

Variational inference

Partly based on material developed together with Helge Langseth

Andrés Masegosa and Thomas Dyhre Nielsen

Day 1: Probabilistic programming

- Introduction to probabilistic programming
- Probabilistic programming in Pyro

Day 2: Variational inference

- Recap of variational inference (variational inference as optimization)
- Derivation and implementation of selected examples
 - Bayesian linear regression
 - Factor analysis
 - ...

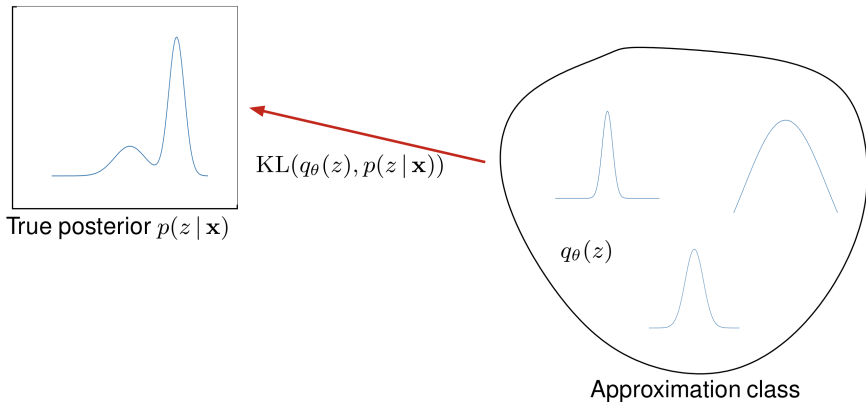
Day 3: Variational inference – cont'd

- Black box variational inference
- Variational inference in Pyro
- Variational auto-encoders

Introduction

What is variational inference?

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



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

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

We can rearrange the KL divergence as follows:

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

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

We can rearrange the KL divergence as follows:

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

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

VI focuses on the ELBO:

$$\log p(\mathbf{x}) = \mathcal{L}(q) + \text{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$$

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

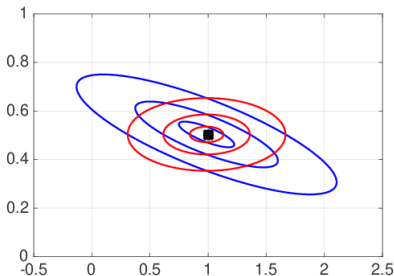
- We can minimize $\text{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$ by maximizing $\mathcal{L}(q)$
- This is **computationally simpler** because it uses $p(\mathbf{z}, \mathbf{x})$ instead of $p(\mathbf{z}|\mathbf{x})$.
- $\mathcal{L}(q)$ is a *lower bound* of the marginal data log likelihood $\log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

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

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

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



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.

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 KL ($q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})$) as our objective.

Algorithm:

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

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

Solving the VB optimization

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.

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

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.

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

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

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

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.

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

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.

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

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

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

because $\mathbb{E}_{q_{\neg j}} [\log q_{\neg j}(\mathbf{z}_{\neg j})]$ 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.

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

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.

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

We get the following result:

The ELBO is maximized wrt. q_j by choosing

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

... 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 KL $(q(\mathbf{z}) || p(\mathbf{z} | \mathbf{x}))$ as our objective.
- We posit a *variational family* of distributions $q_j(z_j | \lambda_j)$, i.e., we choose the distributional form, while wanting to **optimize the parameterization λ_j** .
- The optimal λ_j **will** depend on \mathbf{x} – in fact λ_j encodes all the information about the other variables in the domain that Z_j is “aware of”.

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 KL ($q(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})$) as our objective.
- We posit a *variational family* of distributions $q_j(z_j | \lambda_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization λ_j .
- The optimal λ_j **will** depend on \mathbf{x} – in fact λ_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 $\mathcal{L}(q)$:

- For each j :
 - Calculate $\mathbb{E}_{q_{-j}} [\log p(\mathbf{z}, \mathbf{x})]$ using current estimates for $q_i(\cdot | \lambda_i)$, $i \neq j$.
 - Choose λ_j so that $q_j(z_j | \lambda_j) \propto \exp(\mathbb{E}_{q_{-j}} [\log p(\mathbf{z}, \mathbf{x})])$.
- Calculate the new $\mathcal{L}(q)$.

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 $q_i(\cdot | \boldsymbol{\lambda}_i)$, $i \neq j$.
 - Choose $\boldsymbol{\lambda}_j$ so that $q_j(z_j | \boldsymbol{\lambda}_j) \propto \exp(\mathbb{E}_{q_{\neg j}} [\log p(\mathbf{z}, \mathbf{x})])$.
- Calculate the new $\mathcal{L}(q)$.

Calculating $q_j(z_j | \boldsymbol{\lambda}_j)$ - an observation

The update-rule can equivalently be expressed as

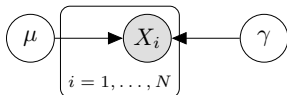
$$\begin{aligned} \log q_j(z_j | \boldsymbol{\lambda}_j) &= \mathbb{E}_{q_{\neg j}} [\ln p(\mathbf{z}, \mathbf{x})] + c. \\ &= \sum_{x \in \text{mb}(z_j)} \mathbb{E}_{q_{\neg j}} \log p(x | \text{pa}(x)) + \sum_{z \in \text{mb}(z_j)} \mathbb{E}_{q_{\neg j}} \log p(z | \text{pa}(z)) + c'. \end{aligned}$$

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!

A simple Gaussian model

A Gaussian model with unknown mean and precision

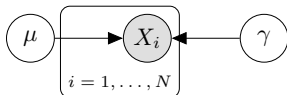


- $X_i \mid \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau^{-1})$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

The probability model

$$p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) = \prod_{i=1}^N p(x_i | \mu, \gamma^{-1}) p(\mu | 0, \tau^{-1}) p(\gamma | \alpha, \beta)$$

A Gaussian model with unknown mean and precision



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

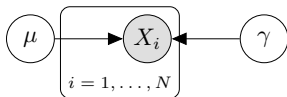
The probability model

$$p(\mathbf{x}, \mu, \gamma \mid \tau, \alpha, \beta) = \prod_{i=1}^N p(x_i \mid \mu, \gamma^{-1}) p(\mu \mid 0, \tau^{-1}) p(\gamma \mid \alpha, \beta)$$

...after taking the log

$$\log p(\mathbf{x}, \mu, \gamma \mid \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i \mid \mu, \gamma^{-1}) + \log p(\mu \mid 0, \tau^{-1}) + \log p(\gamma \mid \alpha, \beta)$$

A Gaussian model with unknown mean and precision



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

The probability model

$$p(\mathbf{x}, \mu, \gamma \mid \tau, \alpha, \beta) = \prod_{i=1}^N p(x_i \mid \mu, \gamma^{-1}) p(\mu \mid 0, \tau^{-1}) p(\gamma \mid \alpha, \beta)$$

The variational model (full mean field)

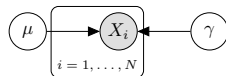
$$q(\mu, \gamma) = q(\mu)q(\gamma),$$

where

- $q(\mu) = \mathcal{N}(\nu_p, \tau_p^{-1})$
- $q(\gamma) = \text{Gamma}(\alpha_p, \beta_p)$

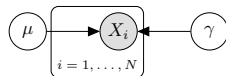
We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c =$$



We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c =$$

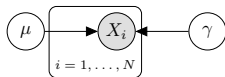


Recall

$$\log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu | 0, \tau^{-1}) + \log p(\gamma | \alpha, \beta)$$

We choose the variational distribution so that

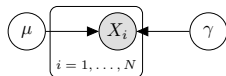
$$\log q(\mu) = \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c =$$



Recall

$$\log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu | 0, \tau^{-1}) + \log p(\gamma | \alpha, \beta)$$

We choose the variational distribution so that



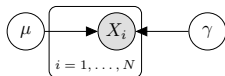
$$\log q(\mu) = \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c$$

Recall

$$\log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu | 0, \tau^{-1}) + \log p(\gamma | \alpha, \beta)$$

We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c$$



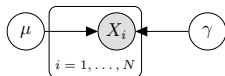
Recall

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

$$\log p(\mu | 0, \tau^{-1}) = \mathcal{N}(0, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (\mu)^2$$

We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c$$

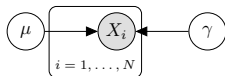


Recall

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

$$\log p(\mu | 0, \tau^{-1}) = \mathcal{N}(0, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (\mu)^2$$

We choose the variational distribution so that

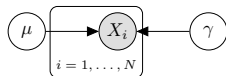


$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c \\ &= \sum_{i=1}^N \mathbb{E}_{\gamma} \left(-\frac{\gamma}{2} (x_i - \mu)^2 \right) - \frac{\tau}{2} (\mu)^2 + c\end{aligned}$$

Recall

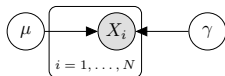
$$\begin{aligned}\log p(x_i | \mu, \gamma^{-1}) &= \mathcal{N}(\mu, \gamma^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma) - \frac{\gamma}{2} (x_i - \mu)^2 \\ \log p(\mu | 0, \tau^{-1}) &= \mathcal{N}(0, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (\mu)^2\end{aligned}$$

We choose the variational distribution so that



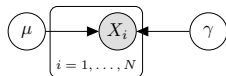
$$\begin{aligned}
 \log q(\mu) &= \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c \\
 &= \sum_{i=1}^N \mathbb{E}_{\gamma} \left(-\frac{\gamma}{2} (x_i - \mu)^2 \right) - \frac{\tau}{2} (\mu)^2 + c \\
 &= -\frac{1}{2} \mathbb{E}_{\gamma}(\gamma) \left(\sum_{i=1}^N x_i^2 + N \cdot \mu^2 - 2\mu \sum_{i=1}^N x_i \right) - \frac{\tau}{2} (\mu)^2 + c
 \end{aligned}$$

We choose the variational distribution so that



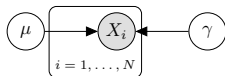
$$\begin{aligned}
 \log q(\mu) &= \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c \\
 &= \sum_{i=1}^N \mathbb{E}_{\gamma} \left(-\frac{\gamma}{2} (x_i - \mu)^2 \right) - \frac{\tau}{2} (\mu)^2 + c \\
 &= -\frac{1}{2} \mathbb{E}_{\gamma}(\gamma) \left(\sum_{i=1}^N x_i^2 + N \cdot \mu^2 - 2\mu \sum_{i=1}^N x_i \right) - \frac{\tau}{2} (\mu)^2 + c \\
 &= -\frac{1}{2} \left(\mathbb{E}_{\gamma}(\gamma) \cdot N + \tau \right) \mu^2 + \left(\mathbb{E}_{\gamma}(\gamma) \sum_{i=1}^N x_i \right) \mu + c
 \end{aligned}$$

We choose the variational distribution so that



$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c \\ &= -\frac{1}{2} \left(\mathbb{E}_{\gamma}(\gamma) \cdot N + \tau \right) \mu^2 + \left(\mathbb{E}_{\gamma}(\gamma) \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$

We choose the variational distribution so that

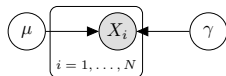


$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c \\ &= -\frac{1}{2} \left(\mathbb{E}_{\gamma}(\gamma) \cdot N + \tau \right) \mu^2 + \left(\mathbb{E}_{\gamma}(\gamma) \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$

Recall the normal distribution

$$\begin{aligned}\log q(\mu | \nu_p, \tau_p^{-1}) &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_p) - \frac{\tau_p}{2} (\mu - \nu_p)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_p) - \tau_p \nu_p^2 - \frac{1}{2} \tau_p \mu^2 + \tau_p \nu_p \mu\end{aligned}$$

We choose the variational distribution so that

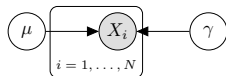


$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c \\ &= -\frac{1}{2} \left(\mathbb{E}_{\gamma}(\gamma) \cdot N + \tau \right) \mu^2 + \left(\mathbb{E}_{\gamma}(\gamma) \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$

Recall the normal distribution

$$\begin{aligned}\log q(\mu | \nu_p, \tau_p^{-1}) &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_p) - \frac{\tau_p}{2} (\mu - \nu_p)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_p) - \tau_p \nu_p^2 - \frac{1}{2} \tau_p \mu^2 + \tau_p \nu_p \mu\end{aligned}$$

We choose the variational distribution so that



$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{\gamma} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\gamma} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\mu | 0, \tau^{-1}) + c \\ &= -\frac{1}{2} \left(\mathbb{E}_{\gamma}(\gamma) \cdot N + \tau \right) \mu^2 + \left(\mathbb{E}_{\gamma}(\gamma) \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$

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

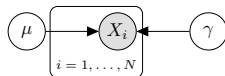
- precision $\tau_p \leftarrow \mathbb{E}_{\gamma}(\gamma) \cdot N + \tau$
- mean $\nu_p \leftarrow \tau_p^{-1} \left(\mathbb{E}_{\gamma}(\gamma) \sum_{i=1}^N x_i \right)$

Recall the normal distribution

$$\begin{aligned}\log q(\mu | \nu_p, \tau_p^{-1}) &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_p) - \frac{\tau_p}{2} (\mu - \nu_p)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_p) - \tau_p \nu_p^2 - \frac{1}{2} \tau_p \mu^2 + \tau_p \nu_p \mu\end{aligned}$$

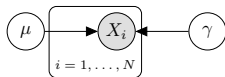
We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c =$$



We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c =$$

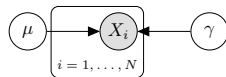


Recall

$$\log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu | 0, \tau^{-1}) + \log p(\gamma | \alpha, \beta)$$

We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c =$$

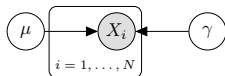


Recall

$$\log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu | 0, \tau^{-1}) + \log p(\gamma | \alpha, \beta)$$

We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c$$

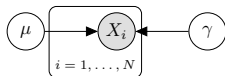


Recall

$$\log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu | 0, \tau^{-1}) + \log p(\gamma | \alpha, \beta)$$

We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c$$



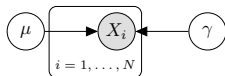
Recall

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

$$\log p(\gamma | \alpha, \beta) = \text{Gamma}(\alpha, \beta) = \alpha \cdot \log(\beta) + (\alpha - 1) \log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha))$$

We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c$$

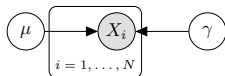


Recall

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

$$\log p(\gamma | \alpha, \beta) = \text{Gamma}(\alpha, \beta) = \alpha \cdot \log(\beta) + (\alpha - 1) \log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha))$$

We choose the variational distribution so that

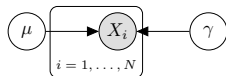


$$\begin{aligned}\log q(\gamma) &= \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c \\ &= \frac{N}{2} \log(\gamma) - \frac{\gamma}{2} \sum_{i=1}^N \mathbb{E}_{\mu} (x_i - \mu)^2 + (\alpha - 1) \log(\gamma) - \beta \cdot \gamma + c\end{aligned}$$

Recall

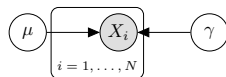
$$\begin{aligned}\log p(x_i | \mu, \gamma^{-1}) &= \mathcal{N}(\mu, \gamma^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma) - \frac{\gamma}{2} (x_i - \mu)^2 \\ \log p(\gamma | \alpha, \beta) &= \text{Gamma}(\alpha, \beta) = \alpha \cdot \log(\beta) + (\alpha - 1) \log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha))\end{aligned}$$

We choose the variational distribution so that



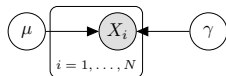
$$\begin{aligned}
 \log q(\gamma) &= \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c \\
 &= \frac{N}{2} \log(\gamma) - \frac{\gamma}{2} \sum_{i=1}^N \mathbb{E}_{\mu} (x_i - \mu)^2 + (\alpha - 1) \log(\gamma) - \beta \cdot \gamma + c \\
 &= \left(\frac{N}{2} + \alpha - 1 \right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^N \mathbb{E}_{\mu} (x_i - \mu)^2 + \beta \right) \cdot \gamma + c
 \end{aligned}$$

We choose the variational distribution so that



$$\begin{aligned} \log q(\gamma) &= \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c \\ &= \left(\frac{N}{2} + \alpha - 1 \right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^N \mathbb{E}_{\mu} (x_i - \mu)^2 + \beta \right) \cdot \gamma + c \end{aligned}$$

We choose the variational distribution so that

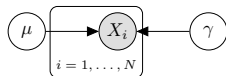


$$\begin{aligned}\log q(\gamma) &= \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c \\ &= \left(\frac{N}{2} + \alpha - 1 \right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^N \mathbb{E}_{\mu} (x_i - \mu)^2 + \beta \right) \cdot \gamma + c\end{aligned}$$

Recall

$$\log q(\gamma | \alpha_p, \beta_p) = \alpha_p \cdot \log(\beta_p) + (\alpha_p - 1) \log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha_p))$$

We choose the variational distribution so that

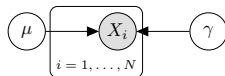


$$\begin{aligned}\log q(\gamma) &= \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c \\ &= \left(\frac{N}{2} + \alpha - 1 \right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^N \mathbb{E}_{\mu} (x_i - \mu)^2 + \beta \right) \cdot \gamma + c\end{aligned}$$

Recall

$$\log q(\gamma | \alpha_p, \beta_p) = \alpha_p \cdot \log(\beta_p) + (\alpha_p - 1) \log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha_p))$$

We choose the variational distribution so that



$$\begin{aligned} \log q(\gamma) &= \mathbb{E}_{\mu} \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) + c = \sum_{i=1}^N \mathbb{E}_{\mu} (\log p(x_i | \mu, \gamma^{-1})) + \log p(\gamma | \alpha, \beta) + c \\ &= \left(\frac{N}{2} + \alpha - 1 \right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^N \mathbb{E}_{\mu} (x_i - \mu)^2 + \beta \right) \cdot \gamma + c \end{aligned}$$

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

- $\alpha_p \leftarrow \frac{N}{2} + \alpha$
- $\beta_p \leftarrow \beta + \frac{1}{2} \sum_{i=1}^N \mathbb{E}_q(x_i - \mu)^2$

Note that:

- $\mathbb{E}_q(x_i - \mu)^2 = x_i^2 + \mathbb{E}_q(\mu^2) - 2 \cdot x_i \cdot \mathbb{E}_q(\mu)$
- $\mathbb{E}_q(\mu^2) = \text{Var}(\mu) + \mathbb{E}_q(\mu)^2$

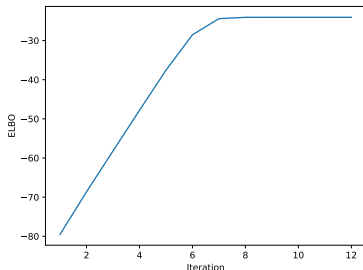
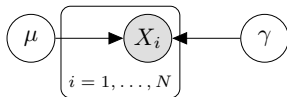
The variational updating rules are guaranteed to never decrease the ELBO $\mathcal{L}(q)$:

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q \log p(\mathbf{x}, \mu, \gamma \mid \tau, \alpha, \beta) - \mathbb{E}_q \log q(\mu, \gamma) \\ &= \sum_{i=1}^N \mathbb{E}_q \log p(x_i \mid \mu, \gamma) + \mathbb{E}_q \log p(\mu \mid 0, \tau) + \mathbb{E}_q \log p(\gamma \mid \alpha, \beta) - \mathbb{E}_q \log q(\mu) - \mathbb{E}_q \log q(\gamma)\end{aligned}$$

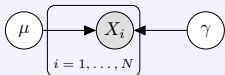
at any updating step. With some pencil pushing we arrive at a somewhat complicated but closed form expression (not show here).

Monitoring the ELBO can be useful for

- Assessing convergence
- Doing debugging
- ...



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 $q(\mu, \gamma) = q(\mu) \cdot q(\gamma)$ that best approximates $p(\mu, \gamma \mid x_1, \dots, x_N)$ wrt. the VB measure $\text{KL}(q \parallel p)$.

- Go through the notebook

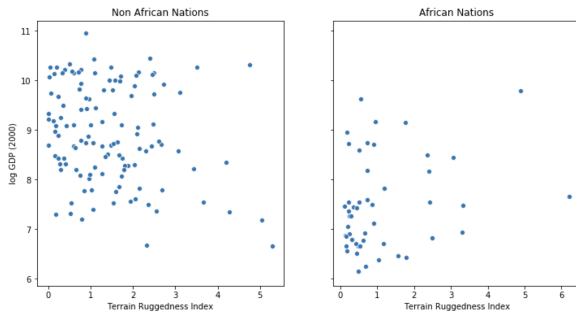
`students_simple_model.ipynb`

and try to link the code to the derivations in the slides.

- Implement the update rules for $q(\mu)$ and $q(\gamma)$ (from the slides) in the notebook.
- Experiment with the model and the data set; try changing the prior and the data generating process.

Bayesian linear regression

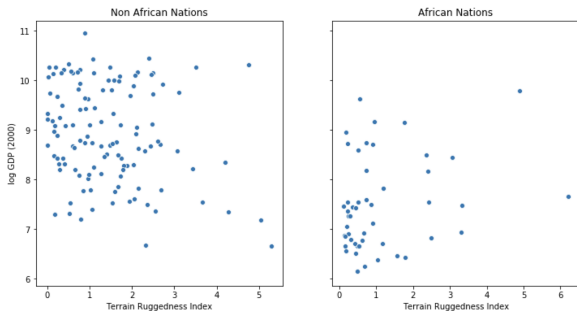
Real Data Example



Relationship between topographic heterogeneity and GDP per capita

- Terrain ruggedness or bad geography is related to poorer economic performance outside of Africa.

Real Data Example

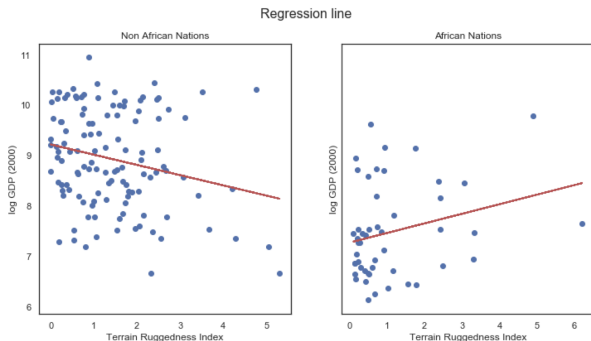


Relationship between topographic heterogeneity and GDP per capita

- Terrain ruggedness or bad geography is related to poorer economic performance outside of Africa.
- Rugged terrains have had a reverse effect on income for African nations.

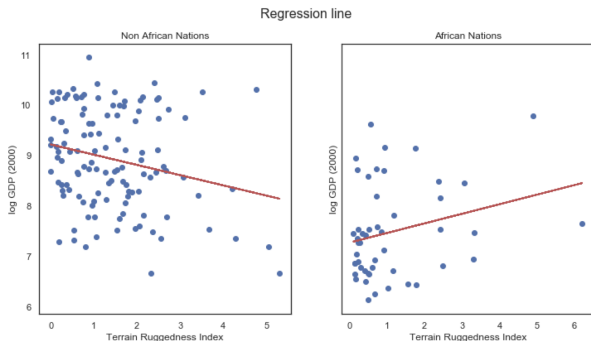
Day2/students_bayesian_regression_VI.ipynb

Real Data Example



Linear Regression Model

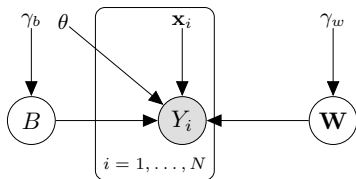
- Negative slope for Non African Nations.
- Positive slope for African Nations.



Bayesian Linear Regression Model

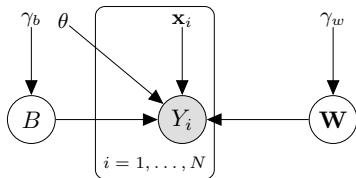
- Modeling data noise (aleatoric uncertainty). To be addressed tomorrow.
- Modeling uncertainty about the linear coefficients (epistemic uncertainty).

The Bayesian linear regression model



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

The Bayesian linear regression model

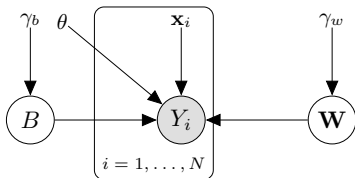


- Num. of data dim: M
- Num. of data inst: N
- $Y_i \mid \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

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

The Bayesian linear regression model



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

The probability model

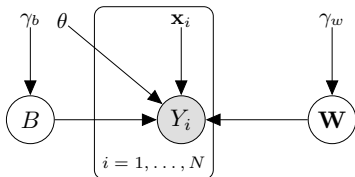
$$p(\cdot \mid \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^N p(y_i \mid \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} \mid \gamma_w) p(b \mid \gamma_b)$$

The variational updating rules (full mean field) - with some pencil pushing

$q(w_j)$ is normally distributed with

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

The Bayesian linear regression model



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

The probability model

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

The variational updating rules (full mean field) - with some pencil pushing

$q(b)$ is normally distributed with

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

Exercise: Implement the updating rules

Now it is your turn!

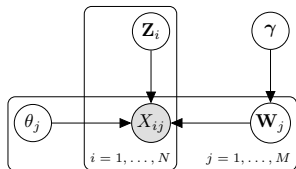
- Implement the updating rules in the notebook

`students_bayesian_regression_VI.ipynb`

- Play around with the code

Factor analysis

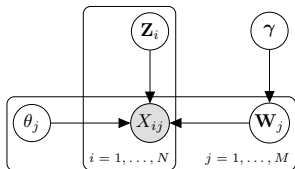
The factor analysis model



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

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

The factor analysis model



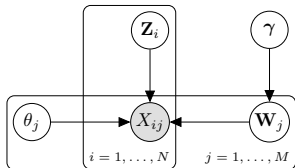
- $X_{ij} \mid \{\mathbf{w}_j, \mathbf{z}_i, \theta_j\} \sim \mathcal{N}(\mathbf{w}_j^\top \mathbf{z}_i, 1/\theta_j)$
- $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D \times D})$
- $\mathbf{W}_j \sim \mathcal{N}(\mathbf{0}, \gamma^{-1} \mathbf{I}_{D \times D})$
- $\theta_j \sim \text{Gamma}(\alpha_\theta,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

- 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 factor analysis model



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

- 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

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

The updating rule for $q(\gamma)$

By choosing the variational distribution so that

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

we find that $q(\gamma | \cdot)$ is gamma distributed with

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

The updating rule for $q(\gamma)$

By choosing the variational distribution so that

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

we find that $q(\gamma | \cdot)$ is gamma distributed with

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

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

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

The updating rule for $q(\gamma)$

By choosing the variational distribution so that

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

we find that $q(\gamma | \cdot)$ is gamma distributed with

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

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

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

Compare this to the Gibbs sampler:

$$\alpha \leftarrow \alpha_\gamma + \frac{DM}{2} \quad \beta \leftarrow \beta_\gamma + \frac{1}{2} \sum_{j=1}^M \mathbf{w}_j^T \mathbf{w}_j$$

VB uses posterior expectations where Gibbs uses samples!

By choosing the variational distribution so that

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

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

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

By choosing the variational distribution so that

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

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

- precision $\mathbf{Q} \leftarrow \mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j) \sum_{i=1}^N \mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top)$
- 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 $\mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top)$

$$\mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top) = \text{Cov}(\mathbf{Z}_i) + \mathbb{E}(\mathbf{Z}_i) \mathbb{E}(\mathbf{Z}_i)^\top$$

The variational updating rules

By choosing the variational distribution so that

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

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

- precision $\mathbf{Q} \leftarrow \mathbb{E}(\gamma)\mathbf{I} + \mathbb{E}(\theta_j) \sum_{i=1}^N \mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top)$
- 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 $\mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top)$

$$\mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top) = \text{Cov}(\mathbf{Z}_i) + \mathbb{E}(\mathbf{Z}_i) \mathbb{E}(\mathbf{Z}_i)^\top$$

Compare this to the Gibbs sampler:

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

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

Exercise

- Play around with the notebook

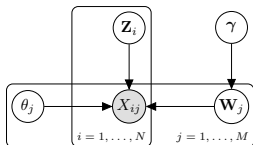
FA.ipynb

- Compare the implementation to the updating rules on the slides.
- Do you see any structure in the latent space (try with different seed values)
- Try investigating the learned distributions.

Data

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

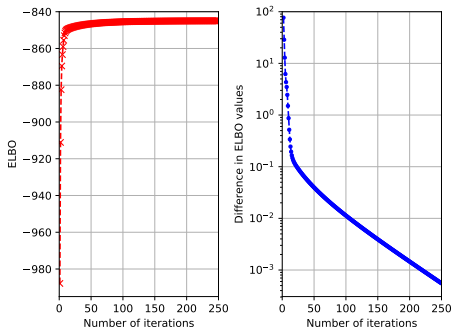
Global model



Local model



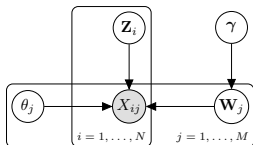
Monitoring convergence



Data

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

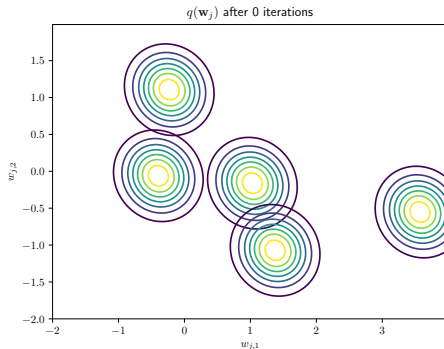
Global model



Local model



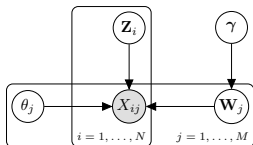
Variational posteriors



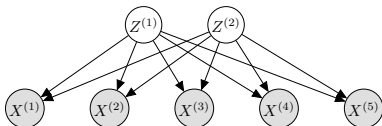
Data

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

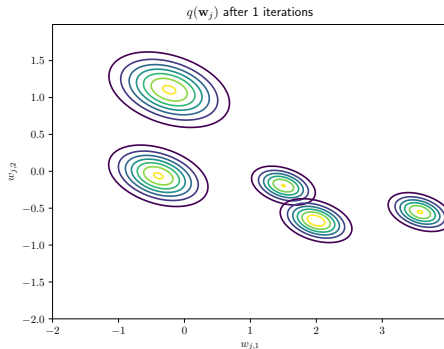
Global model



Local model



Variational posteriors



Data

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

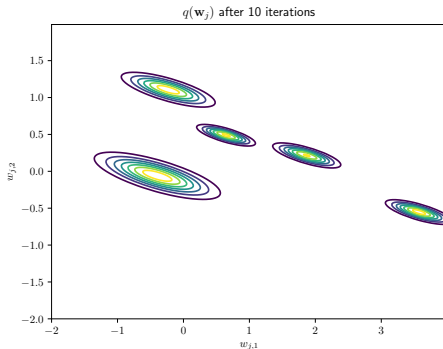
Global model



Local model



Variational posteriors



Data

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

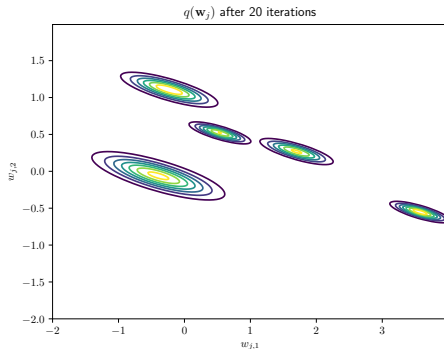
Global model



Local model



Variational posteriors



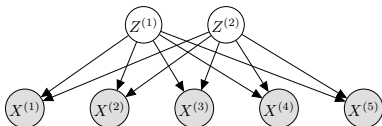
Data

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

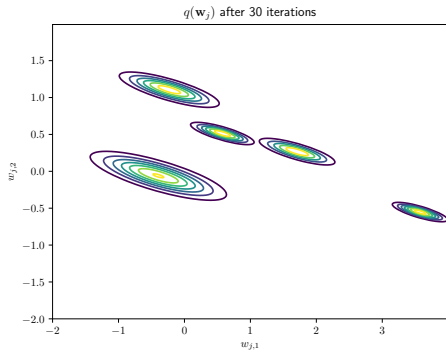
Global model



Local model



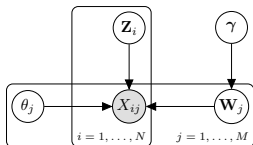
Variational posteriors



Data

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

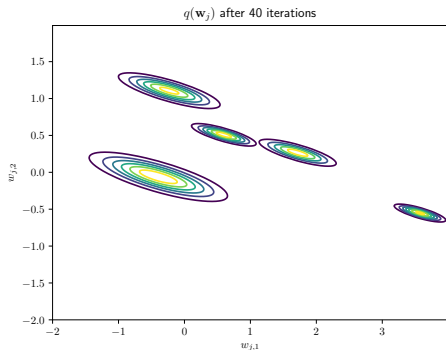
Global model



Local model



Variational posteriors



Data

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

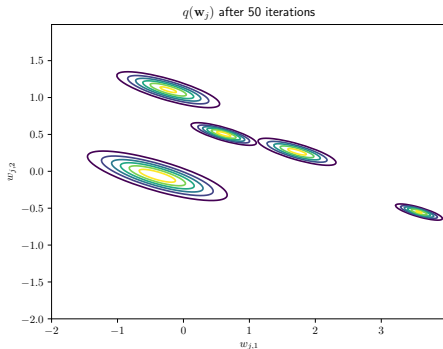
Global model



Local model



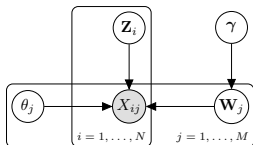
Variational posteriors



Data

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

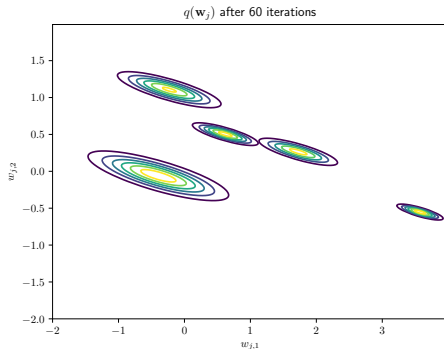
Global model



Local model



Variational posteriors



Data

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

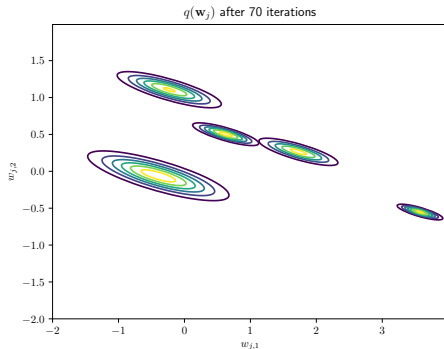
Global model



Local model



Variational posteriors



Data

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

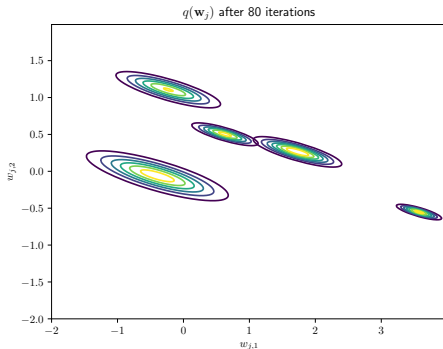
Global model



Local model



Variational posteriors



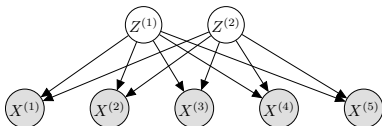
Data

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

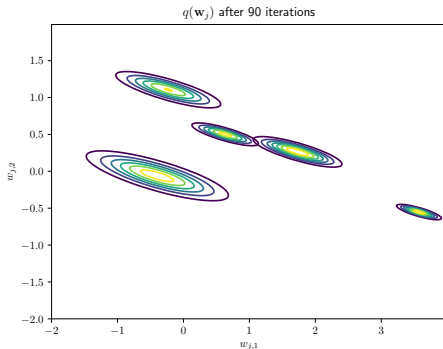
Global model



Local model



Variational posteriors



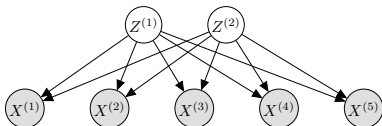
Data

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

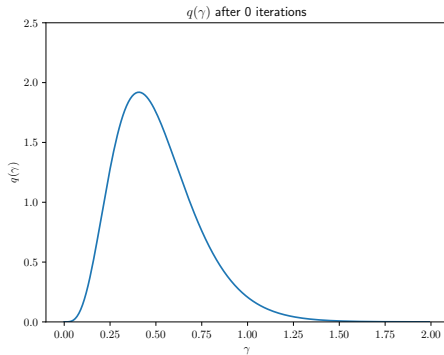
Global model



Local model



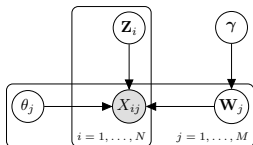
Variational posteriors



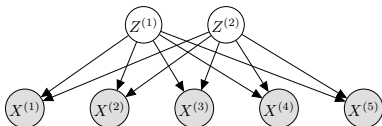
Data

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

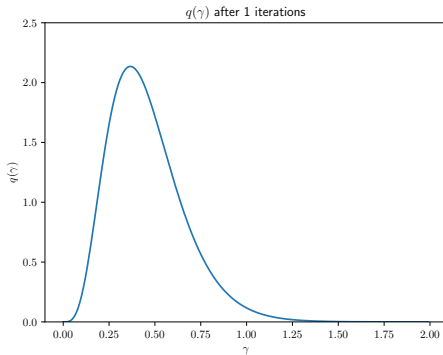
Global model



Local model



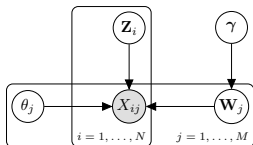
Variational posteriors



Data

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

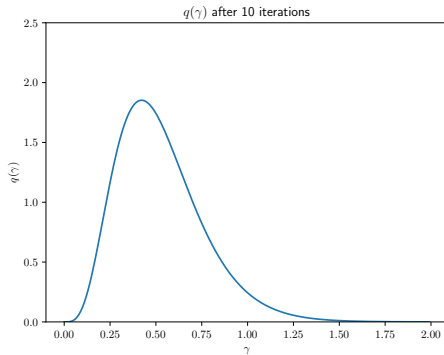
Global model



Local model



Variational posteriors



Data

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

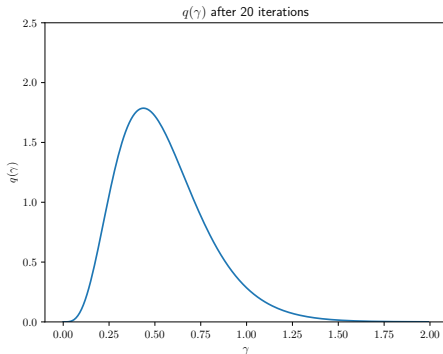
Global model



Local model



Variational posteriors



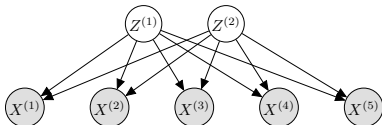
Data

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

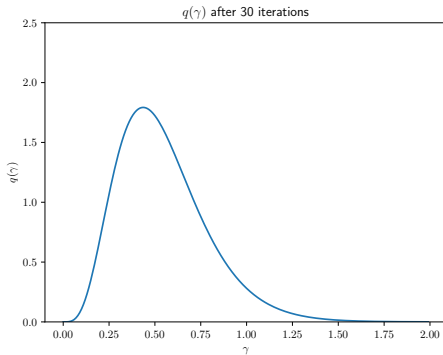
Global model



Local model



Variational posteriors



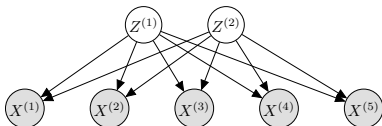
Data

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

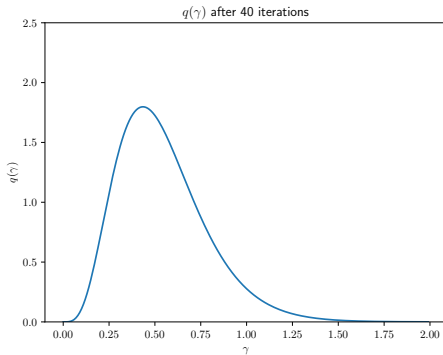
Global model



Local model



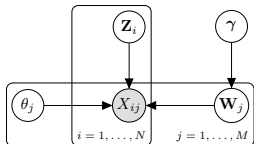
Variational posteriors



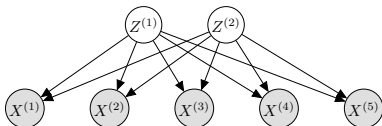
Data

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

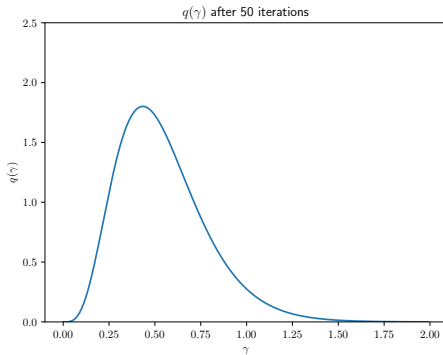
Global model



Local model



Variational posteriors



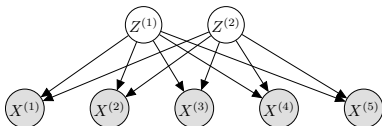
Data

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

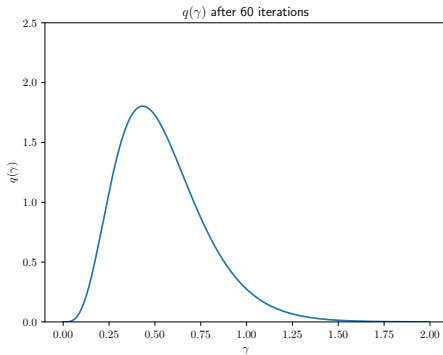
Global model



Local model



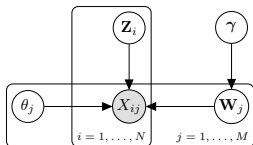
Variational posteriors



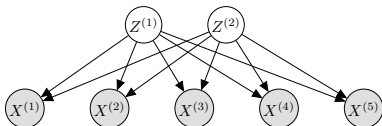
Data

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

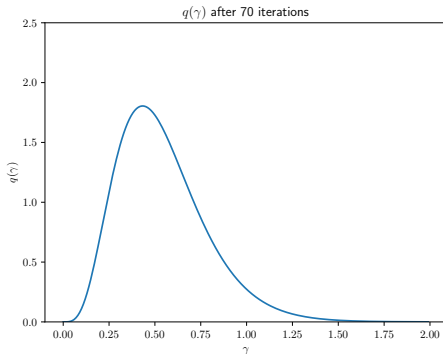
Global model



Local model



Variational posteriors



Data

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

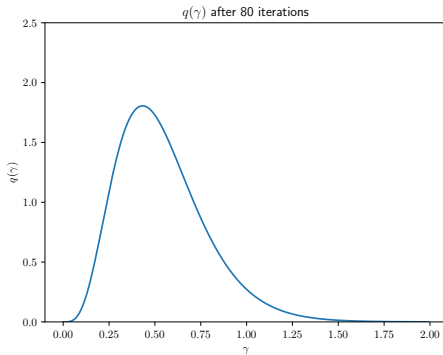
Global model



Local model



Variational posteriors



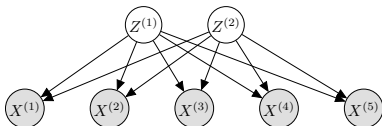
Data

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

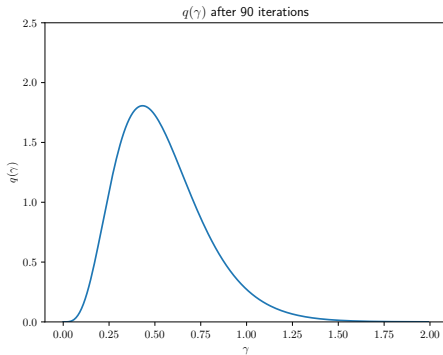
Global model



Local model



Variational posteriors



Data

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

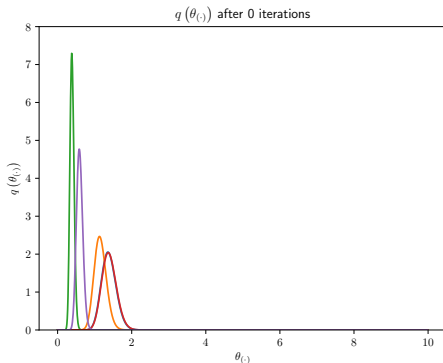
Global model



Local model



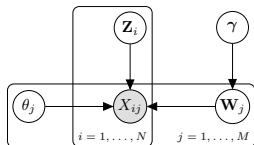
Variational posteriors



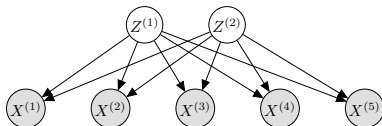
Data

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

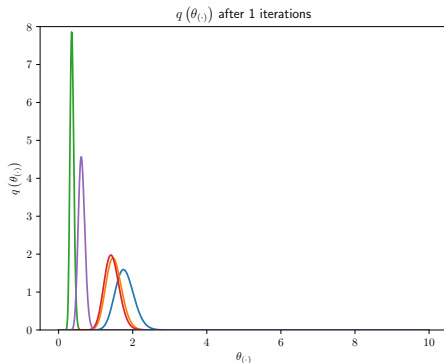
Global model



Local model



Variational posteriors



Data

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

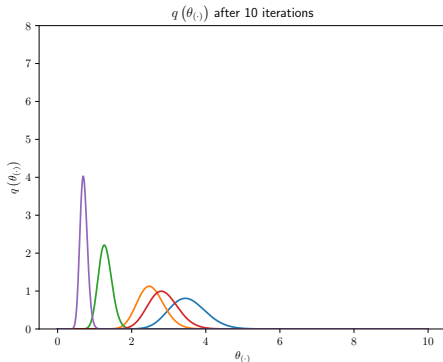
Global model



Local model



Variational posteriors



Data

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

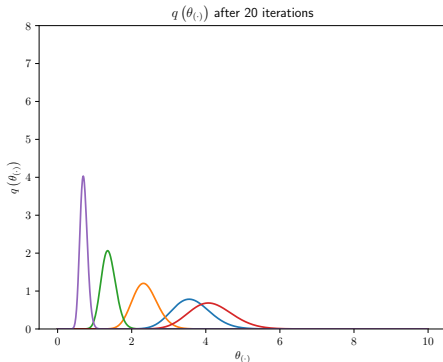
Global model



Local model



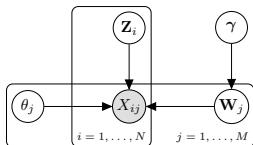
Variational posteriors



Data

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

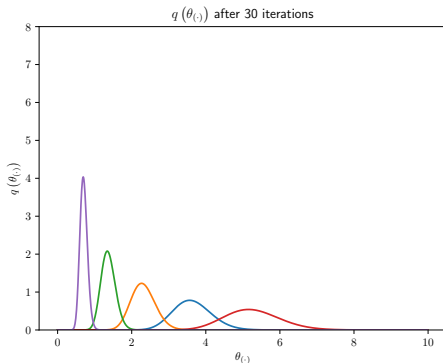
Global model



Local model



Variational posteriors



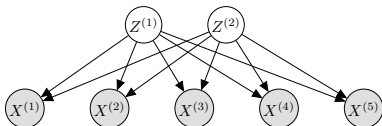
Data

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

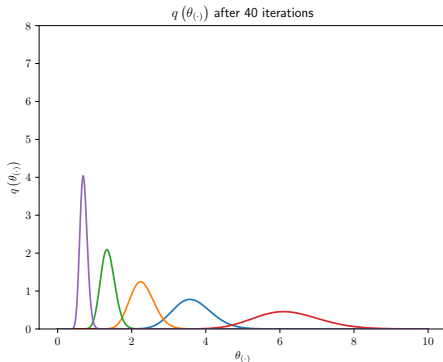
Global model



Local model



Variational posteriors



Data

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

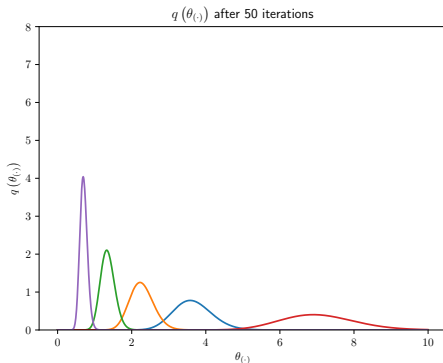
Global model



Local model



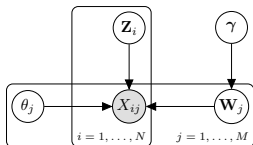
Variational posteriors



Data

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

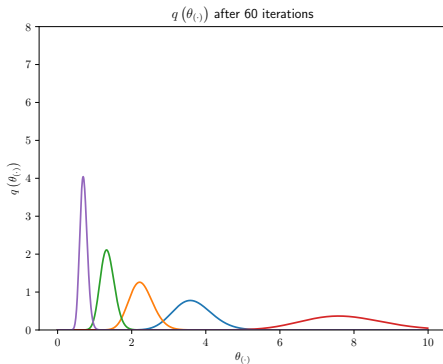
Global model



Local model



Variational posteriors



Data

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

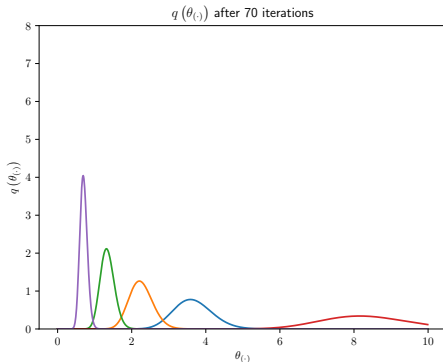
Global model



Local model



Variational posteriors



Data

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

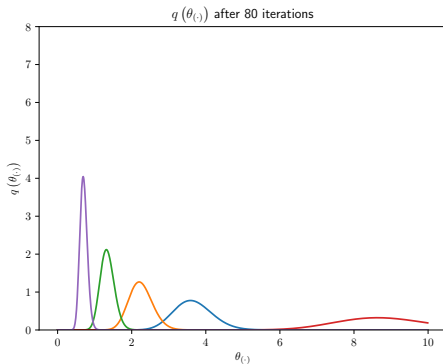
Global model



Local model



Variational posteriors



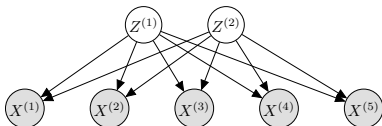
Data

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

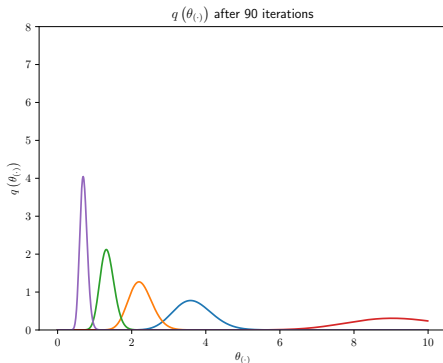
Global model



Local model



Variational posteriors



Data

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

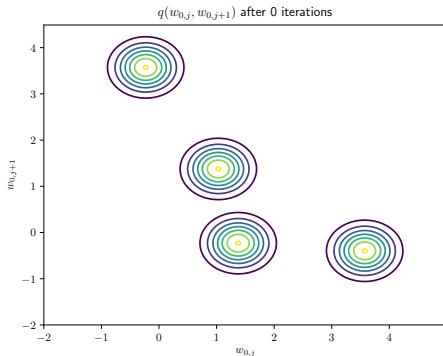
Global model



Local model



Variational posteriors



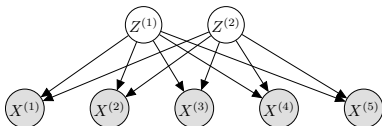
Data

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

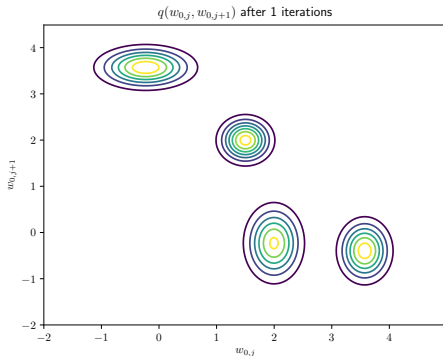
Global model



Local model



Variational posteriors



Data

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

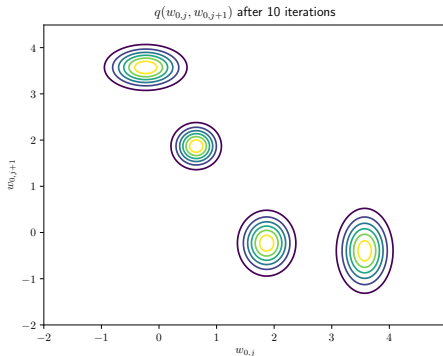
Global model



Local model



Variational posteriors



Data

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

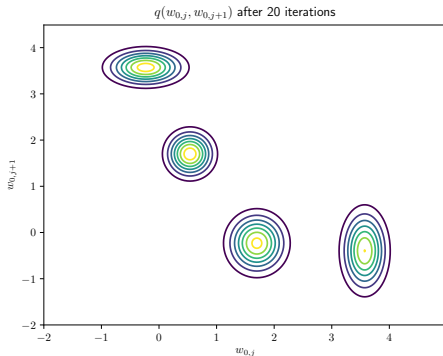
Global model



Local model



Variational posteriors



Data

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

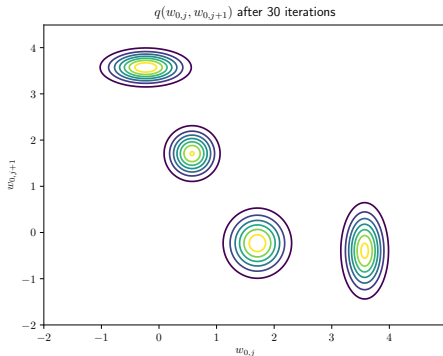
Global model



Local model



Variational posteriors



Data

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

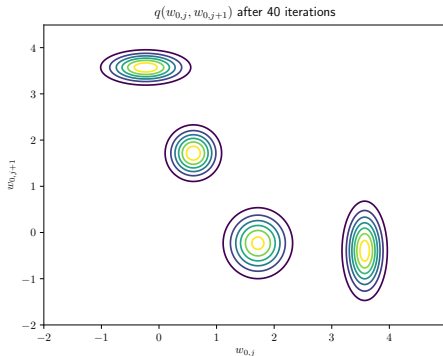
Global model



Local model



Variational posteriors



Data

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

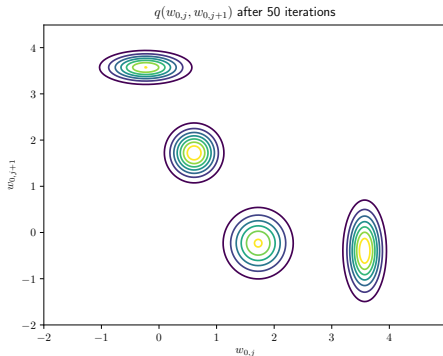
Global model



Local model



Variational posteriors



Data

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

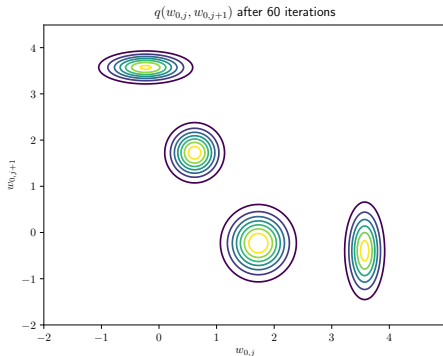
Global model



Local model



Variational posteriors



Data

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

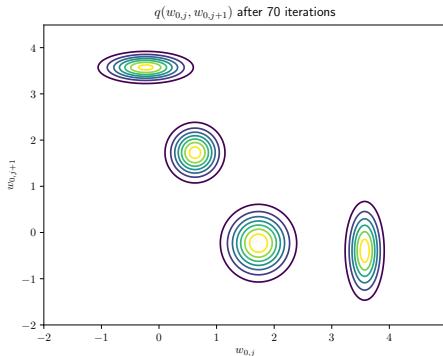
Global model



Local model



Variational posteriors



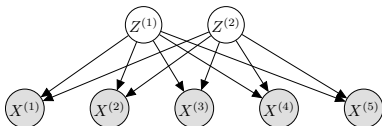
Data

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

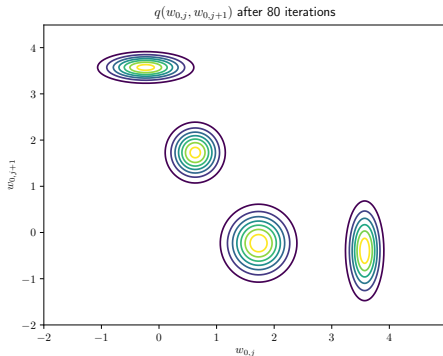
Global model



Local model



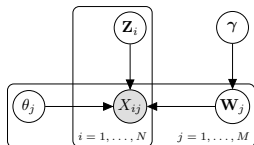
Variational posteriors



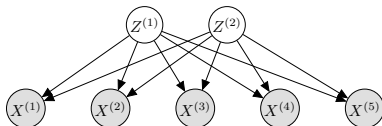
Data

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

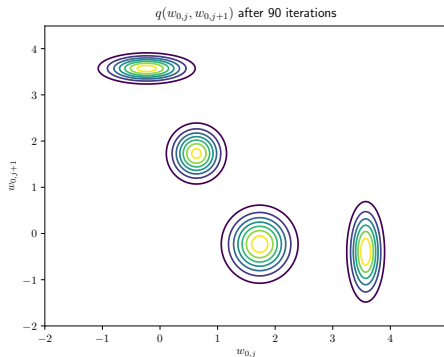
Global model



Local model



Variational posteriors



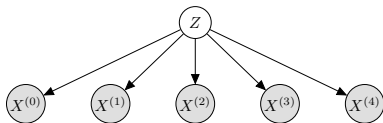
Data

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

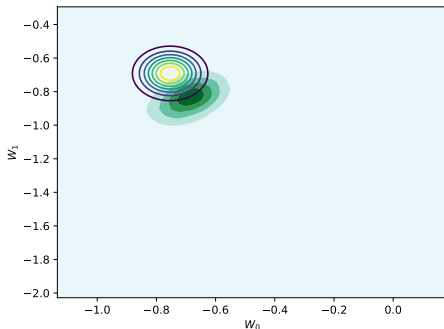
Global model



Local model



Comparison with Gibbs sampling



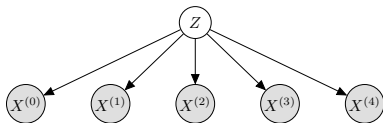
Data

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

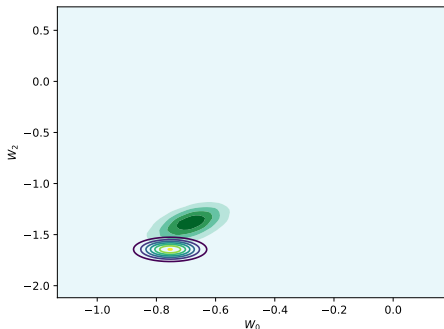
Global model



Local model



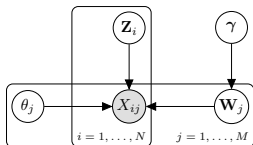
Comparison with Gibbs sampling



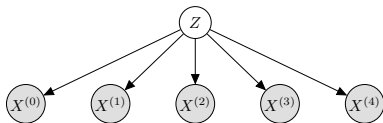
Data

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

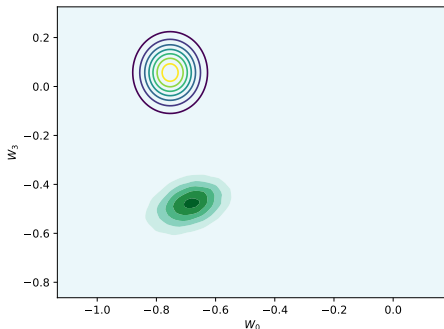
Global model



Local model



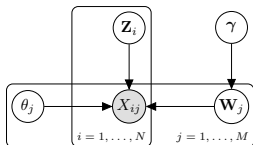
Comparison with Gibbs sampling



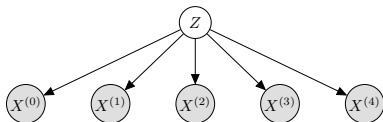
Data

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

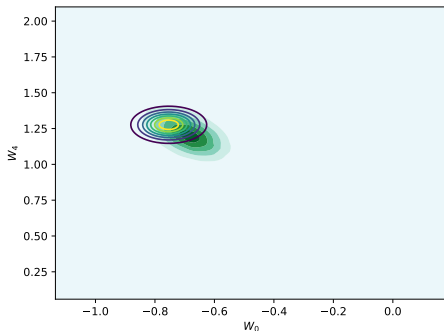
Global model



Local model



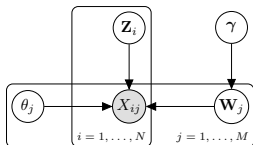
Comparison with Gibbs sampling



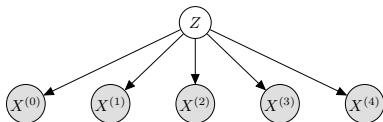
Data

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

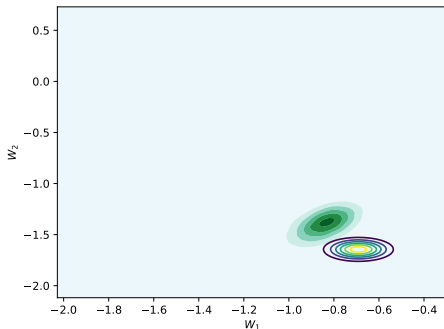
Global model



Local model



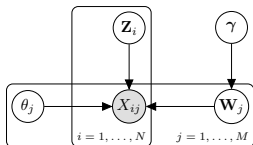
Comparison with Gibbs sampling



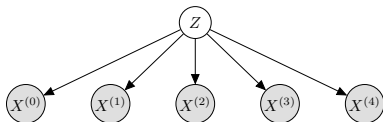
Data

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

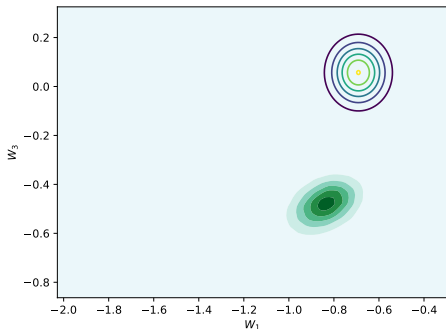
Global model



Local model



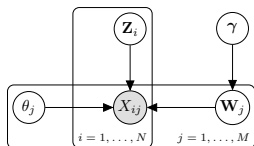
Comparison with Gibbs sampling



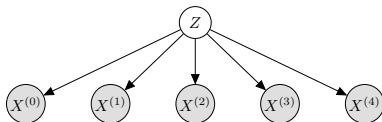
Data

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

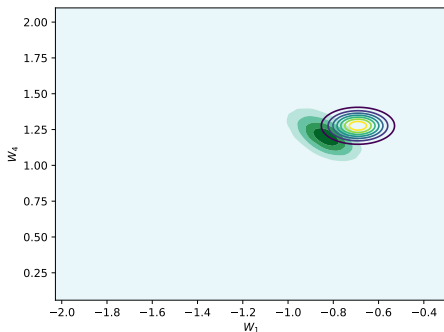
Global model



Local model



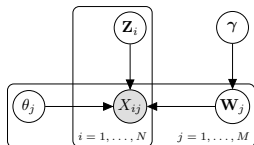
Comparison with Gibbs sampling



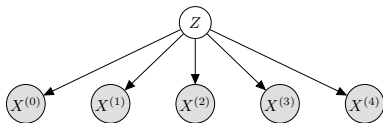
Data

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

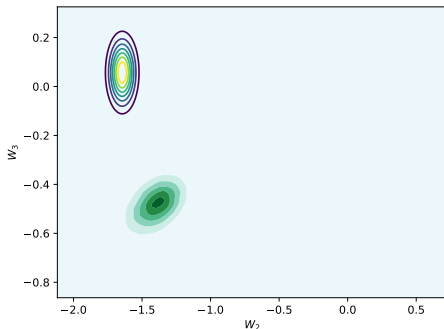
Global model



Local model



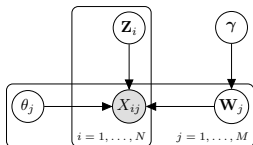
Comparison with Gibbs sampling



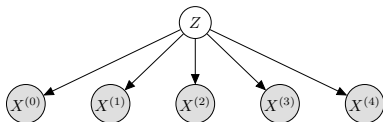
Data

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

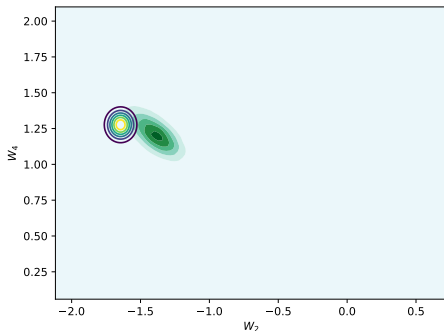
Global model



Local model



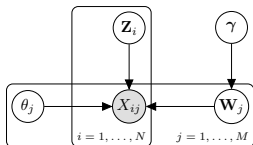
Comparison with Gibbs sampling



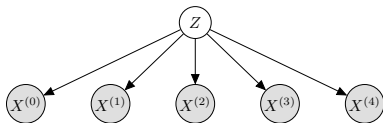
Data

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

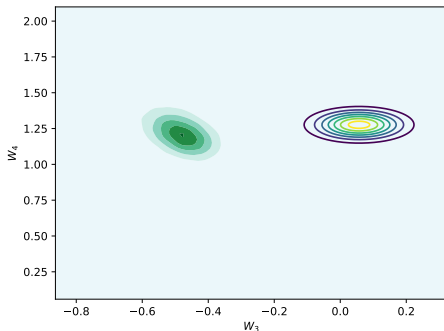
Global model



Local model



Comparison with Gibbs sampling



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

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 KL ($q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})$) as our objective.

Algorithm:

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

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

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 KL ($q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})$) as our objective.

Algorithm:

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

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

As we just realized, calculations of $\mathbb{E}_{q_{-j}} [\log p(\mathbf{z}, \mathbf{x})]$ 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.