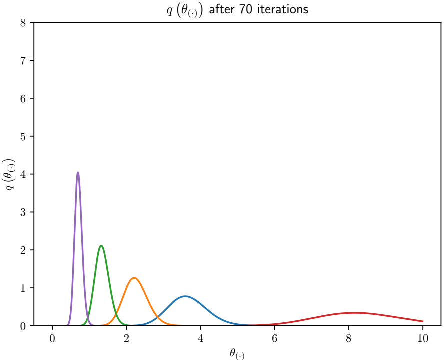
Global model



Local model



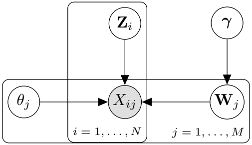
Variational posteriors



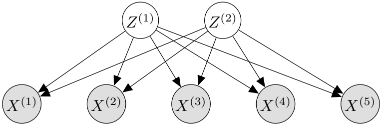
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

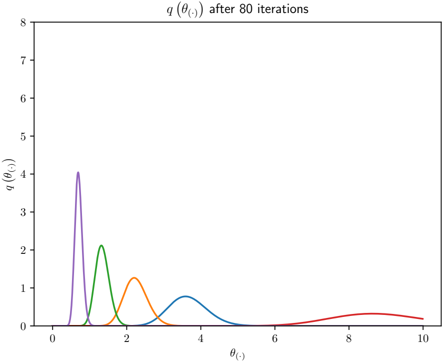
Global model



Local model



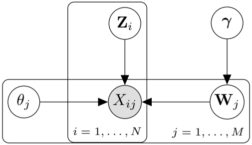
Variational posteriors



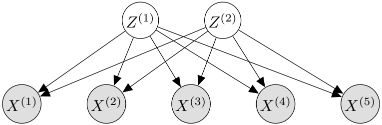
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

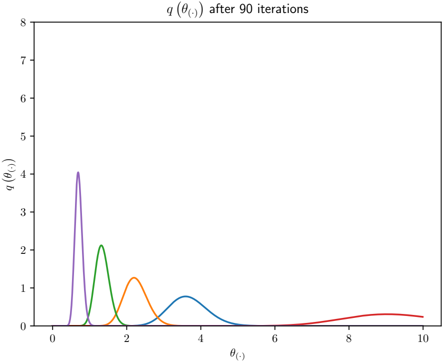
Global model



Local model



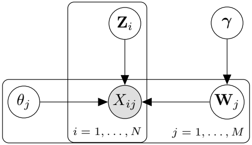
Variational posteriors



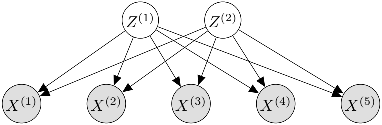
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

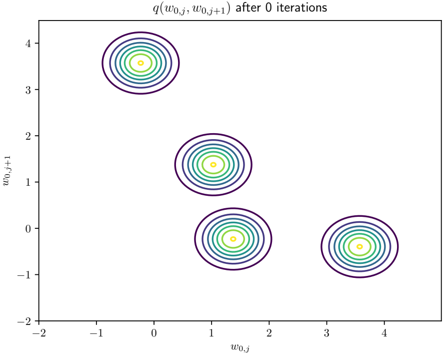
Global model



Local model



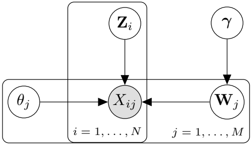
Variational posteriors



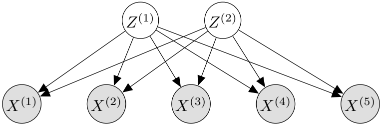
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

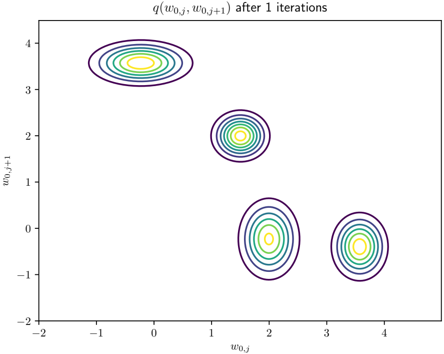
Global model



Local model



Variational posteriors

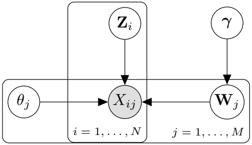


VB for the factor analysis model

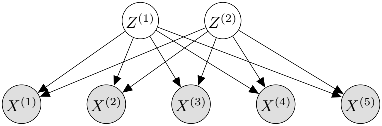
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

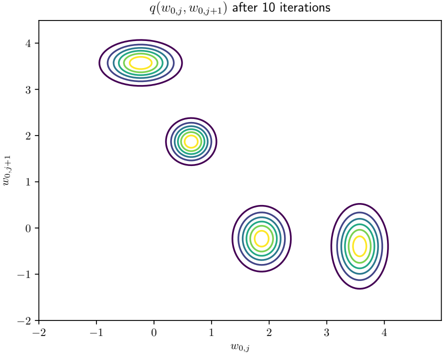
Global model



Local model



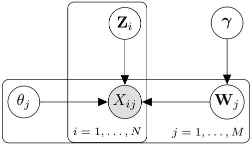
Variational posteriors



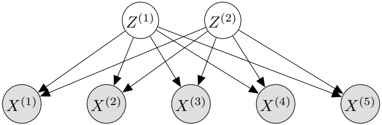
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

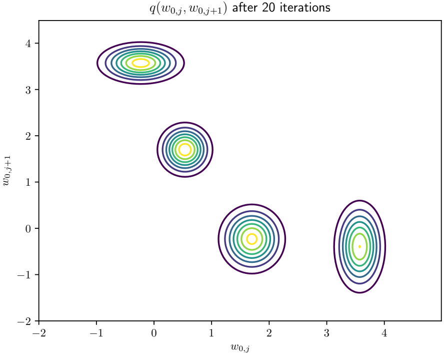
Global model



Local model



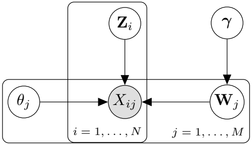
Variational posteriors



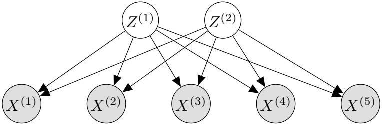
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

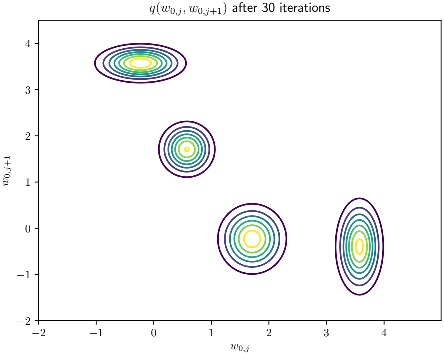
Global model



Local model



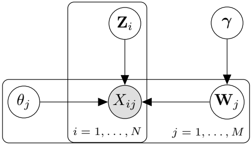
Variational posteriors



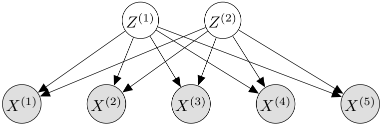
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

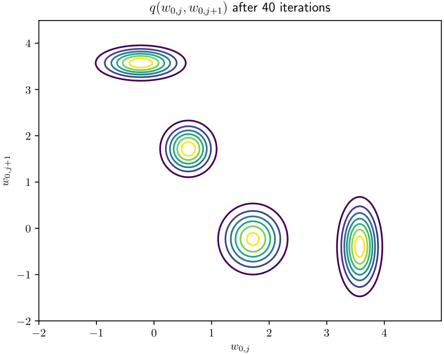
Global model



Local model



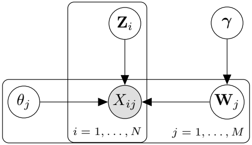
Variational posteriors



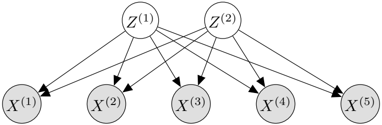
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

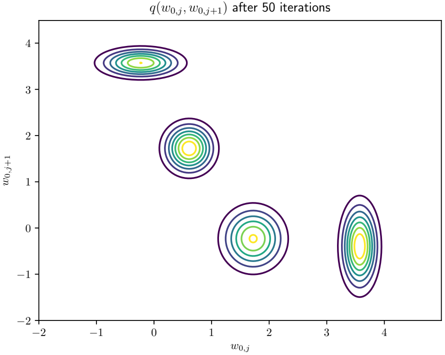
Global model



Local model



Variational posteriors

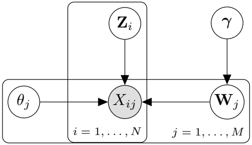


VB for the factor analysis model

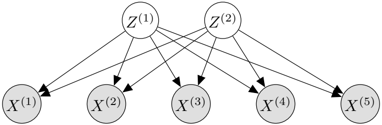
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

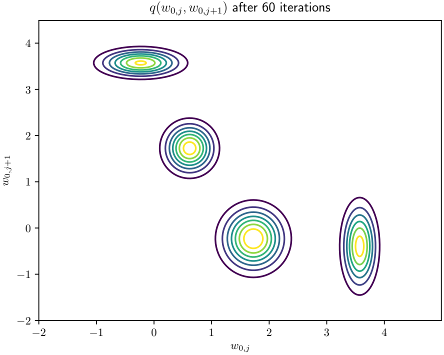
Global model



Local model



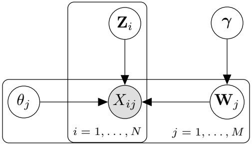
Variational posteriors



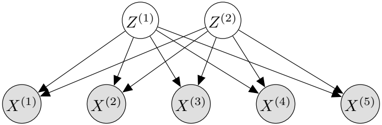
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

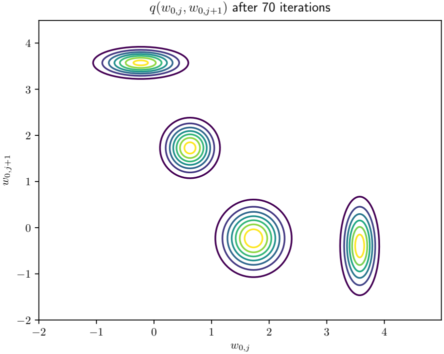
Global model



Local model



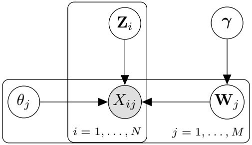
Variational posteriors



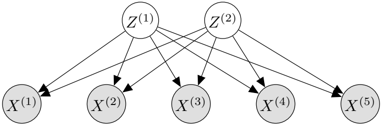
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

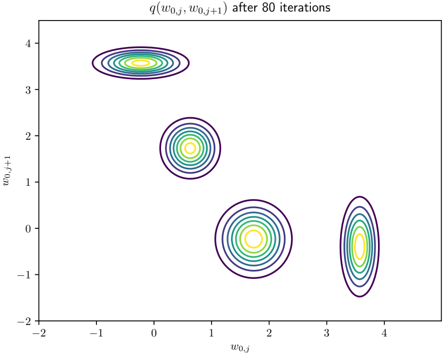
Global model



Local model



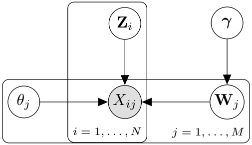
Variational posteriors



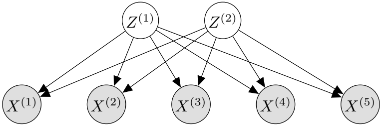
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

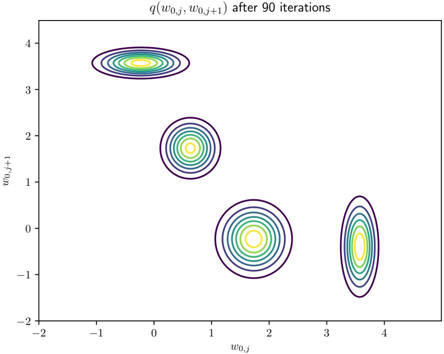
Global model



Local model



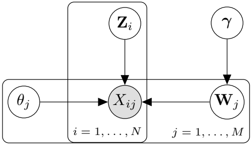
Variational posteriors



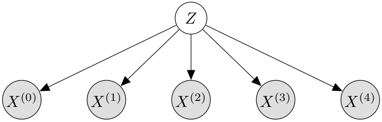
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

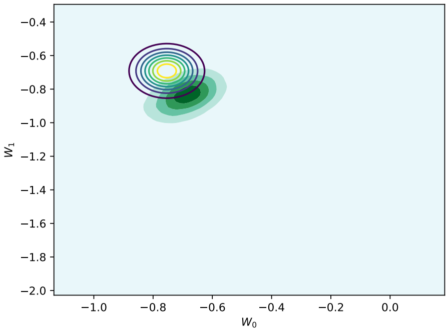
Global model



Local model



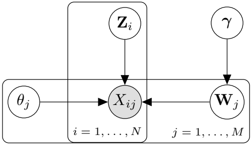
Comparison with Gibbs sampling



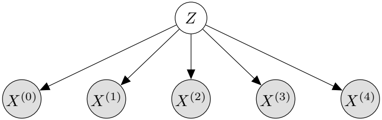
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

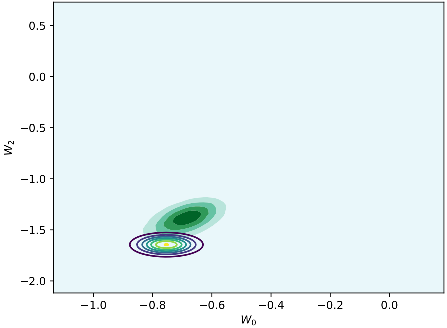
Global model



Local model



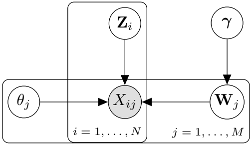
Comparison with Gibbs sampling



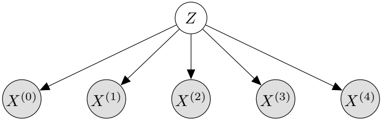
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

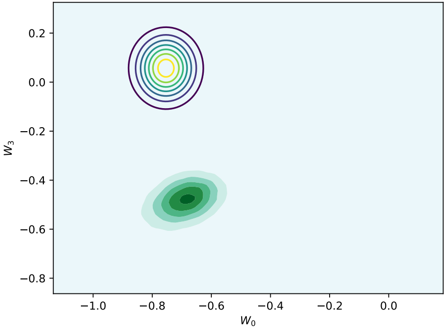
Global model



Local model



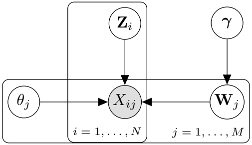
Comparison with Gibbs sampling



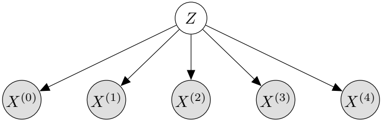
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

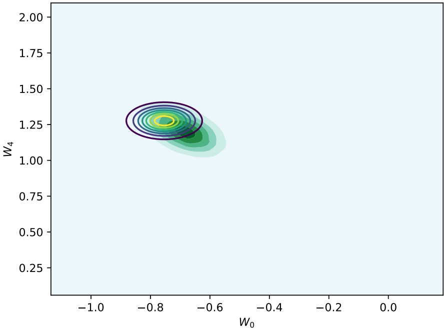
Global model



Local model



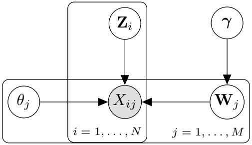
Comparison with Gibbs sampling



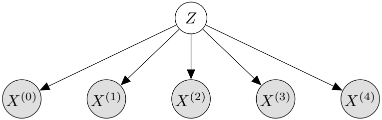
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

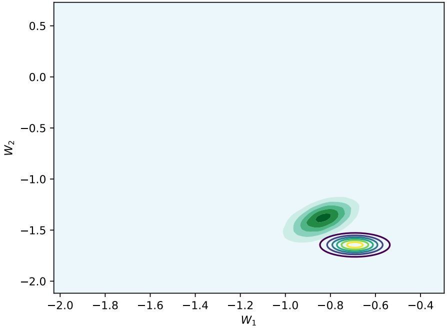
Global model



Local model



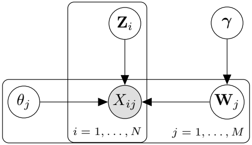
Comparison with Gibbs sampling



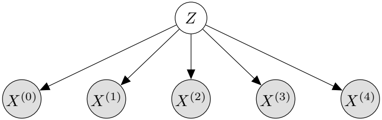
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

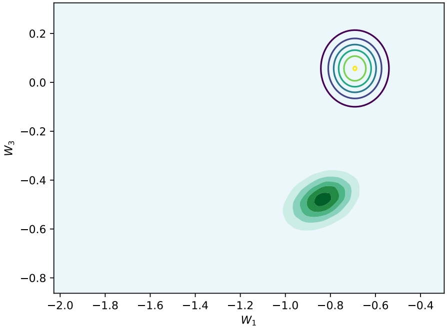
Global model



Local model



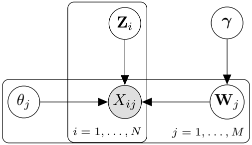
Comparison with Gibbs sampling



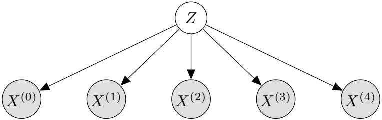
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

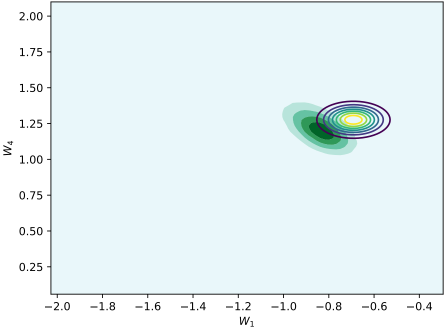
Global model



Local model



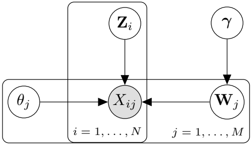
Comparison with Gibbs sampling



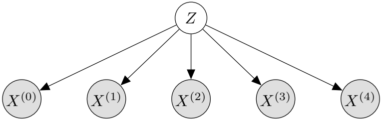
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

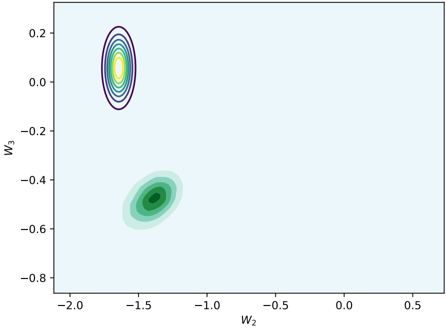
Global model



Local model



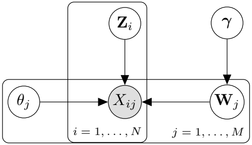
Comparison with Gibbs sampling



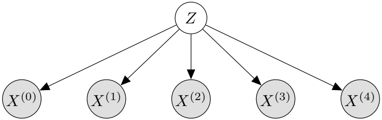
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

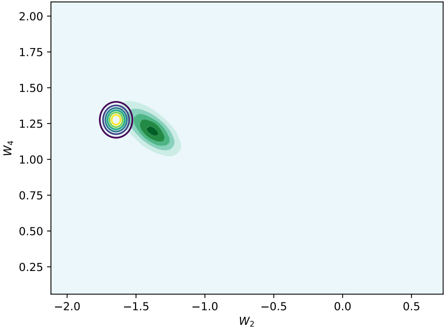
Global model



Local model



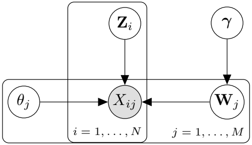
Comparison with Gibbs sampling



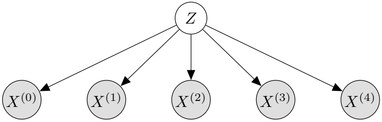
Data

100 data points was randomly sampled from a 5-dim multivariate Gaussian distribution.

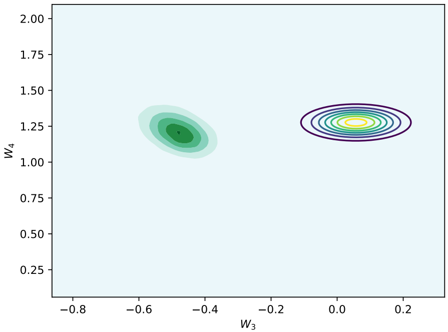
Global model



Local model



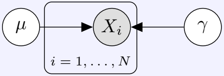
Comparison with Gibbs sampling



Not seen from the plot, **but** the results strongly dependent on the VI initialization.

Code-task: VB for a simple Gaussian model

Code Task: VB for a simple Gaussian model



- $X_i \mid \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau)$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

In this task you need to use mean-field, and look for $q(\mu, \gamma) = q(\mu) \cdot q(\tau)$ that best approximates $p(\mu, \tau \mid x_1, \ldots, x_N)$ wrt. the VB measure $\mathrm{KL}\left(q||p\right)$.

- Calculate the update rules for $q(\mu)$ and $q(\gamma)$.
  - Hint: $q(\mu)$ is Gaussian with mean ν^* and precision τ^*; $q(\gamma)$ is Gamma-distributed with parameters α^* and β^*.
- Implement the update rules you find in the notebook
  `students_simple_model.ipynb`

It may be useful to recall the definition of pdfs:

- Gamma: $\log p(x \mid \alpha, \beta) = \alpha \log(\beta) + (\alpha - 1)\log(x) - \beta \cdot x - \log(\Gamma(\alpha))$.
- Gauss: $\log p(x \mid \mu, 1/\gamma) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\gamma) - \frac{\gamma}{2}\left(x - \mu\right)^2$.

Wrapping it all up: The VB algorithm under MF

Algorithm:

- We have observed $\mathbf{X} = \mathbf{x}$, and have access to the full joint $p(\mathbf{z}, \mathbf{x})$.
- We posit a *variational family* of distributions $q_j(\cdot \mid \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.
- The posterior approximation is assumed to factorize according to the mean-field assumption, and we use the $\mathrm{KL}\left(q(\mathbf{z}) || p(\mathbf{z} \mid \mathbf{x})\right)$ as our objective.

Algorithm:

Repeat until negligible improvement in terms of $\mathcal{L}\left(q\right)$:

- For each j:
  - Calculate $\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]$ using current estimates for $q_i(\cdot \mid \boldsymbol{\lambda}_i)$, $i \neq j$.
  - Choose $\boldsymbol{\lambda}_j$ so that $q_j(z_j \mid \boldsymbol{\lambda}_j) \propto \exp\left(\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right)$.
- Calculate the new $\mathcal{L}\left(q\right)$.

Algorithm:

- We have observed $\mathbf{X} = \mathbf{x}$, and have access to the full joint $p(\mathbf{z}, \mathbf{x})$.
- We posit a *variational family* of distributions $q_j(\cdot \mid \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.
- The posterior approximation is assumed to factorize according to the mean-field assumption, and we use the $\mathrm{KL}\left(q(\mathbf{z}) || p(\mathbf{z} \mid \mathbf{x})\right)$ as our objective.

Algorithm:

Repeat until negligible improvement in terms of $\mathcal{L}(q)$:

- For each j:
  - Calculate $\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]$ using current estimates for $q_i(\cdot \mid \boldsymbol{\lambda}_i)$, $i \neq j$.
  - Choose $\boldsymbol{\lambda}_j$ so that $q_j(z_j \mid \boldsymbol{\lambda}_j) \propto \exp\left(\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right)$.
- Calculate the new $\mathcal{L}(q)$.

As we just realized, calculations of $\mathbb{E}_{q_{\neg j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]$ and $\mathcal{L}(q)$ are quite tedious – and apparently must be done separately for each model we make.

This **harms the applicability** of variational inference, even under the **quite restrictive** mean field assumption.

The Exponential Family of distributions

Why consider the Exponential Family

Positives with the Exponential Family:

- It is the only family of distributions with finite-sized sufficient statistics*;
- It is the only family of distributions that has conjugate priors;
- It simplifies the operations of variational inference;
- It has simple mathematical procedures for calculating moments, MLEs, Bayesian posteriors, ...

* Under certain regularity conditions...

Negatives:

- Standard distributions are defined using a new parameterization. Can be "unnatural".
- While all the calculations we will do are in principle simple, they are sometimes a bit abstract – due to the massive generality of the construction.

Definition – univariate exponential family model

Consider a univariate distribution $f_X(x \mid \boldsymbol{\theta})$, written as:

$$f_X(x \mid \boldsymbol{\theta}) = \exp\left(h(x) + \boldsymbol{\eta}(\boldsymbol{\theta})^\intercal \mathbf{t}(x) - A(\boldsymbol{\theta})\right)$$

Here we define:

- $h(x)$: log base measure
- $\boldsymbol{\eta}(\boldsymbol{\theta})$: the natural parameters
- $\mathbf{t}(x)$: the sufficient statistics
- $A(\boldsymbol{\theta})$: the log partition function

Definition – univariate exponential family model

Consider a univariate distribution $f_X(x \,|\, \boldsymbol{\theta})$, written as:

$$f_X(x \,|\, \boldsymbol{\theta}) = \exp\left(h(x) + \boldsymbol{\eta}(\boldsymbol{\theta})^\intercal \mathbf{t}(x) - A(\boldsymbol{\theta})\right)$$

Here we define:

- $h(x)$: log base measure
- $\boldsymbol{\eta}(\boldsymbol{\theta})$: the natural parameters
- $\mathbf{t}(x)$: the sufficient statistics
- $A(\boldsymbol{\theta})$: the log partition function

Members:

- Bernoulli (and Multinomial)

Example:

- $h(x) = 0$
- $\boldsymbol{\eta}(p) = \log(p/1-p)$
- $\mathbf{t}(x) = x$
- $A(p) = -\log(1-p)$

$$\begin{aligned} f_X(x \,|\, \boldsymbol{\theta}) &= \exp\left(h(x) + \boldsymbol{\eta}(\boldsymbol{\theta})^\intercal \mathbf{t}(x) - A(\boldsymbol{\theta})\right) \\ &= \exp\left\{\log\left[p/(1-p)\right] \cdot x + \log(1-p)\right\} \\ &= \left[p/(1-p)\right]^x \cdot (1-p) \\ &= p^x \,(1-p)^{1-x} \end{aligned}$$

The Bernoulli is in the exponential family

Definition – univariate exponential family model

Consider a univariate distribution $f_X(x \mid \boldsymbol{\theta})$, written as:

$$f_X(x \mid \boldsymbol{\theta}) = \exp\left(h(x) + \boldsymbol{\eta}(\boldsymbol{\theta})^\intercal \mathbf{t}(x) - A(\boldsymbol{\theta})\right)$$

Here we define:

- $h(x)$: log base measure
- $\boldsymbol{\eta}(\boldsymbol{\theta})$: the natural parameters
- $\mathbf{t}(x)$: the sufficient statistics
- $A(\boldsymbol{\theta})$: the log partition function

Members:

- Bernoulli (and Multinomial)
- Gaussians

Example:

- $h(x) = -\frac{1}{2}\log(2\pi)$
- $\boldsymbol{\eta}(\mu, \tau) = \left[\tau\mu, -\frac{\tau}{2}\right]^\intercal$
- $\mathbf{t}(x) = \left[x, x^2\right]^\intercal$
- $A(\boldsymbol{\theta}) = \frac{\tau\mu^2}{2} - \frac{1}{2}\log|\tau|$.

$$\begin{aligned} f_X(x \mid \boldsymbol{\theta}) &= \exp\left\{\left[\tau\mu, -\frac{\tau}{2}\right]\begin{bmatrix} x \\ x^2 \end{bmatrix} \right. \\ &\qquad \left. -\frac{1}{2}\log(2\pi) - (\tau\mu^2/2 - \frac{1}{2}\log\tau)\right\} \\ &= \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left(-\tau(x-\mu)^2/2\right) \end{aligned}$$

The Gaussian is in the exponential family

Definition – univariate exponential family model

Consider a univariate distribution $f_X(x \mid \boldsymbol{\theta})$, written as:

$$f_X(x \mid \boldsymbol{\theta}) = \exp\left(h(x) + \boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{t}(x) - A(\boldsymbol{\theta})\right)$$

Here we define:

- $h(x)$: log base measure
- $\boldsymbol{\eta}(\boldsymbol{\theta})$: the natural parameters
- $\mathbf{t}(x)$: the sufficient statistics
- $A(\boldsymbol{\theta})$: the log partition function

Members:

- Bernoulli (and Multinomial)
- Gaussians
- Gamma + Inverse Gamma
- ...

Example:

- $h(x) = 0$
- $\boldsymbol{\eta}(\alpha, \beta) = [-\beta, (\alpha - 1)]^{\mathsf{T}}$
- $\mathbf{t}(x) = [x, \log(x)]^{\mathsf{T}}$
- $A(\boldsymbol{\theta}) = \log(\Gamma(\alpha)) - \alpha\log(\beta)$

$$\begin{aligned} f_X(x \mid \theta) &= \exp\left(0 + [-\beta, (\alpha - 1)]\,[x, \log(x)]^{\mathsf{T}}\right. \\ &\qquad \left. + \alpha\log(\beta) - \log(\Gamma(\alpha))\right) \\ &= \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{(\alpha-1)} \exp(-\beta x) \end{aligned}$$

The Gamma is in the exponential family

Multivariate distributions

Consider a multi-variate distribution $f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})$, and assume it can be written as:

$$f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) = \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{t}(\mathbf{x}) - A(\boldsymbol{\theta})\right)$$

We define $h(\mathbf{x})$ (the log base measure) and $\mathbf{t}(\mathbf{x})$ (the sufficient statistics) as taking vector-inputs, but otherwise the definition is identical to the univariate case.

Example:
Consider a d-dimensional $\mathbf{x}$ with expectation $\boldsymbol{\mu}$ and inverse covariance $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$.

Define

- $h(\mathbf{x}) = -\frac{d}{2}\log(2\pi)$
- $\boldsymbol{\eta}(\boldsymbol{\mu}, \mathbf{Q}) = \left[\mathbf{Q}\boldsymbol{\mu}, -\frac{1}{2}\mathbf{Q}\right]^{\mathsf{T}}$
- $\mathbf{t}(\mathbf{x}) = \left[\mathbf{x}, \mathbf{x}\mathbf{x}^{\mathsf{T}}\right]^{\mathsf{T}}$
- $A(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{Q}\boldsymbol{\mu} - \frac{1}{2}\log|\mathbf{Q}|$.

This gives us the multivariate Gaussian distribution; $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{Q}^{-1})$.

Multivariate exponential family model

Multivariate distributions

Consider a multi-variate distribution $f_{\mathbf{X}}(\mathbf{x} \,|\, \boldsymbol{\theta})$, and assume it can be written as:

$$f_{\mathbf{X}}(\mathbf{x} \,|\, \boldsymbol{\theta}) = \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{t}(\mathbf{x}) - A(\boldsymbol{\theta})\right)$$

We define $h(\mathbf{x})$ (the log base measure) and $\mathbf{t}(\mathbf{x})$ (the sufficient statistics) as taking vector-inputs, but otherwise the definition is identical to the univariate case.

Notational simplification:

Notice how $\boldsymbol{\eta}$ take the role of the parameters $\boldsymbol{\theta}$, and – for the model we consider – a one-to-one mapping between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$.

For instance, $\boldsymbol{\eta} = [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2]^{\mathsf{T}} = \left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right]$ in the Gaussian distribution, meaning

$$\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]^{\mathsf{T}} = \left[-\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1, -\frac{1}{2}\boldsymbol{\eta}_2^{-1}\right]^{\mathsf{T}}.$$

Given that $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are interchangable, we will simplify notation, and use this form:

$$f_{\mathbf{X}}(\mathbf{x} \,|\, \boldsymbol{\eta}) = \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^{\mathsf{T}}\mathbf{t}(\mathbf{x}) - A(\boldsymbol{\eta})\right)$$

Code Task: Translation between moment-based and ExpFam representations

In this task you will translate between moment-based representations and the corresponding ExpFam representations. The task is fairly straight-forward, but is intended to give you "hands-on" experience with ExpFam representation.

Start from

`students_translator.ipynb`

You will see that there are some supporting functions there, and a translation for univariate Gaussians (to show how the supporting functions work).

Your task is to implement the same for the Gamma distribution. If time, other ExpFam distributions can of course be given the same treatment.

Remember the definition or the Exponential Family:

$$f_X(x \mid \boldsymbol{\eta}) = \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^\intercal \mathbf{t}(\mathbf{x}) - A(\boldsymbol{\eta})\right)$$

Derivatives of the log normalizer:

$\nabla^k A(\boldsymbol{\eta})$ has an interesting form:

- $\nabla A(\boldsymbol{\eta}) = \frac{\mathrm{d}A(\boldsymbol{\eta})}{\mathrm{d}\boldsymbol{\eta}} = \mathbb{E}\left[\mathbf{t}\left(\mathbf{X}\right)\right]$

Proof:

$$\begin{aligned}
\frac{\mathrm{d}A(\boldsymbol{\eta})}{\mathrm{d}\boldsymbol{\eta}} &= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\eta}} \log \int_{\mathbf{x}} \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^\intercal \mathbf{t}(\mathbf{x})\right) \mathrm{d}\mathbf{x} \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\eta}} \int_{\mathbf{x}} \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^\intercal \mathbf{t}(\mathbf{x})\right) \mathrm{d}\mathbf{x}}{\int_{\mathbf{x}} \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^\intercal \mathbf{t}(\mathbf{x})\right) \mathrm{d}\mathbf{x}} \\
&= \frac{\int_{\mathbf{x}} \mathbf{t}(\mathbf{x})\, \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^\intercal \mathbf{t}(\mathbf{x}))\right) \mathrm{d}\mathbf{x}}{\exp\left(A(\boldsymbol{\eta})\right)} \\
&= \int_{\mathbf{x}} \mathbf{t}(\mathbf{x})\, \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^\intercal \mathbf{t}(\mathbf{x}) - A(\boldsymbol{\eta})\right) \mathrm{d}\mathbf{x} = \mathbb{E}\left[\mathbf{t}\left(\mathbf{X}\right)\right]
\end{aligned}$$

Some nice features

Remember the definition or the Exponential Family:

$$f_X(x \,|\, \boldsymbol{\eta}) = \exp\left(h(\mathbf{x}) + \boldsymbol{\eta}^\intercal \mathbf{t}(\mathbf{x}) - A(\boldsymbol{\eta})\right)$$

Derivatives of the log normalizer:

$\nabla^k A(\boldsymbol{\eta})$ has an interesting form:

- $\nabla A(\boldsymbol{\eta}) = \frac{\mathrm{d}A(\boldsymbol{\eta})}{\mathrm{d}\boldsymbol{\eta}} = \mathbb{E}\left[\mathbf{t}\left(\mathbf{X}\right)\right]$
- $\nabla^2 A(\boldsymbol{\eta}) = \mathbb{E}\left[\left(\mathbf{t}\left(\mathbf{X}\right) - \mathbb{E}\left[\mathbf{t}\left(\mathbf{X}\right)\right]\right)^2\right] = \mathbb{V}\left[\mathbf{t}\left(\mathbf{X}\right)\right]$
  - ... which also shows that $A(\boldsymbol{\eta})$ is *convex*.
- $\nabla^k A(\boldsymbol{\eta}) = \mathbb{E}\left[\left(\mathbf{t}\left(\mathbf{X}\right) - \mathbb{E}\left[\mathbf{t}\left(\mathbf{X}\right)\right]\right)^k\right]$

Proofs for $k > 1$ are more of the same manipulations (left out for simplicity).