

A Beginner's Guide to Variational Inference

Haziq Jamil

Social Statistics
London School of Economics and Political Science

1 February 2018

Social Statistics Meeting

<http://socialstats.haziqj.ml>

Outline

① Introduction

- Idea

- Comparison to EM

- Mean-field distributions

- Coordinate ascent algorithm

② Examples

- Univariate Gaussian

- Gaussian mixtures

③ Discussion

- Exponential families

- Zero-forcing vs Zero-avoiding

- Quality of approximation

- Advanced topics

Introduction

- Consider a statistical model where we have observations $\mathbf{y} = (y_1, \dots, y_n)$ and also some latent variables $\mathbf{z} = (z_1, \dots, z_m)$.

Introduction

- Consider a statistical model where we have observations $\mathbf{y} = (y_1, \dots, y_n)$ and also some latent variables $\mathbf{z} = (z_1, \dots, z_m)$.
- Want to evaluate the intractable integral

$$\mathcal{I} := \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \, d\mathbf{z}$$

- ▶ Bayesian posterior analysis
- ▶ Random effects models
- ▶ Mixture models

Introduction

- Consider a statistical model where we have observations $\mathbf{y} = (y_1, \dots, y_n)$ and also some latent variables $\mathbf{z} = (z_1, \dots, z_m)$.
- Want to evaluate the intractable integral

$$\mathcal{I} := \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \, d\mathbf{z}$$

- ▶ Bayesian posterior analysis
 - ▶ Random effects models
 - ▶ Mixture models
- Variational inference approximates the “posterior” \mathcal{I} by a tractably close distribution in the Kullback-Leibler sense.

Introduction

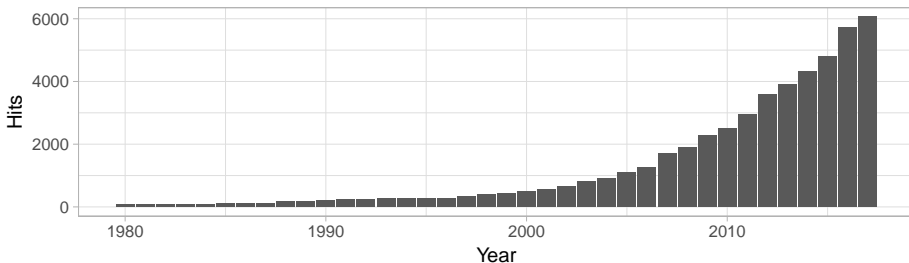
- Consider a statistical model where we have observations $\mathbf{y} = (y_1, \dots, y_n)$ and also some latent variables $\mathbf{z} = (z_1, \dots, z_m)$.
- Want to evaluate the intractable integral

$$\mathcal{I} := \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \, d\mathbf{z}$$

- ▶ Bayesian posterior analysis
- ▶ Random effects models
- ▶ Mixture models
- Variational inference approximates the “posterior” \mathcal{I} by a tractably close distribution in the Kullback-Leibler sense.
- Advantages:
 - ▶ Computationally fast
 - ▶ Convergence easily assessed
 - ▶ Works well in practice

In the literature

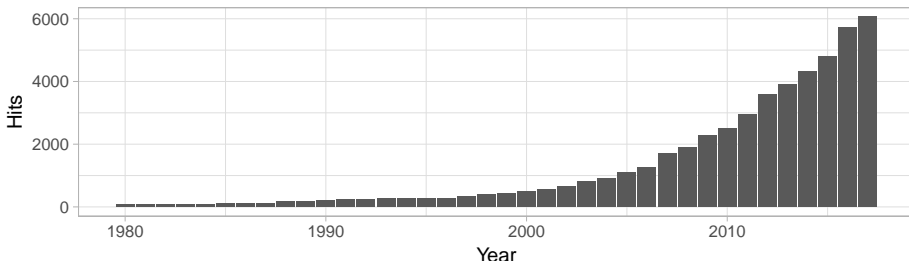
Google Scholar results for 'variational inference'



- Well known in the machine learning community.

In the literature

Google Scholar results for 'variational inference'



- Well known in the machine learning community.
- In social statistics:
 - ▶ E. A. Erosheva et al. (2007). "Describing disability through individual-level mixture models for multivariate binary data". *Ann. Appl. Stat.*, 1.2, p. 346
 - ▶ J. Grimmer (2010). "An introduction to Bayesian inference via variational approximations". *Political Analysis* 19.1, pp. 32–47
 - ▶ Y. S. Wang et al. (2017). "A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences". *arXiv: 1512.08731*

Recommended texts

- M. J. Beal and Z. Ghahramani (2003). “The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures”. In: *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*. Ed. by J. M. Bernardo et al. Oxford: Oxford University Press, pp. 453–464
- C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer
- K. P. Murphy (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press
- D. M. Blei et al. (2017). “Variational inference: A review for statisticians”. *J. Am. Stat. Assoc.*, to appear

Idea

$$p(\mathbf{z}|\mathbf{y})$$

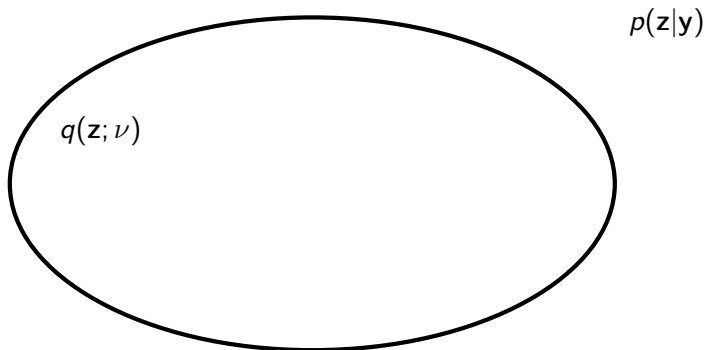
$$q(\mathbf{z})$$

- Minimise Kullback-Leibler divergence (using calculus of variations)

$$\text{KL}(q\|p) = - \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z}.$$

D. M. Blei (2017). “Variational Inference: Foundations and Innovations”. URL: <https://simons.berkeley.edu/talks/david-blei-2017-5-1>

Idea

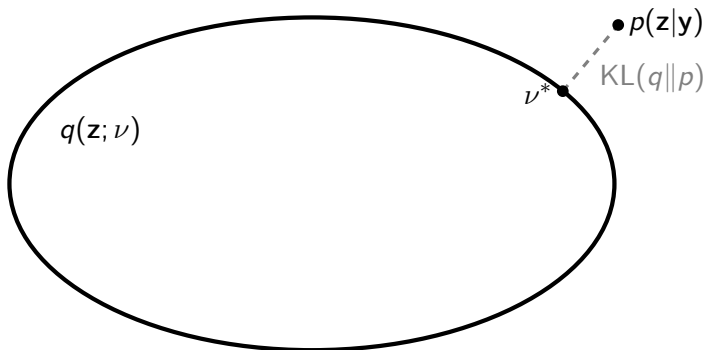


- Minimise Kullback-Leibler divergence (using calculus of variations)

$$\text{KL}(q\|p) = - \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z}.$$

D. M. Blei (2017). “Variational Inference: Foundations and Innovations”. URL: <https://simons.berkeley.edu/talks/david-blei-2017-5-1>

Idea

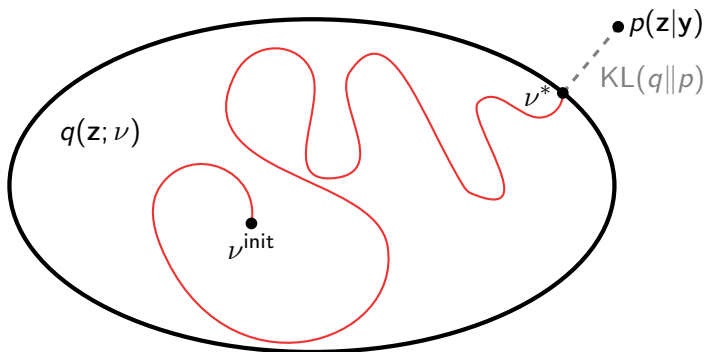


- Minimise Kullback-Leibler divergence (using calculus of variations)

$$\text{KL}(q||p) = - \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z}.$$

D. M. Blei (2017). “Variational Inference: Foundations and Innovations”. URL: <https://simons.berkeley.edu/talks/david-blei-2017-5-1>

Idea

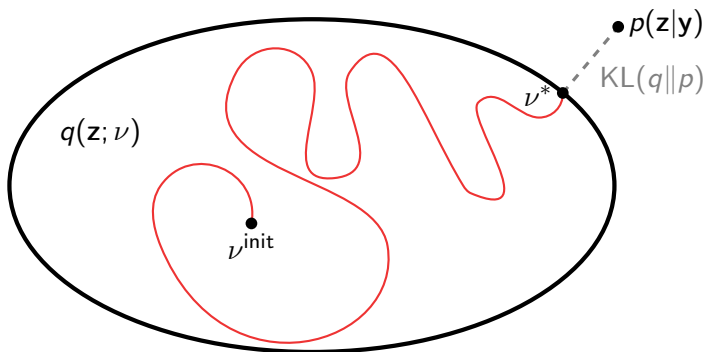


- Minimise Kullback-Leibler divergence (using calculus of variations)

$$KL(q||p) = - \int \log \frac{p(z|y)}{q(z)} q(z) dz.$$

D. M. Blei (2017). “Variational Inference: Foundations and Innovations”. URL: <https://simons.berkeley.edu/talks/david-blei-2017-5-1>

Idea



- Minimise Kullback-Leibler divergence (using calculus of variations)

$$\text{KL}(q||p) = - \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z}.$$

- **ISSUE:** $\text{KL}(q||p)$ is intractable.

D. M. Blei (2017). “Variational Inference: Foundations and Innovations”. URL: <https://simons.berkeley.edu/talks/david-blei-2017-5-1>

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$.

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\log p(\mathbf{y}) = \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y})$$

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\ &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) d\mathbf{z}\end{aligned}$$

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\ &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) d\mathbf{z} \\ &= \mathcal{L}(q) + \text{KL}(q\|p) \\ &\geq \mathcal{L}(q)\end{aligned}$$

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\ &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) d\mathbf{z} \\ &= \mathcal{L}(q) + \text{KL}(q\|p) \\ &\geq \mathcal{L}(q)\end{aligned}$$

- \mathcal{L} is referred to as the “lower-bound”, and it serves as a surrogate function to the marginal.
- Maximising $\mathcal{L}(q)$ is equivalent to minimising $\text{KL}(q\|p)$.

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\begin{aligned}
 \log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\
 &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) d\mathbf{z} \\
 &= \mathcal{L}(q) + \text{KL}(q\|p) \\
 &\geq \mathcal{L}(q)
 \end{aligned}$$

- \mathcal{L} is referred to as the “lower-bound”, and it serves as a surrogate function to the marginal.
- Maximising $\mathcal{L}(q)$ is equivalent to minimising $\text{KL}(q\|p)$.
- ISSUE:** $\mathcal{L}(q)$ is (generally) not convex.

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ .

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ .
- In the EM algorithm, the true posterior density is used, i.e. $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y}, \theta)$.

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ .
- In the EM algorithm, the true posterior density is used, i.e. $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y}, \theta)$.
- Thus,

$$\log p(\mathbf{y}|\theta) = \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y}, \theta)} - \cancel{\log \frac{p(\mathbf{z}|\mathbf{y}, \theta)}{p(\mathbf{z}|\mathbf{y}, \theta)}} \right\} p(\mathbf{z}|\mathbf{y}, \theta^{(t)}) d\mathbf{z}$$

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ .
- In the EM algorithm, the true posterior density is used, i.e. $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y}, \theta)$.
- Thus,

$$\begin{aligned} \log p(\mathbf{y}|\theta) &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y}, \theta)} - \cancel{\log \frac{p(\mathbf{z}|\mathbf{y}, \theta)}{p(\mathbf{z}|\mathbf{y}, \theta)}} \right\} p(\mathbf{z}|\mathbf{y}, \theta^{(t)}) d\mathbf{z} \\ &= E_{\theta^{(t)}}[\log p(\mathbf{y}, \mathbf{z}|\theta)] - E_{\theta^{(t)}}[\log p(\mathbf{z}|\mathbf{y}, \theta)] \end{aligned}$$

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ .
- In the EM algorithm, the true posterior density is used, i.e. $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y}, \theta)$.
- Thus,

$$\begin{aligned} \log p(\mathbf{y}|\theta) &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y}, \theta)} - \cancel{\log \frac{p(\mathbf{z}|\mathbf{y}, \theta)}{p(\mathbf{z}|\mathbf{y}, \theta)}} \right\} p(\mathbf{z}|\mathbf{y}, \theta^{(t)}) d\mathbf{z} \\ &= E_{\theta^{(t)}}[\log p(\mathbf{y}, \mathbf{z}|\theta)] - E_{\theta^{(t)}}[\log p(\mathbf{z}|\mathbf{y}, \theta)] \\ &= Q(\theta|\theta^{(t)}) + \text{entropy}. \end{aligned}$$

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ .
- In the EM algorithm, the true posterior density is used, i.e. $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y}, \theta)$.
- Thus,

$$\begin{aligned} \log p(\mathbf{y}|\theta) &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y}, \theta)} - \cancel{\log \frac{p(\mathbf{z}|\mathbf{y}, \theta)}{p(\mathbf{z}|\mathbf{y}, \theta)}} \right\} p(\mathbf{z}|\mathbf{y}, \theta^{(t)}) d\mathbf{z} \\ &= E_{\theta^{(t)}}[\log p(\mathbf{y}, \mathbf{z}|\theta)] - E_{\theta^{(t)}}[\log p(\mathbf{z}|\mathbf{y}, \theta)] \\ &= Q(\theta|\theta^{(t)}) + \text{entropy}. \end{aligned}$$

- Minimising the KL divergence corresponds to the E-step.

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ .
- In the EM algorithm, the true posterior density is used, i.e. $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y}, \theta)$.
- Thus,

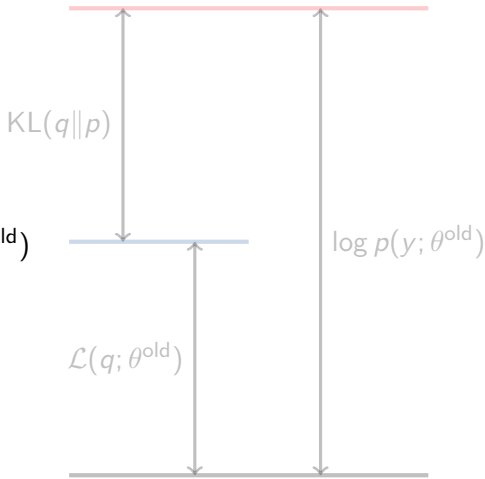
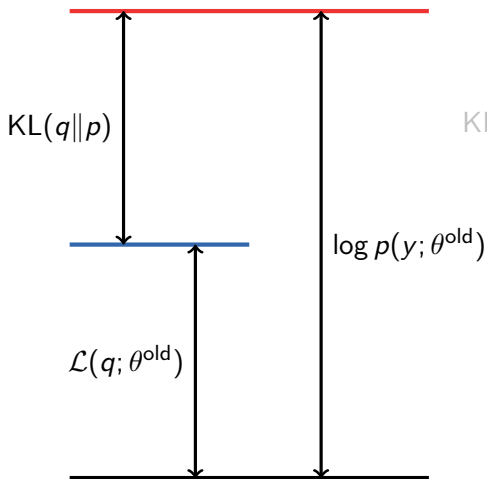
$$\begin{aligned} \log p(\mathbf{y}|\theta) &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y}, \theta)} - \log \frac{p(\mathbf{z}|\mathbf{y}, \theta)}{p(\mathbf{z}|\mathbf{y}, \theta)} \right\} p(\mathbf{z}|\mathbf{y}, \theta^{(t)}) d\mathbf{z} \\ &= E_{\theta^{(t)}}[\log p(\mathbf{y}, \mathbf{z}|\theta)] - E_{\theta^{(t)}}[\log p(\mathbf{z}|\mathbf{y}, \theta)] \\ &= Q(\theta|\theta^{(t)}) + \text{entropy}. \end{aligned}$$

- Minimising the KL divergence corresponds to the E-step.
- For any θ ,

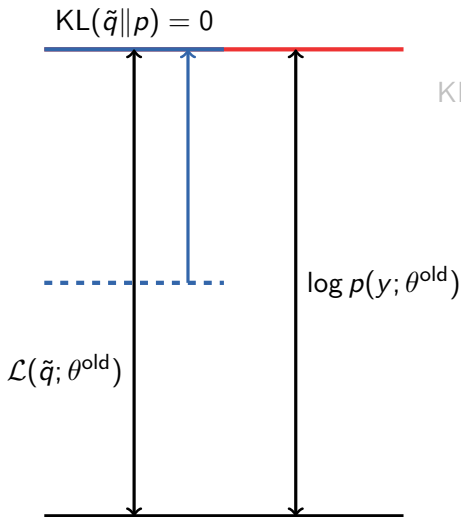
$$\begin{aligned} \log p(\mathbf{y}|\theta) - \log p(\mathbf{y}|\theta^{(t)}) &= Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + \Delta\text{entropy} \\ &\geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}). \end{aligned}$$

EM Algorithm

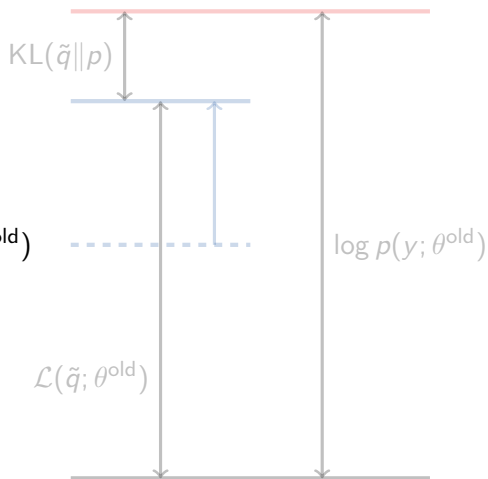
Variational Inference



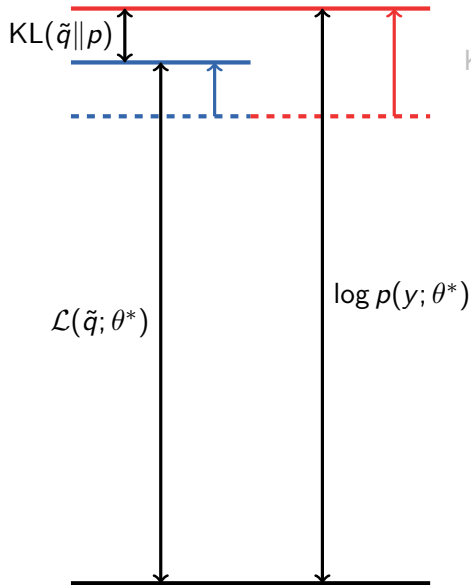
EM (E-step)



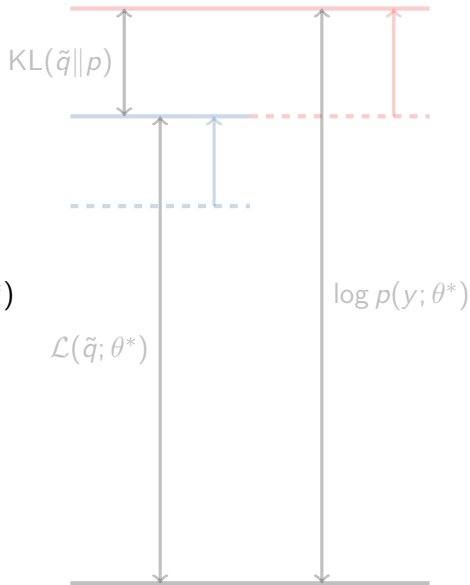
VI (E-step)



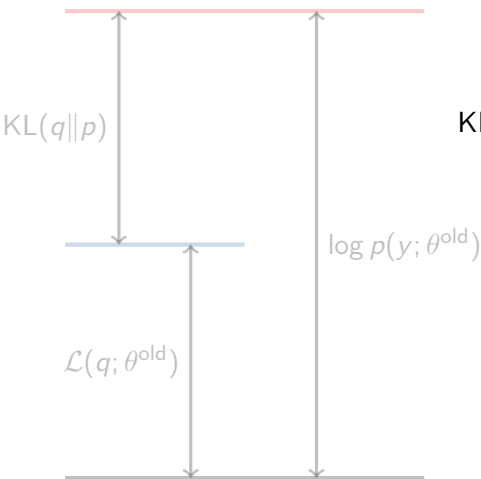
EM (M-step)



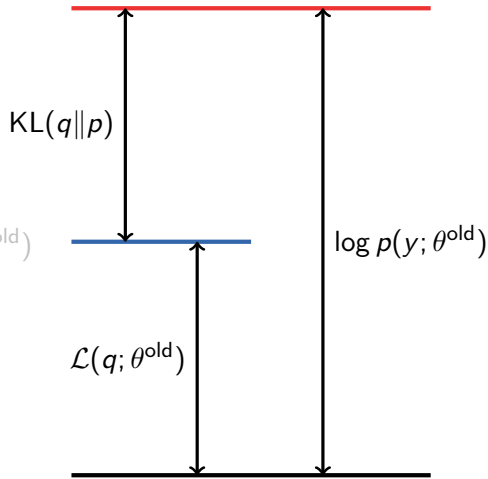
VI (M-step)



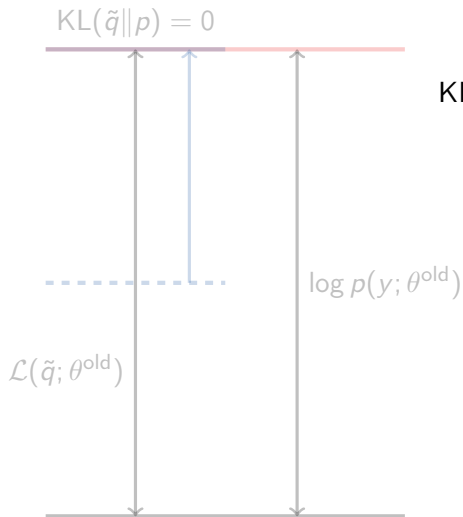
EM Algorithm



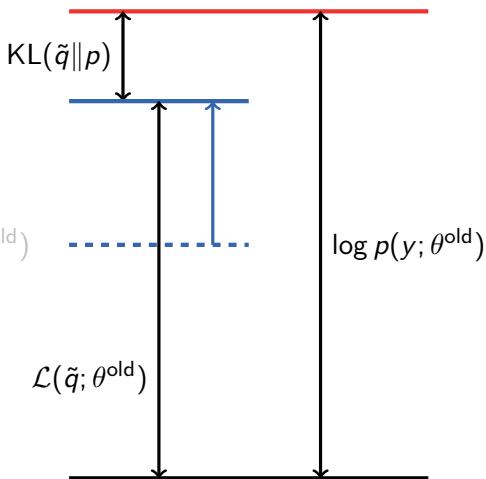
Variational Inference



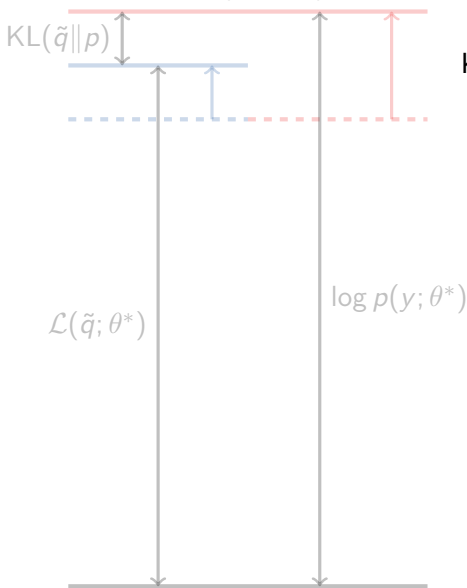
EM (E-step)



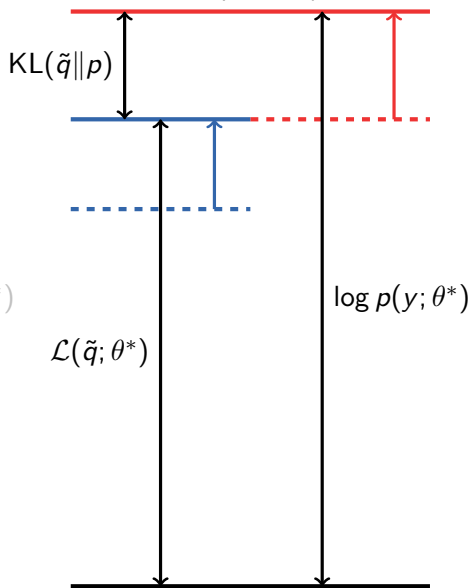
VI (E-step)



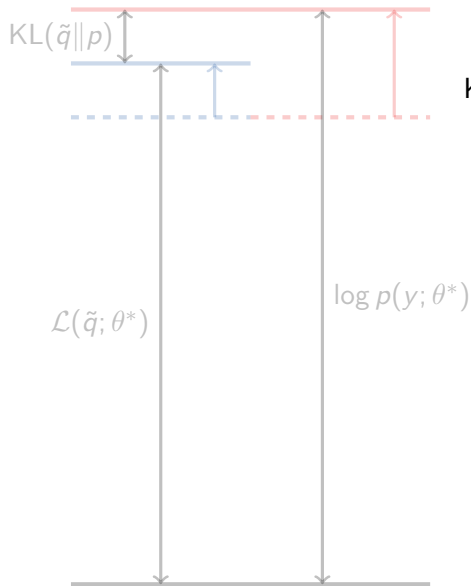
EM (M-step)



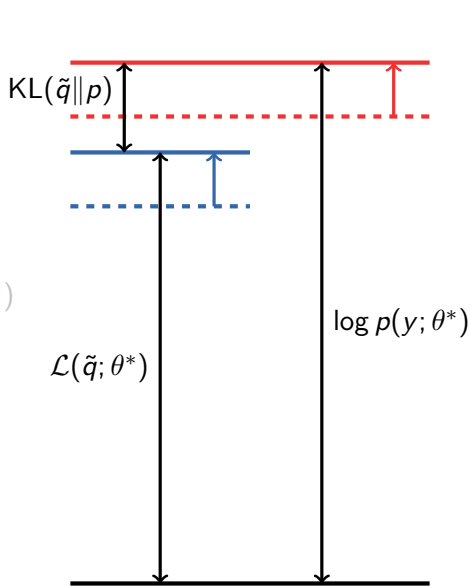
VI (M-step)



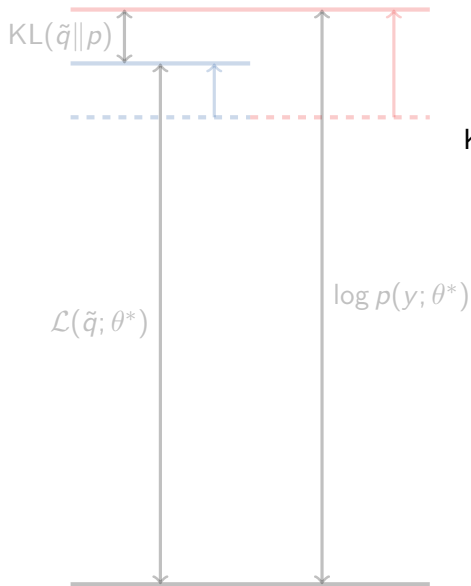
EM (M-step)



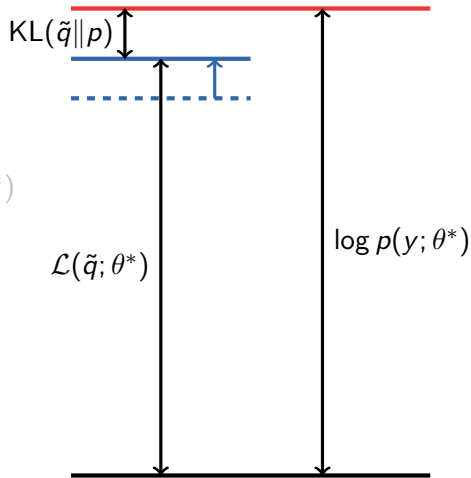
VI (M-step)



EM (M-step)



VI (M-step)



Factorised distributions (Mean-field theory)

- Maximising \mathcal{L} over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of \mathbf{z} into M disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^M q_j(\mathbf{z}^{(j)}).$$

Factorised distributions (Mean-field theory)

- Maximising \mathcal{L} over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of \mathbf{z} into M disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^M q_j(\mathbf{z}^{(j)}).$$

- Under this restriction, the solution to $\arg \max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp \left(\mathbb{E}_{-j} [\log p(\mathbf{y}, \mathbf{z})] \right) \quad (1)$$

for $j \in \{1, \dots, m\}$.

Factorised distributions (Mean-field theory)

- Maximising \mathcal{L} over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of \mathbf{z} into M disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^M q_j(\mathbf{z}^{(j)}).$$

- Under this restriction, the solution to $\arg \max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp \left(\mathbb{E}_{-j} [\log p(\mathbf{y}, \mathbf{z})] \right) \quad (1)$$

for $j \in \{1, \dots, m\}$.

- In practice, these unnormalised densities are of recognisable form (especially if conjugacy is considered).

Coordinate ascent mean-field variational inference (CAVI)

- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}^{(j)})$ depends on the optimal moments of $\mathbf{z}^{(k)}$, $k \in \{1, \dots, M : k \neq j\}$.

Coordinate ascent mean-field variational inference (CAVI)

- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}^{(j)})$ depends on the optimal moments of $\mathbf{z}^{(k)}$, $k \in \{1, \dots, M : k \neq j\}$.
- One way around this to employ an iterative procedure.

Coordinate ascent mean-field variational inference (CAVI)

- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}^{(j)})$ depends on the optimal moments of $\mathbf{z}^{(k)}$, $k \in \{1, \dots, M : k \neq j\}$.
- One way around this to employ an iterative procedure.
- Assess convergence by monitoring the lower bound

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z})].$$

Coordinate ascent mean-field variational inference (CAVI)

- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}^{(j)})$ depends on the optimal moments of $\mathbf{z}^{(k)}$, $k \in \{1, \dots, M : k \neq j\}$.
- One way around this to employ an iterative procedure.
- Assess convergence by monitoring the lower bound

$$\mathcal{L}(q) = E_q[\log p(\mathbf{y}, \mathbf{z})] - E_q[\log q(\mathbf{z})].$$

Algorithm 4 CAVI

```

1: initialise Variational factors  $q_j(\mathbf{z}^{(j)})$ 
2: while  $\mathcal{L}(q)$  not converged do
3:   for  $j = 1, \dots, M$  do
4:      $\log q_j(\mathbf{z}^{(j)}) \leftarrow E_{-j}[\log p(\mathbf{y}, \mathbf{z})] + \text{const.}$  ▷ from (1)
5:   end for
6:    $\mathcal{L}(q) \leftarrow E_q[\log p(\mathbf{y}, \mathbf{z})] - E_q[\log q(\mathbf{z})]$ 
7: end while
8: return  $\tilde{q}(\mathbf{z}) = \prod_{j=1}^M \tilde{q}_j(\mathbf{z}^{(j)})$ 

```

① Introduction

② Examples

③ Discussion

Estimation of a 1-dim Gaussian mean and variance

- **GOAL:** Bayesian inference of mean μ and variance ψ^{-1}

$$y_i \stackrel{\text{iid}}{\sim} \text{N}(\mu, \psi^{-1}) \quad \text{Data}$$

$$\mu | \psi \sim \text{N}(\mu_0, (\kappa_0 \psi)^{-1}) \quad \text{Priors}$$

$$\psi \sim \Gamma(a_0, b_0)$$

$$i = 1, \dots, n$$

Estimation of a 1-dim Gaussian mean and variance

- **GOAL:** Bayesian inference of mean μ and variance ψ^{-1}

$$y_i \stackrel{\text{iid}}{\sim} \text{N}(\mu, \psi^{-1}) \quad \text{Data}$$

$$\mu | \psi \sim \text{N}(\mu_0, (\kappa_0 \psi)^{-1}) \quad \text{Priors}$$

$$\psi \sim \Gamma(a_0, b_0)$$

$$i = 1, \dots, n$$

- Substitute $p(\mu, \psi | \mathbf{y})$ with the mean-field approximation

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi).$$

Estimation of a 1-dim Gaussian mean and variance

- **GOAL:** Bayesian inference of mean μ and variance ψ^{-1}

$$y_i \stackrel{\text{iid}}{\sim} \text{N}(\mu, \psi^{-1}) \quad \text{Data}$$

$$\mu | \psi \sim \text{N}(\mu_0, (\kappa_0 \psi)^{-1}) \quad \text{Priors}$$

$$\psi \sim \Gamma(a_0, b_0)$$

$$i = 1, \dots, n$$

- Substitute $p(\mu, \psi | \mathbf{y})$ with the mean-field approximation

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi).$$

- From (1), we can work out the solutions

Estimation of a 1-dim Gaussian mean and variance

- **GOAL:** Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg \max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp \left(\mathbb{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})] \right) \quad (1)$$

- for $j \in \{1, \dots, m\}$.

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi).$$

- From (1), we can work out the solutions

Estimation of a 1-dim Gaussian mean and variance

- **GOAL:** Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg \max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp \left(\mathbb{E}_{-j} [\log p(\mathbf{y}, \mathbf{z})] \right) \quad (1)$$

for $j \in \{1, \dots, m\}$.

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi).$$

- From (1), we can work out the solutions

$$\log \tilde{q}_\mu(\mu) = \mathbb{E}_\psi [\log p(\mathbf{y} | \mu, \psi)] + \mathbb{E}_\psi [\log p(\mu | \psi)] + \text{const.}$$

$$\begin{aligned} \log \tilde{q}_\psi(\psi) &= \mathbb{E}_\mu [\log p(\mathbf{y} | \mu, \psi)] + \mathbb{E}_\mu [\log p(\mu | \psi)] + \log p(\psi) \\ &\quad + \text{const.} \end{aligned}$$

Estimation of a 1-dim Gaussian mean and variance

- **GOAL:** Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg \max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp \left(\mathbb{E}_{-j} [\log p(\mathbf{y}, \mathbf{z})] \right) \quad (1)$$

for $j \in \{1, \dots, m\}$.

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi).$$

- From (1), we can work out the solutions

$$\tilde{q}_\mu(\mu) \equiv \mathcal{N} \left(\frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \frac{1}{(\kappa_0 + n) \mathbb{E}_q[\psi]} \right)$$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg \max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp \left(\mathbb{E}_{-j} [\log p(\mathbf{y}, \mathbf{z})] \right) \quad (1)$$

for $j \in \{1, \dots, m\}$.

$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi).$$

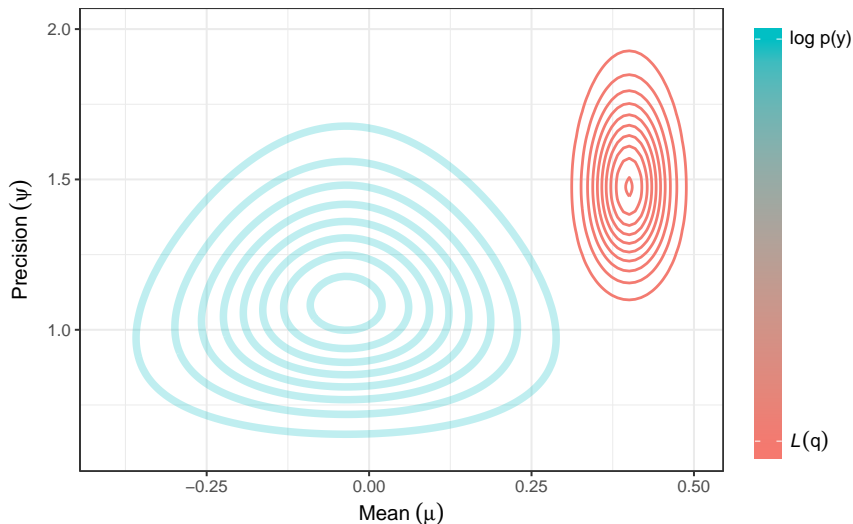
- From (1), we can work out the solutions

$$\tilde{q}_\mu(\mu) \equiv \mathcal{N} \left(\frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \frac{1}{(\kappa_0 + n) \mathbb{E}_q[\psi]} \right) \quad \text{and} \quad \tilde{q}_\psi(\psi) \equiv \Gamma(\tilde{a}, \tilde{b})$$

$$\tilde{a} = a_0 + \frac{n}{2} \quad \tilde{b} = b_0 + \frac{1}{2} \mathbb{E}_q \left[\sum_{i=1}^n (y_i - \mu)^2 + \kappa_0 (\mu - \mu_0)^2 \right]$$

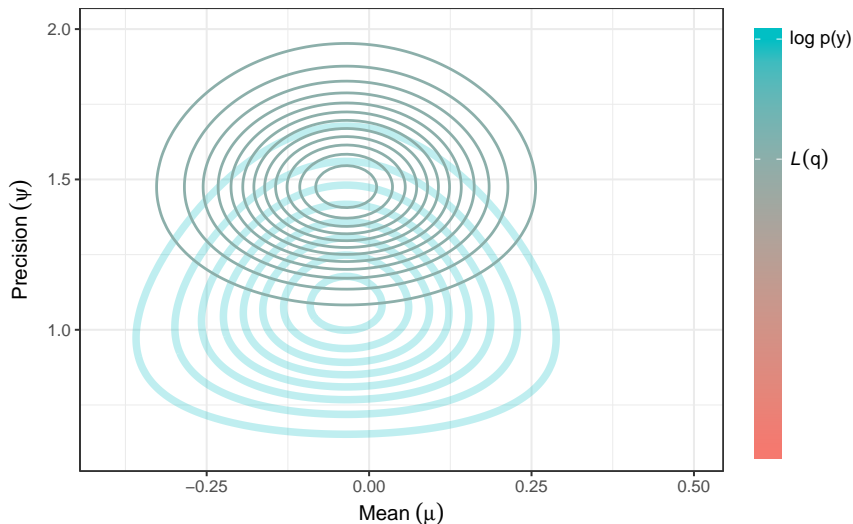
Estimation of a 1-dim Gaussian mean and variance (cont.)

Iteration 0 (initialisation)



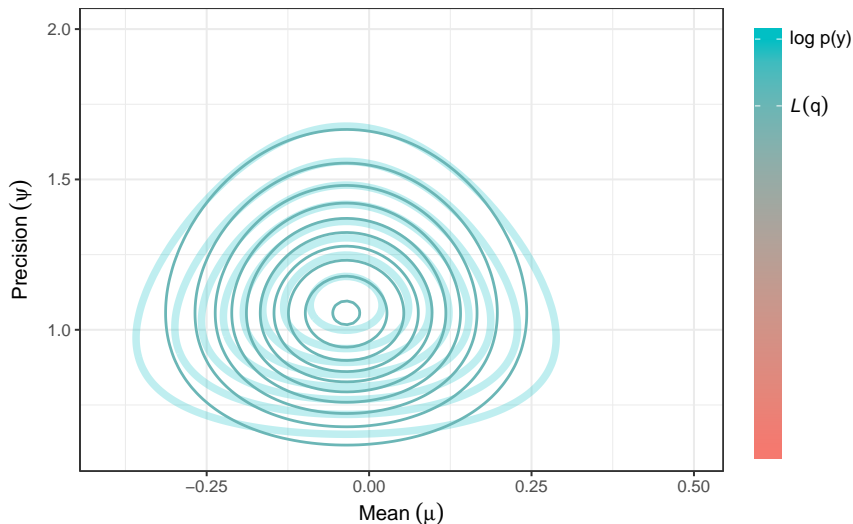
Estimation of a 1-dim Gaussian mean and variance (cont.)

Iteration 1 (μ update)



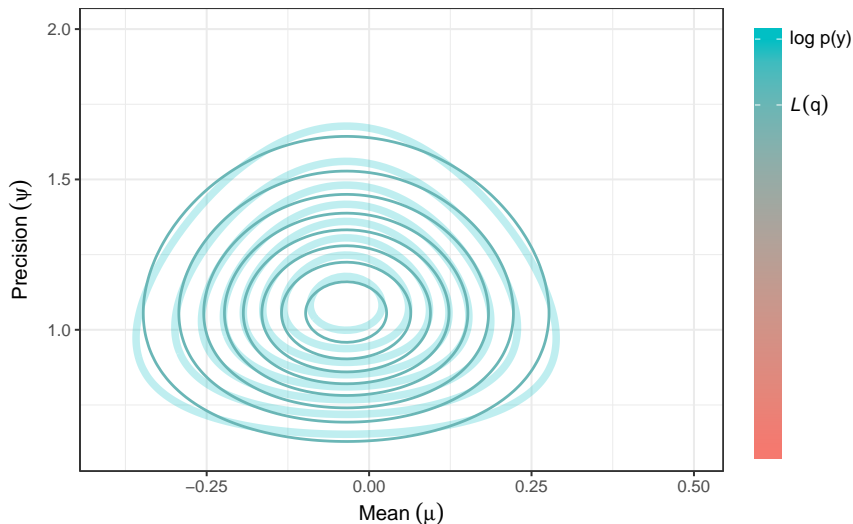
Estimation of a 1-dim Gaussian mean and variance (cont.)

Iteration 1 (ψ update)



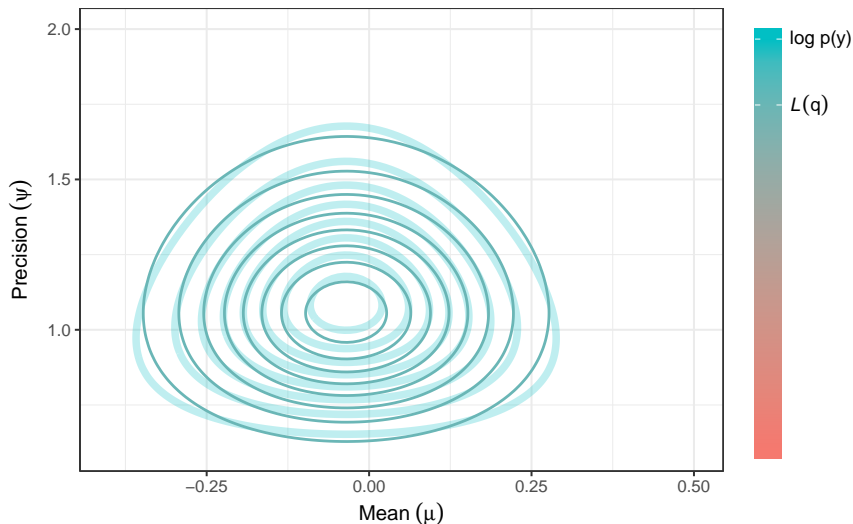
Estimation of a 1-dim Gaussian mean and variance (cont.)

Iteration 2 (μ update)



Estimation of a 1-dim Gaussian mean and variance (cont.)

Iteration 2 (ψ update)



Comparison of solutions

Variational posterior

$$\mu \sim \text{N} \left(\frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \frac{1}{(\kappa_0 + n) \text{E}[\psi]} \right)$$

$$\psi \sim \Gamma \left(a_0 + \frac{n}{2}, b_0 + \frac{1}{2} c \right)$$

$$c = \text{E} \left[\sum_{i=1}^n (y_i - \mu)^2 + \kappa_0 (\mu - \mu_0)^2 \right]$$

True posterior

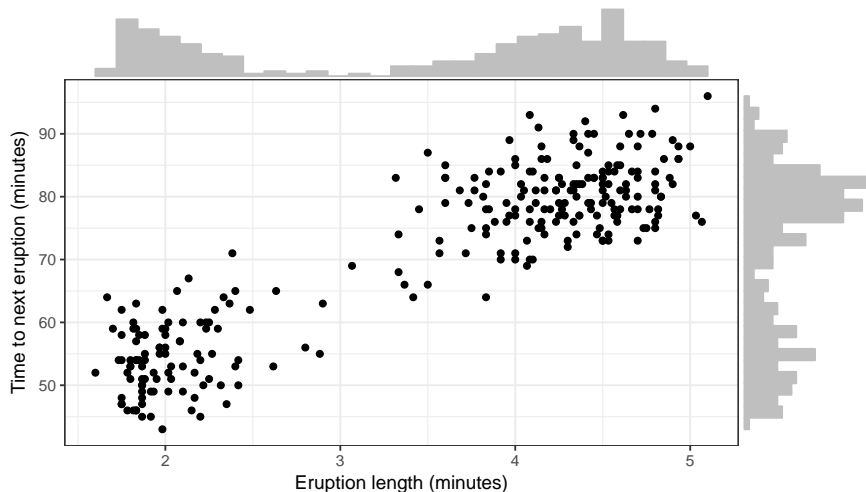
$$\mu | \psi \sim \text{N} \left(\frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \frac{1}{(\kappa_0 + n) \psi} \right)$$

$$\psi \sim \Gamma \left(a_0 + \frac{n}{2}, b_0 + \frac{1}{2} c' \right)$$

$$c' = \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{\kappa_0}{\kappa_0 + n} (\bar{y} - \mu_0)^2$$

- $\text{Cov}(\mu, \psi) = 0$ by design in VI solutions.
- For this simple example, it is possible to decouple and solve explicitly.
- VI solutions leads to unbiased MLE if $\kappa_0 = \mu_0 = a_0 = b_0 = 0$.

Gaussian mixture model (Old Faithful data set)



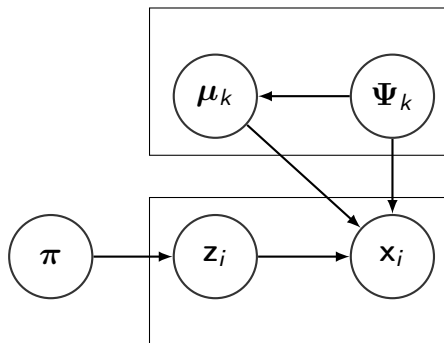
- Let $\mathbf{x}_i \in \mathbb{R}^d$ and assume $\mathbf{x}_i \stackrel{\text{iid}}{\sim} \sum_{k=1}^K \pi_k \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k^{-1})$ for $i = 1, \dots, n$.

Gaussian mixture model

- Introduce $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$, a 1-of- K binary vector, where each $z_{ik} \sim \text{Bern}(\pi_k)$.
- Assuming $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ are observed along with $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$,

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \prod_{i=1}^n \prod_{k=1}^K \text{N}_d(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Psi}_k^{-1})^{z_{ik}}.$$

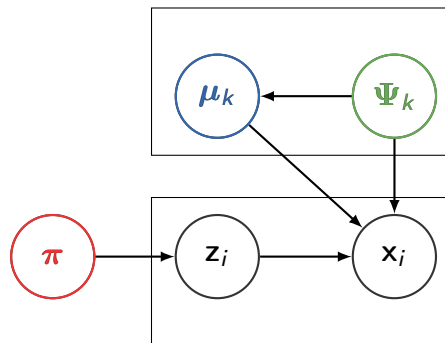
Gaussian mixture model



- Introduce $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$, a 1-of- K binary vector, where each $z_{ik} \sim \text{Bern}(\pi_k)$.
- Assuming $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ are observed along with $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$,

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \prod_{i=1}^n \prod_{k=1}^K \text{N}_d(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Psi}_k^{-1})^{z_{ik}}.$$

Gaussian mixture model



$$\begin{aligned}
 p(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi}) &= p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Psi})p(\mathbf{z}|\boldsymbol{\pi}) \\
 &\quad \times p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Psi})p(\boldsymbol{\Psi}) \\
 &= p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Psi})p(\mathbf{z}|\boldsymbol{\pi}) \\
 &\quad \times \text{Dir}_K(\boldsymbol{\pi}|\alpha_{01}, \dots, \alpha_{0K}) \\
 &\quad \times \prod_{k=1}^K N_d(\boldsymbol{\mu}_k|\mathbf{m}_0, (\kappa_0 \boldsymbol{\Psi}_k)^{-1}) \\
 &\quad \times \prod_{k=1}^K \text{Wis}_d(\boldsymbol{\Psi}_k|\mathbf{W}_0, \nu_0)
 \end{aligned}$$

- Introduce $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$, a 1-of- K binary vector, where each $z_{ik} \sim \text{Bern}(\pi_k)$.
- Assuming $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ are observed along with $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$,

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \prod_{i=1}^n \prod_{k=1}^K N_d(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k^{-1})^{z_{ik}}.$$

Variational inference for GMM

- Assume the mean-field posterior density

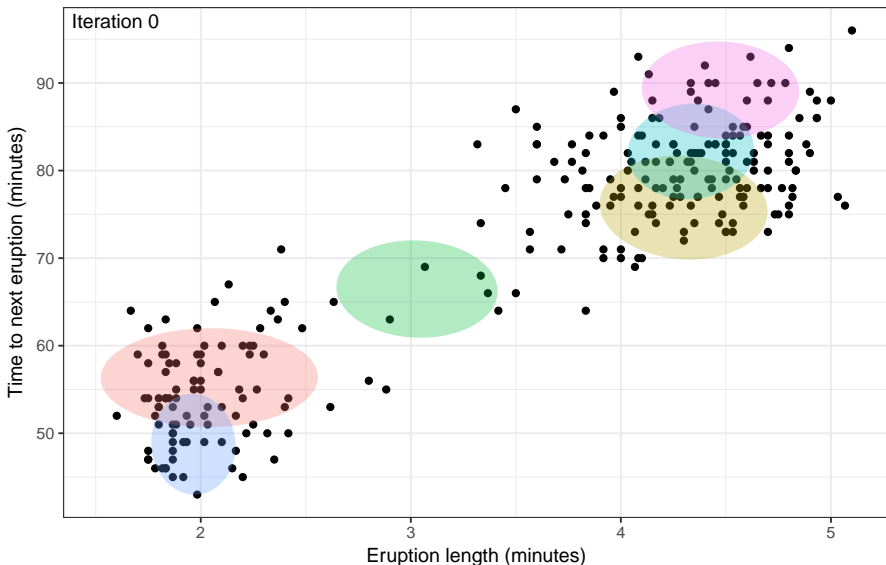
$$\begin{aligned}q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi}) &= q(\mathbf{z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi}) \\ &= q(\mathbf{z})q(\boldsymbol{\pi})q(\boldsymbol{\mu}|\boldsymbol{\Psi})q(\boldsymbol{\Psi}).\end{aligned}$$

Algorithm 5 CAVI for GMM

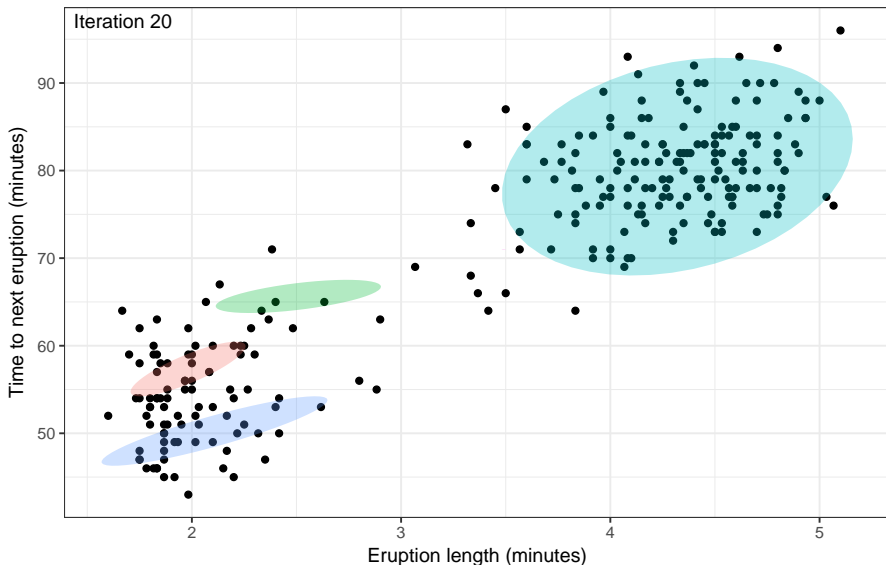
[details](#)

- 1: **initialise** Variational factors $q(\mathbf{z})$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\mu}, \boldsymbol{\Psi})$
 - 2: **while** $\mathcal{L}(q)$ not converged **do**
 - 3: $q(z_{ik}) \leftarrow \text{Bern}(\cdot)$
 - 4: $q(\boldsymbol{\pi}) \leftarrow \text{Dir}_K(\cdot)$
 - 5: $q(\boldsymbol{\mu}|\boldsymbol{\Psi}) \leftarrow \text{N}_d(\cdot, \cdot)$
 - 6: $q(\boldsymbol{\Psi}) \leftarrow \text{Wis}_d(\cdot, \cdot)$
 - 7: $\mathcal{L}(q) \leftarrow \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi})] - \mathbb{E}_q[\log q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi})]$
 - 8: **end while**
 - 9: **return** $\tilde{q}(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \tilde{q}(\mathbf{z})\tilde{q}(\boldsymbol{\pi})\tilde{q}(\boldsymbol{\mu}|\boldsymbol{\Psi})\tilde{q}(\boldsymbol{\Psi})$
-

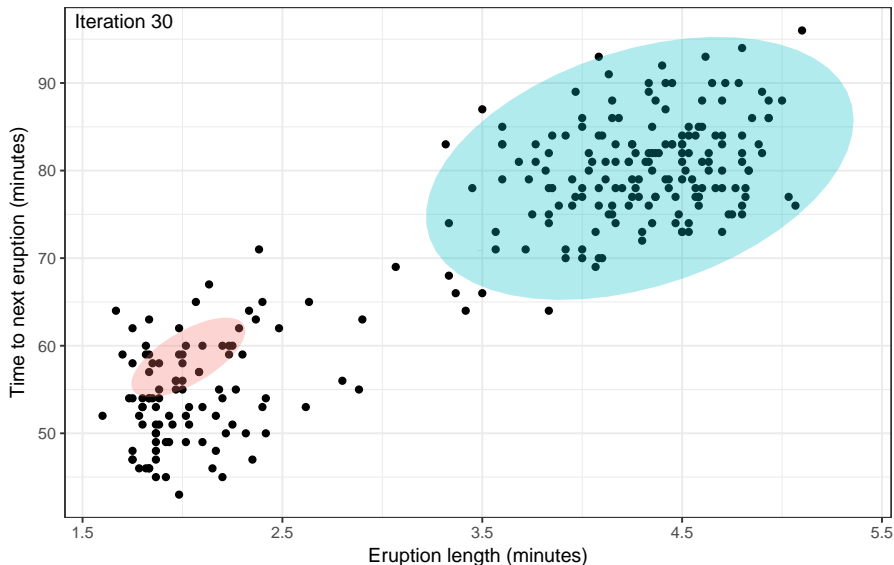
Variational inference for GMM (cont.)



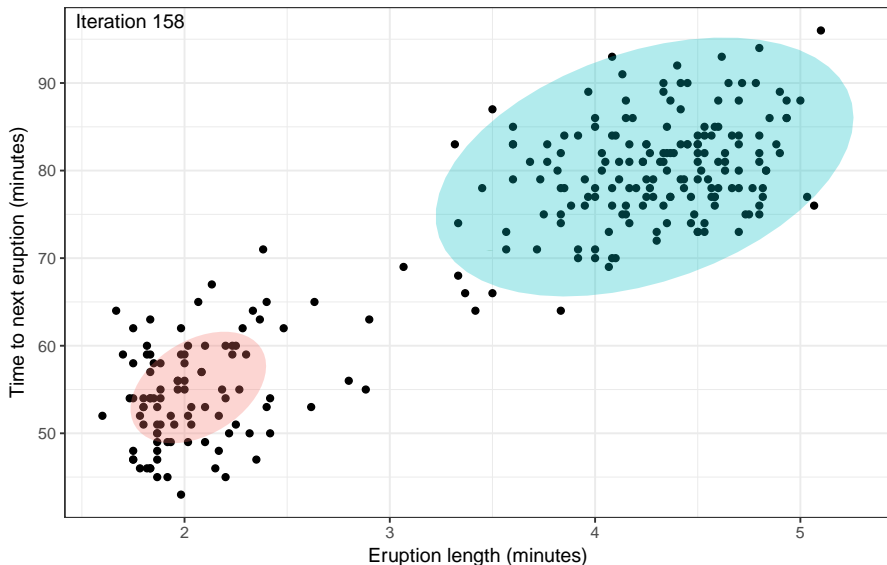
Variational inference for GMM (cont.)



Variational inference for GMM (cont.)



Variational inference for GMM (cont.)



Final thoughts on variational GMM

- Similar algorithm to the EM, and therefore similar computational time.
- Can extend to mixture of bernoullis a.k.a. latent class analysis.
- **PROS:**
 - ▶ Automatic selection of number of mixture components.
 - ▶ Less pathological special cases compared to EM solutions because regularised by prior information.
 - ▶ Less sensitive to number of parameters/components.
- **CONS:**
 - ▶ Hyperparameter tuning.

① Introduction

② Examples

③ Discussion

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)} | \mathbf{z}_{-j}, \mathbf{y}) = h(\mathbf{z}^{(j)}) \exp(\eta_j(\mathbf{z}_{-j}, \mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)).$$

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)} | \mathbf{z}_{-j}, \mathbf{y}) = h(\mathbf{z}^{(j)}) \exp(\eta_j(\mathbf{z}_{-j}, \mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)).$$

- Then, from (1),

$$\begin{aligned}\tilde{q}_j(\mathbf{z}^{(j)}) &\propto \exp(E_{-j}[\log p(\mathbf{z}^{(j)} | \mathbf{z}_{-j}, \mathbf{y})]) \\ &= \exp(\log h(\mathbf{z}^{(j)}) + E[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)} - E[A(\eta_j)]) \\ &\propto h(\mathbf{z}^{(j)}) \exp(E[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)})\end{aligned}$$

is also in the same exponential family.

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)} | \mathbf{z}_{-j}, \mathbf{y}) = h(\mathbf{z}^{(j)}) \exp(\eta_j(\mathbf{z}_{-j}, \mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)).$$

- Then, from (1),

$$\begin{aligned}\tilde{q}_j(\mathbf{z}^{(j)}) &\propto \exp(E_{-j}[\log p(\mathbf{z}^{(j)} | \mathbf{z}_{-j}, \mathbf{y})]) \\ &= \exp(\log h(\mathbf{z}^{(j)}) + E[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)} - E[A(\eta_j)]) \\ &\propto h(\mathbf{z}^{(j)}) \exp(E[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)})\end{aligned}$$

is also in the same exponential family.

- C.f. Gibbs conditional densities.

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)} | \mathbf{z}_{-j}, \mathbf{y}) = h(\mathbf{z}^{(j)}) \exp (\eta_j(\mathbf{z}_{-j}, \mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)).$$

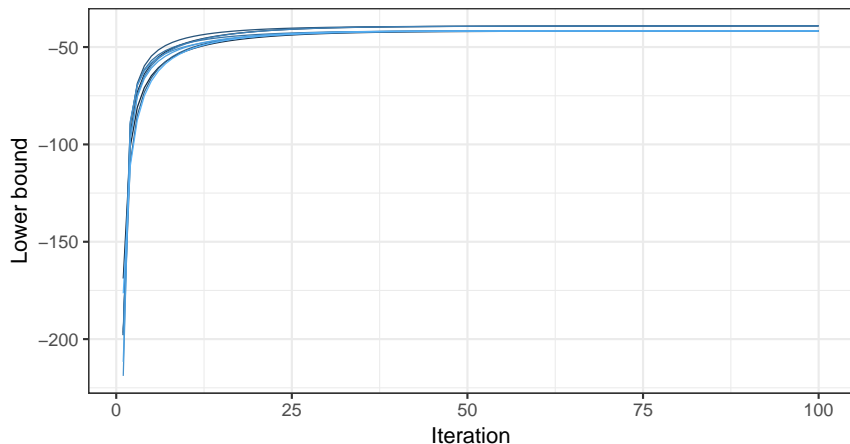
- Then, from (1),

$$\begin{aligned} \tilde{q}_j(\mathbf{z}^{(j)}) &\propto \exp (E_{-j}[\log p(\mathbf{z}^{(j)} | \mathbf{z}_{-j}, \mathbf{y})]) \\ &= \exp \left(\log h(\mathbf{z}^{(j)}) + E[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)} - E[A(\eta_j)] \right) \\ &\propto h(\mathbf{z}^{(j)}) \exp \left(E[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)} \right) \end{aligned}$$

is also in the same exponential family.

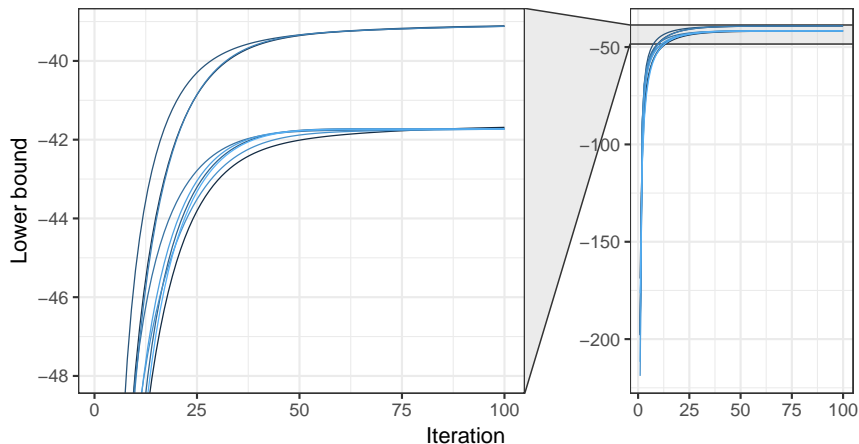
- C.f. Gibbs conditional densities.
- ISSUE:** What if not in exponential family? Importance sampling or Metropolis sampling.

Non-convexity of ELBO



- CAVI only guarantees converges to a local optimum.
- Multiple local optima may exist.

Non-convexity of ELBO



- CAVI only guarantees converges to a local optimum.
- Multiple local optima may exist.

Zero-forcing vs Zero-avoiding

- Back to the KL divergence:

$$\text{KL}(q\|p) = \int \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})} q(\mathbf{z}) d\mathbf{z}$$

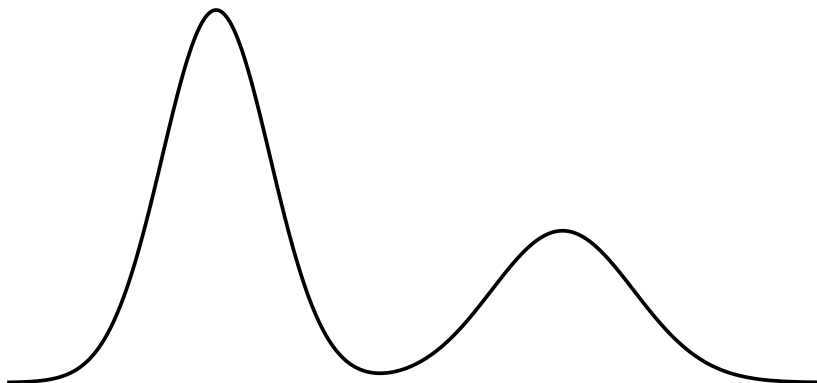
- $\text{KL}(q\|p)$ is large when $p(\mathbf{z}|\mathbf{y})$ is close to zero, unless $q(\mathbf{z})$ is also close to zero (*zero-forcing*).
- **ISSUE:** What about other measures of closeness? For instance,

$$\text{KL}(p\|q) = \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{y})} p(\mathbf{z}|\mathbf{y}) d\mathbf{z}.$$

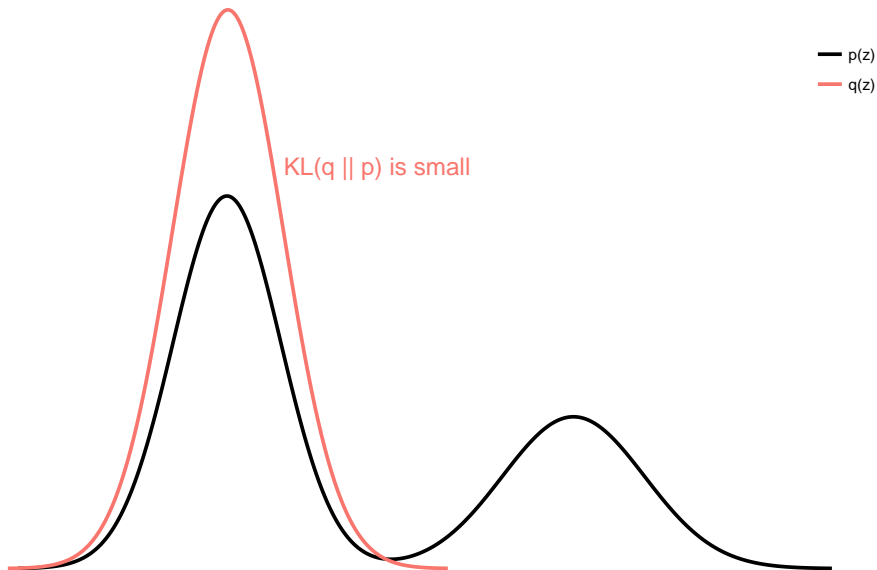
- This gives the Expectation Propagation (EP) algorithm.
- It is *zero-avoiding*, because $\text{KL}(p\|q)$ is small when both $p(\mathbf{z}|\mathbf{y})$ and $q(\mathbf{z})$ are non-zero.

Zero-forcing vs Zero-avoiding (cont.)

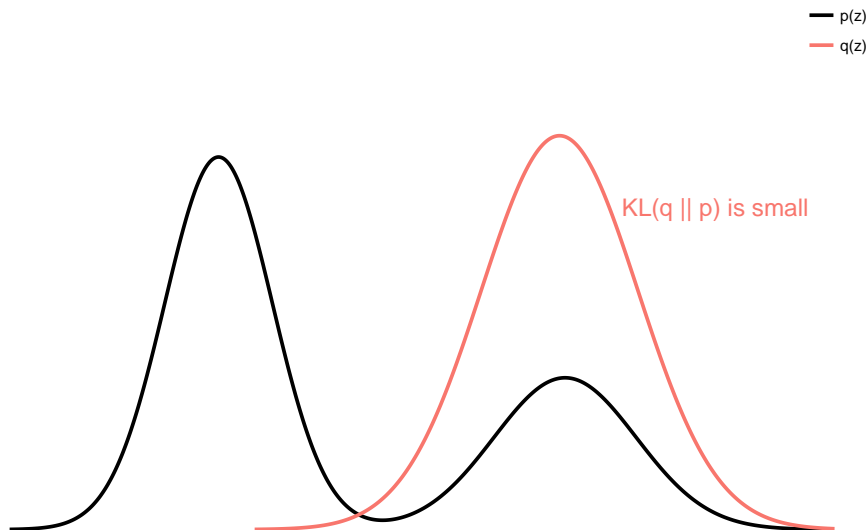
— $p(z)$



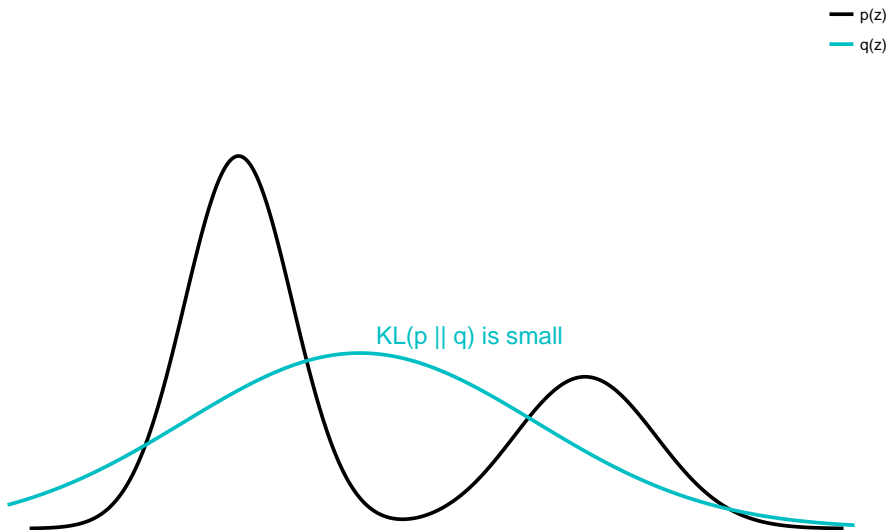
Zero-forcing vs Zero-avoiding (cont.)



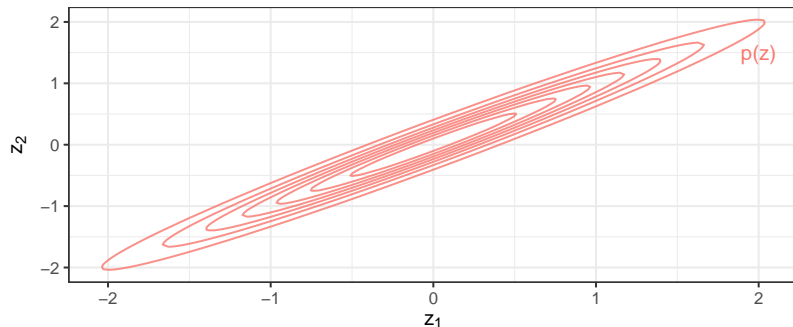
Zero-forcing vs Zero-avoiding (cont.)



Zero-forcing vs Zero-avoiding (cont.)

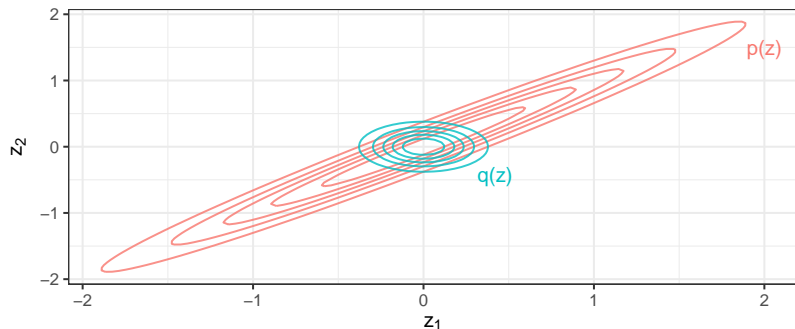


Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\text{Cov}(z_1, z_2) \neq 0$.

Distortion of higher order moments

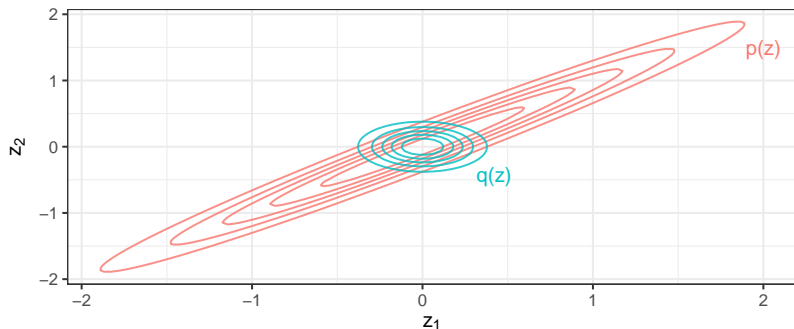


- Consider $\mathbf{z} = (z_1, z_2)^\top \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\text{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q(z_1)q(z_2)$ yields

$$\tilde{q}(z_1) = N(z_1|\mu_1, \boldsymbol{\Psi}_{11}^{-1}) \quad \text{and} \quad \tilde{q}(z_2) = N(z_2|\mu_2, \boldsymbol{\Psi}_{22}^{-1})$$

and by definition, $\text{Cov}(z_1, z_2) = 0$ under \tilde{q} .

Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\text{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q(z_1)q(z_2)$ yields

$$\tilde{q}(z_1) = N(z_1|\mu_1, \boldsymbol{\Psi}_{11}^{-1}) \quad \text{and} \quad \tilde{q}(z_2) = N(z_2|\mu_2, \boldsymbol{\Psi}_{22}^{-1})$$

and by definition, $\text{Cov}(z_1, z_2) = 0$ under \tilde{q} .

- This leads to underestimation of variances (widely reported in the literature—Zhao and Marriott 2013).

Quality of approximation

- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).

Quality of approximation

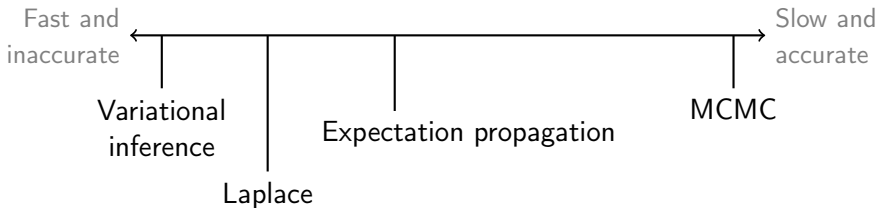
- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).
- But not much can be said about the quality of approximation.

Quality of approximation

- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).
- But not much can be said about the quality of approximation.
- Statistical properties not well understood—what is its statistical profile relative to the exact posterior?

Quality of approximation

- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).
- But not much can be said about the quality of approximation.
- Statistical properties not well understood—what is its statistical profile relative to the exact posterior?
- Speed trumps accuracy?



Advanced topics

- Local variational bounds
 - ▶ Not using the mean-field assumption.
 - ▶ Instead, find a bound for the marginalising integral \mathcal{I} .
 - ▶ Used for Bayesian logistic regression as follows:

$$\mathcal{I} = \int \text{expit}(x^\top \beta) p(\beta) d\beta \geq \int f(x^\top \beta, \xi) p(\beta) d\beta.$$

Advanced topics

- Local variational bounds

- ▶ Not using the mean-field assumption.
- ▶ Instead, find a bound for the marginalising integral \mathcal{I} .
- ▶ Used for Bayesian logistic regression as follows:

$$\mathcal{I} = \int \text{expit}(\mathbf{x}^\top \beta) p(\beta) \, \mathrm{d}\beta \geq \int f(\mathbf{x}^\top \beta, \xi) p(\beta) \, \mathrm{d}\beta.$$

- Stochastic variational inference

- ▶ Use ideas from stochastic optimisation—gradient based improvement of ELBO from subsamples of the data.
- ▶ Scales to massive data.

Advanced topics

- Local variational bounds

- ▶ Not using the mean-field assumption.
- ▶ Instead, find a bound for the marginalising integral \mathcal{I} .
- ▶ Used for Bayesian logistic regression as follows:

$$\mathcal{I} = \int \text{expit}(x^\top \beta) p(\beta) d\beta \geq \int f(x^\top \beta, \xi) p(\beta) d\beta.$$

- Stochastic variational inference

- ▶ Use ideas from stochastic optimisation—gradient based improvement of ELBO from subsamples of the data.
- ▶ Scales to massive data.

- Black box variational inference

- ▶ Beyond exponential families and model-specific derivations.

End

Thank you!

Slides and source code are made available at: <http://socialstats.haziqj.ml>

References I

- Beal, M. J. and Z. Ghahramani (2003). “The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures”. In: *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*. Ed. by J. M. Bernardo, A. P. Dawid, J. O. Berger, M. West, D. Heckerman, M. Bayarri, and A. F. Smith. Oxford: Oxford University Press, pp. 453–464.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Blei, D. M. (2017). “Variational Inference: Foundations and Innovations”. URL: <https://simons.berkeley.edu/talks/david-blei-2017-5-1>.
- Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). “Variational inference: A review for statisticians”. *Journal of the American Statistical Association*, to appear.

References II

- Erosheva, E. A., S. E. Fienberg, and C. Joutard (2007). “Describing disability through individual-level mixture models for multivariate binary data”. *Annals of Applied Statistics*, 1.2, p. 346.
- Grimmer, J. (2010). “An introduction to Bayesian inference via variational approximations”. *Political Analysis* 19.1, pp. 32–47.
- Gunawardana, A. and W. Byrne (2005). “Convergence theorems for generalized alternating minimization procedures”. *Journal of Machine Learning Research* 6, pp. 2049–2073.
- Kass, R. and A. Raftery (1995). “Bayes Factors”. *Journal of the American Statistical Association* 90.430, pp. 773–795.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

References III

- Wang, Y. S., R. Matsueda, and E. A. Erosheva (2017). “A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences”. [arXiv: 1512.08731](#).
- Zhao, H. and P. Marriott (2013). “Diagnostics for variational Bayes approximations”. [arXiv: 1309.5117](#).

④ Additional material

The variational principle

Laplace's method

Solutions to Gaussian mixture

The variational principle

- Name derived from calculus of variations which deals with maximising or minimising functionals.

Functions $p : \theta \mapsto \mathbb{R}$ (standard calculus)

Functionals $\mathcal{H} : p \mapsto \mathbb{R}$ (variational calculus)

The variational principle

- Name derived from calculus of variations which deals with maximising or minimising functionals.

Functions $p : \theta \mapsto \mathbb{R}$ (standard calculus)

Functionals $\mathcal{H} : p \mapsto \mathbb{R}$ (variational calculus)

- Using standard calculus, we can solve

$$\arg \max_{\theta} p(\theta) =: \hat{\theta}$$

e.g. p is a likelihood function, and $\hat{\theta}$ is the ML estimate.

The variational principle

- Name derived from calculus of variations which deals with maximising or minimising functionals.

Functions $p : \theta \mapsto \mathbb{R}$ (standard calculus)

Functionals $\mathcal{H} : p \mapsto \mathbb{R}$ (variational calculus)

- Using standard calculus, we can solve

$$\arg \max_{\theta} p(\theta) =: \hat{\theta}$$

e.g. p is a likelihood function, and $\hat{\theta}$ is the ML estimate.

- Using variational calculus, we can solve

$$\arg \max_p \mathcal{H}(p) =: \tilde{p}$$

e.g. \mathcal{H} is the entropy $\mathcal{H} = - \int p(x) \log p(x) dx$, and \tilde{p} is the entropy maximising distribution.

Laplace's method

- Interested in $p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) =: e^{Q(\mathbf{f})}$, with normalising constant $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} d\mathbf{f}$. The Taylor expansion of Q about its mode $\tilde{\mathbf{f}}$

$$Q(\mathbf{f}) \approx Q(\tilde{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \tilde{\mathbf{f}})^\top \mathbf{A}(\mathbf{f} - \tilde{\mathbf{f}})$$

is recognised as the logarithm of an unnormalised Gaussian density, with $\mathbf{A} = -D^2Q(\mathbf{f})$ being the negative Hessian of Q evaluated at $\tilde{\mathbf{f}}$.

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795, §4.1, pp.777–778.

Laplace's method

- Interested in $p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) =: e^{Q(\mathbf{f})}$, with normalising constant $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} d\mathbf{f}$. The Taylor expansion of Q about its mode $\tilde{\mathbf{f}}$

$$Q(\mathbf{f}) \approx Q(\tilde{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \tilde{\mathbf{f}})^\top \mathbf{A}(\mathbf{f} - \tilde{\mathbf{f}})$$

is recognised as the logarithm of an unnormalised Gaussian density, with $\mathbf{A} = -D^2Q(\mathbf{f})$ being the negative Hessian of Q evaluated at $\tilde{\mathbf{f}}$.

- The posterior $p(\mathbf{f}|\mathbf{y})$ is approximated by $N(\tilde{\mathbf{f}}, \mathbf{A}^{-1})$, and the marginal by

$$p(\mathbf{y}) \approx (2\pi)^{n/2} |\mathbf{A}|^{-1/2} p(\mathbf{y}|\tilde{\mathbf{f}}) p(\tilde{\mathbf{f}})$$

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795, §4.1, pp.777–778.

Laplace's method

- Interested in $p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) =: e^{Q(\mathbf{f})}$, with normalising constant $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} d\mathbf{f}$. The Taylor expansion of Q about its mode $\tilde{\mathbf{f}}$

$$Q(\mathbf{f}) \approx Q(\tilde{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \tilde{\mathbf{f}})^\top \mathbf{A}(\mathbf{f} - \tilde{\mathbf{f}})$$

is recognised as the logarithm of an unnormalised Gaussian density, with $\mathbf{A} = -D^2Q(\mathbf{f})$ being the negative Hessian of Q evaluated at $\tilde{\mathbf{f}}$.

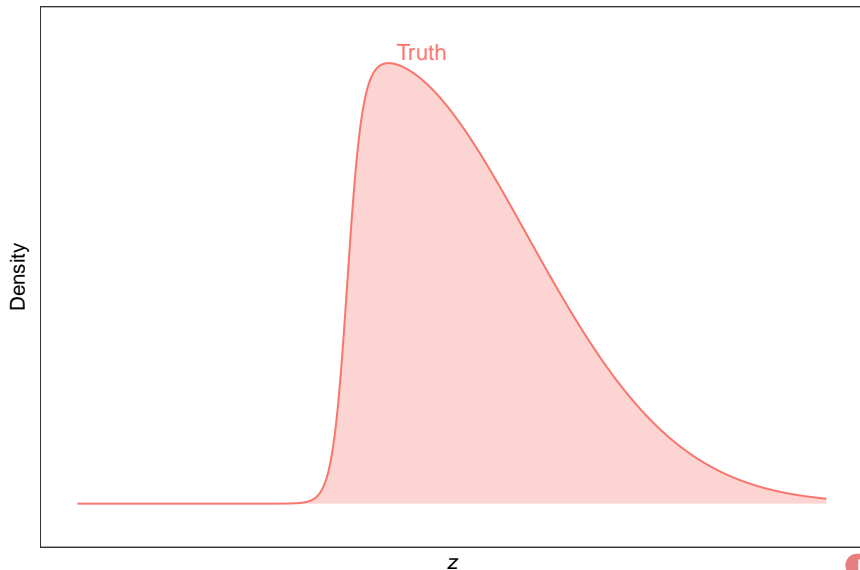
- The posterior $p(\mathbf{f}|\mathbf{y})$ is approximated by $N(\tilde{\mathbf{f}}, \mathbf{A}^{-1})$, and the marginal by

$$p(\mathbf{y}) \approx (2\pi)^{n/2} |\mathbf{A}|^{-1/2} p(\mathbf{y}|\tilde{\mathbf{f}}) p(\tilde{\mathbf{f}})$$

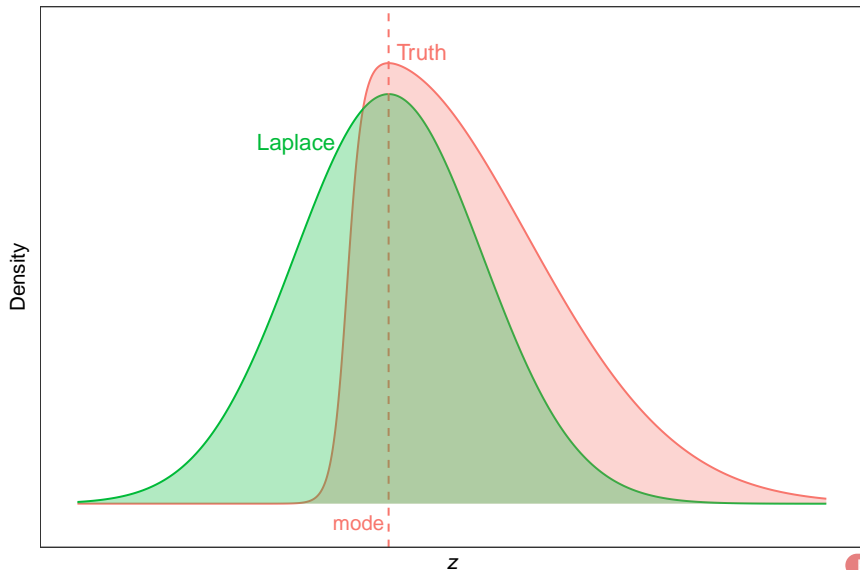
- Won't scale with large n ; difficult to find modes in high dimensions.

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795, §4.1, pp.777–778.

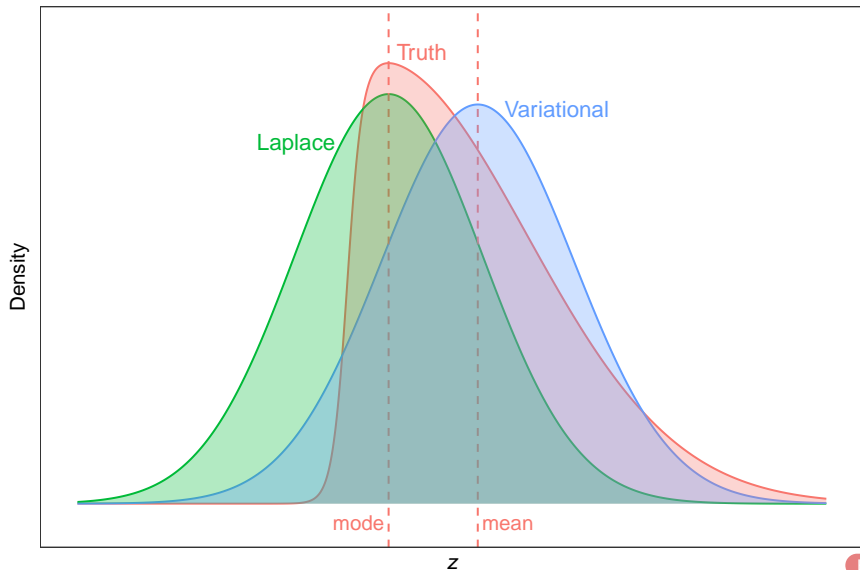
Comparison of approximations (density)



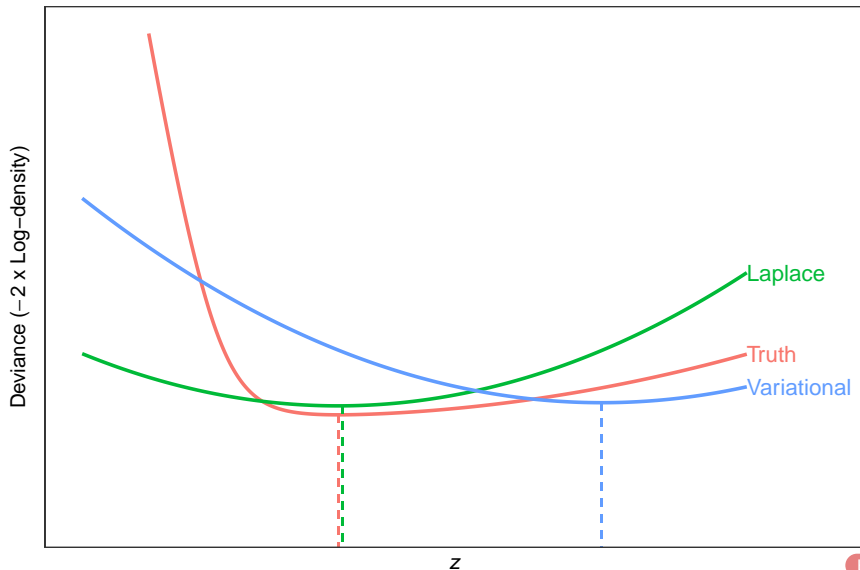
Comparison of approximations (density)



Comparison of approximations (density)



Comparison of approximations (deviance)



Variational solutions to Gaussian mixture model

Variational M-step

$$\tilde{q}(\mathbf{z}) = \prod_{i=1}^n \prod_{k=1}^K r_{ik}^{z_{ik}}, \quad r_{ik} = \rho_{ik} / \sum_{k=1}^K \rho_{ik}$$

$$\begin{aligned} \log \rho_{ik} = & \mathbb{E}[\log \pi_k] + \frac{1}{2} \mathbb{E}[\log |\Psi_k|] - \frac{d}{2} \log 2\pi \\ & - \frac{1}{2} \mathbb{E}[(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \Psi_k (\mathbf{x}_i - \boldsymbol{\mu}_k)] \end{aligned}$$

Variational E-step

$$\tilde{q}(\pi_1, \dots, \pi_K) = \text{Dir}_K(\boldsymbol{\pi} | \tilde{\boldsymbol{\alpha}}), \quad \tilde{\alpha}_k = \alpha_{0k} + \sum_{i=1}^n r_{ik}$$

$$\tilde{q}(\boldsymbol{\mu}, \Psi) = \prod_{k=1}^K \text{N}_d(\boldsymbol{\mu}_k | \tilde{\mathbf{m}}_k, (\tilde{\kappa}_k \Psi_k)^{-1}) \text{Wis}_d(\Psi_k | \tilde{\mathbf{W}}_k, \tilde{\nu}_k)$$

Variational solutions to Gaussian mixture model (cont.)

$$\begin{aligned}\tilde{\kappa}_k &= \kappa_0 + \sum_{i=1}^n r_{ik} \\ \tilde{\mathbf{m}}_k &= (\kappa_0 \mathbf{m}_0 + \sum_{i=1}^n r_{ik} \mathbf{x}_i) / \tilde{\kappa}_k \\ \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + \sum_{i=1}^n r_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top \\ \bar{\mathbf{x}}_k &= \sum_{i=1}^n r_{ik} \mathbf{x}_i / \sum_{i=1}^n r_{ik} \\ \nu_k &= \nu_0 + \sum_{i=1}^n r_{ik}\end{aligned}$$

Also useful

$$\mathbb{E} \left[(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Psi}_k (\mathbf{x}_i - \boldsymbol{\mu}_k) \right] = d / \tilde{\kappa}_k + \nu_k (\mathbf{x}_i - \tilde{\mathbf{m}}_k)^\top \tilde{\mathbf{W}}_k (\mathbf{x}_i - \tilde{\mathbf{m}}_k)$$

$$\mathbb{E}[\log \pi_k] = \sum_{i=1}^d \psi \left(\frac{\nu_k + 1 - i}{2} \right) + d \log 2 + \log |\tilde{\mathbf{W}}_k|$$

$$\mathbb{E}[\log |\boldsymbol{\Psi}_k|] = \psi(\tilde{\alpha}_k) - \psi \left(\sum_{k=1}^K \tilde{\alpha}_k \right), \quad \psi(\cdot) \text{ is the digamma function}$$