A Brief Guide to Variational Inference

Haziq Jamil

PhD (LSE), MSc (LSE), BSc MMORSE (Warw)

4 December 2018

UBD Interview Seminar

http://socialstats.haziqj.ml

Outline

Introduction

- Consider a statistical model parameterised by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ for which 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 parameterised by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ for which 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

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

  - Frequentist likelihood maximisation $\arg\max_\theta \log p(\mathbf{y}|\theta)$
  - Bayesian posterior analysis $p(\mathbf{z}|\mathbf{y}) = p(\mathbf{y}, \mathbf{z})/p(\mathbf{y})$

Introduction

- Consider a statistical model parameterised by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ for which 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

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

  - Frequentist likelihood maximisation $\arg\max_\theta \log p(\mathbf{y}|\theta)$
  - Bayesian posterior analysis $p(\mathbf{z}|\mathbf{y}) = p(\mathbf{y},\mathbf{z})/p(\mathbf{y})$
- Variational inference approximates the "posterior" $p(\mathbf{z}|\mathbf{y})$ by a tractably close distribution in the Kullback-Leibler sense.

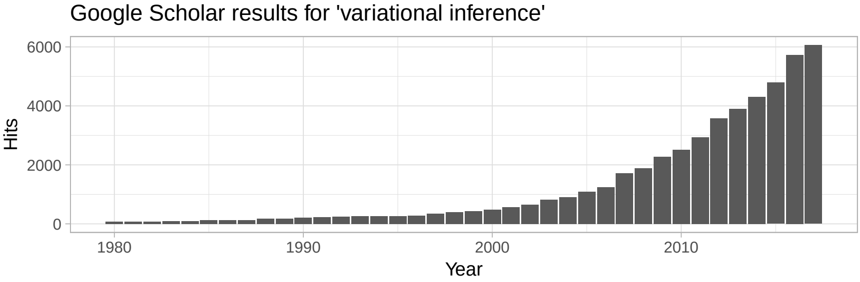
Introduction

- Consider a statistical model parameterised by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ for which 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

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

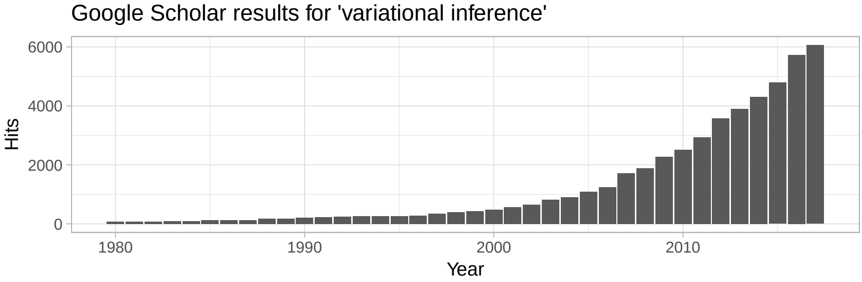
  - Frequentist likelihood maximisation $\arg\max_\theta \log p(\mathbf{y}|\theta)$
  - Bayesian posterior analysis $p(\mathbf{z}|\mathbf{y}) = p(\mathbf{y},\mathbf{z})/p(\mathbf{y})$
- Variational inference approximates the "posterior" $p(\mathbf{z}|\mathbf{y})$ by a tractably close distribution in the Kullback-Leibler sense.
- Advantages:
  - Computationally fast
  - Convergence easily assessed
  - Works well in practice

In the literature



- Well known in machine learning, slowly encroaching other fields.

In the literature



- Well known in machine learning, slowly encroaching other fields.
- Applications (Blei et al., 2017):
  - Computer vision and robotics (image denoising, tracking, recognition)
  - Natural language processing and speech recognition (topic modelling)
  - Social statistics (probit models, latent class models, variable selection)
  - Computational biology (phylogenetic hidden Markov models, population genetics, gene expression analysis)
  - Computational neuroscience (autoregressive processes, hierarchical models, spatial models, artificial neural networks)

Introductory texts

- D. M. Blei et al. (2017). "Variational Inference: A Review for Statisticians". *J. Am. Stat. Assoc*, 112.518, pp. 859–877
- C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer
- K. P. Murphy (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press
- M. J. Beal (2003). "Variational algorithms for approximate Bayesian inference". PhD thesis. Gatsby Computational Neuroscience Unit, University College London
- HJ (Oct. 2018). "Regression modelling using priors depending on Fisher information covariance kernels (I-priors)". PhD thesis. London School of Economics and Political Science

Idea

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

$q(\mathbf{z})$

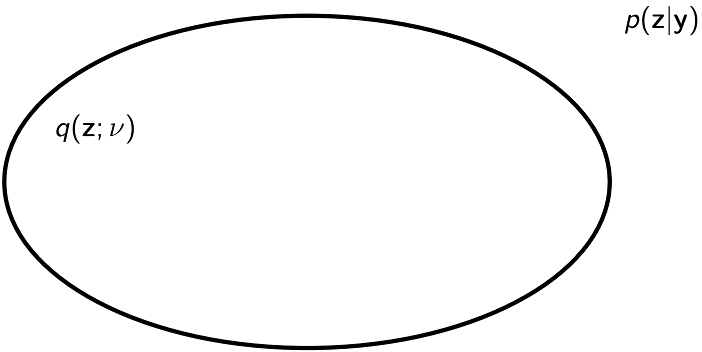
- Minimise Kullback-Leibler divergence (using calculus of variations)

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

- Use $\tilde{q}(\mathbf{z};\nu^*) := \arg\min_q \mathrm{KL}(q\|p)$ as an approximation to $p(\mathbf{z}|\mathbf{y})$.

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

Idea



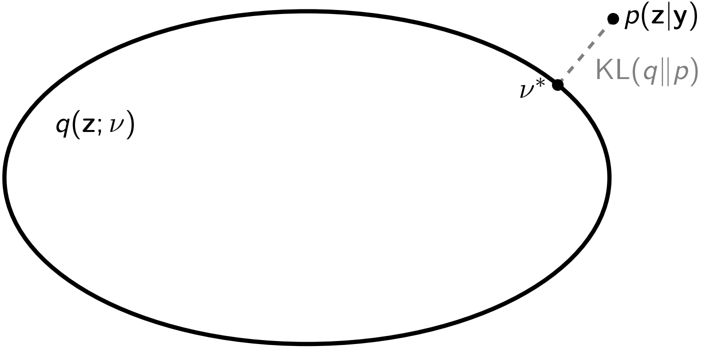
- Minimise Kullback-Leibler divergence (using calculus of variations)

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

- Use $\tilde{q}(\mathbf{z};\nu^*) := \arg\min_q \mathrm{KL}(q\|p)$ as an approximation to $p(\mathbf{z}|\mathbf{y})$.

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

Idea



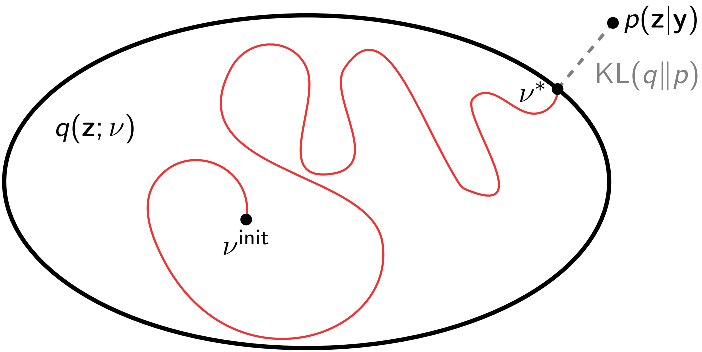
- Minimise Kullback-Leibler divergence (using calculus of variations)

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

- Use $\tilde{q}(\mathbf{z};\nu^*) := \arg\min_q \mathrm{KL}(q\|p)$ as an approximation to $p(\mathbf{z}|\mathbf{y})$.

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)

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

- Use $\tilde{q}(\mathbf{z};\nu^*) := \arg\min_q \mathrm{KL}(q\|p)$ as an approximation to $p(\mathbf{z}|\mathbf{y})$.

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:

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

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:

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

- $\mathcal{L}$ 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 $\mathrm{KL}(q\|p)$.

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:

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

- $\mathcal{L}$ 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 $\mathrm{KL}(q \| p)$.
- N.b. Equality in the bound when $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y})$, and $\mathrm{KL}(q \| p)$ vanishes (c.f. EM algorithm).

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

$$\tilde{q}_j(\mathbf{z}_{[j]}) \propto \exp\big(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\big) \tag{1}$$

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

$$\tilde{q}_j(\mathbf{z}_{[j]}) \propto \exp\big(\operatorname{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\big) \tag{1}$$

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

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

Coordinate ascent mean-field variational inference (CAVI)

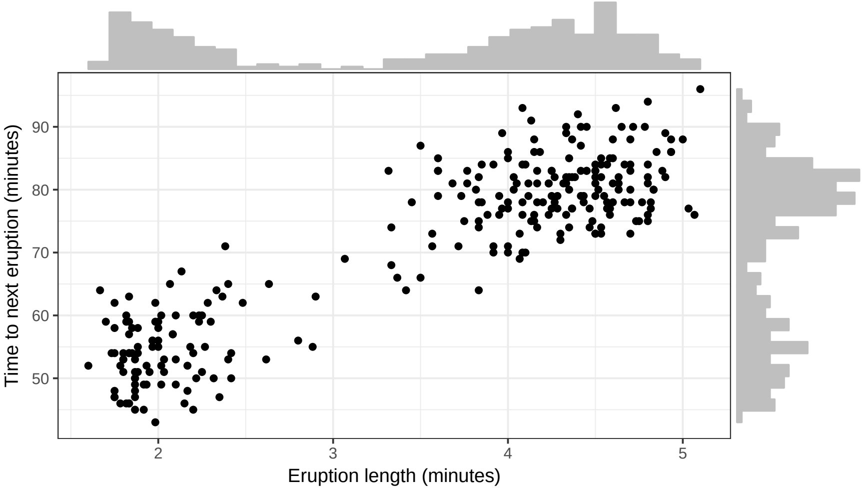
- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}_{[j]})$ depends on the optimal moments of $\mathbf{z}_{[k]}$, $k \in \{1, \dots, M | k \neq j\}$.
- One way around this to employ an iterative procedure.
- Assess convergence by monitoring the lower bound

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

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: $\log q_j(\mathbf{z}_{[j]}) \leftarrow \mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})] + \text{const.}$ ▷ from (1)
5: **end for**
6: $\mathcal{L}(q) \leftarrow \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathrm{E}_q[\log q(\mathbf{z})]$
7: **end while**
8: **return** $\tilde{q}(\mathbf{z}) = \prod_{j=1}^{M} \tilde{q}_j(\mathbf{z}_{[j]})$

Gaussian mixture model (Old Faithful data set)



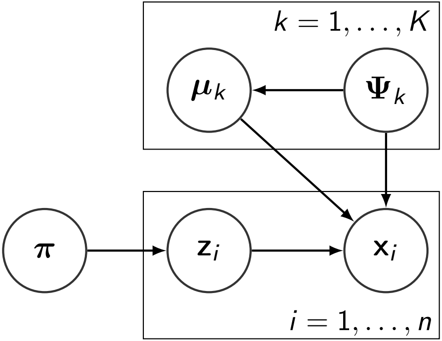
- Let $\mathbf{x}_i \in \mathbb{R}^d$ and assume $\mathbf{x}_i \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k \, \mathrm{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k^{-1})$ for $i = 1, \dots, n$.

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 $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, the conditional likelihood is

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

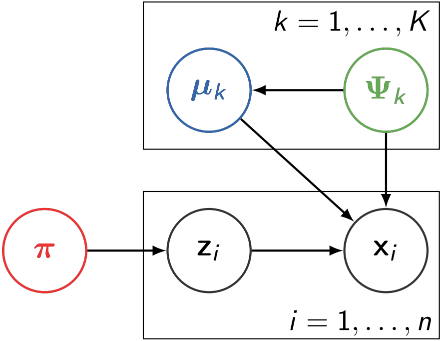
Gaussian mixture model



- Introduce $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$, a 1-of-$K$ binary vector, where each $z_{ik} \sim \mathrm{Bern}(\pi_k)$.
- Assuming $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, the conditional likelihood is

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

Gaussian mixture model



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

- Introduce $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$, a 1-of-$K$ binary vector, where each $z_{ik} \sim \operatorname{Bern}(\pi_k)$.
- Assuming $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, the conditional likelihood is

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

Variational inference for GMM

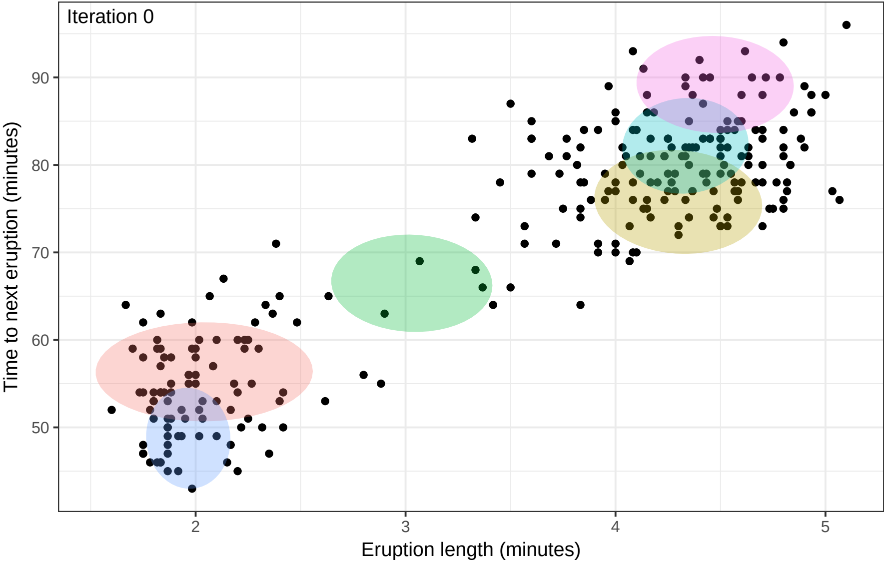
- Assume the mean-field posterior density

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

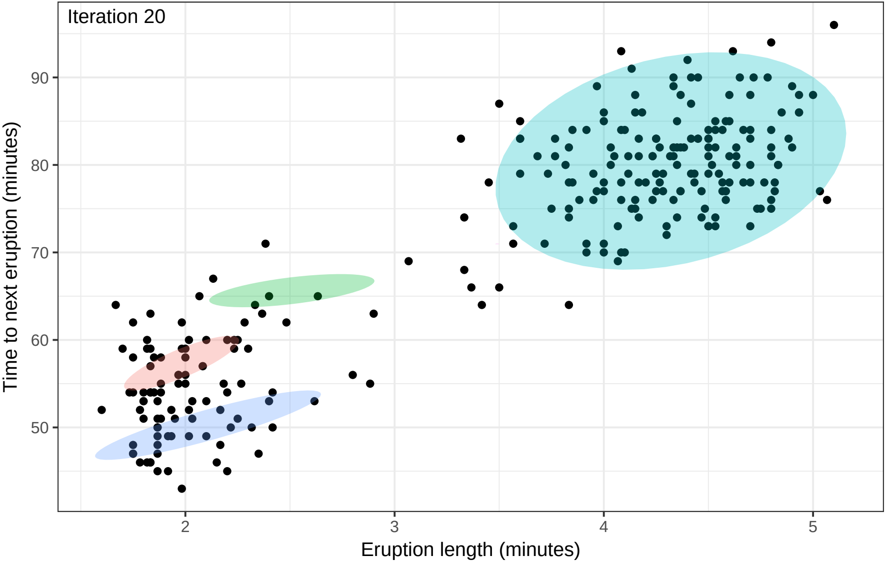
Algorithm 5 CAVI for GMM details

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: $\quad q(z_{ik}) \leftarrow \text{Bern}(\cdot)$
4: $\quad q(\boldsymbol{\pi}) \leftarrow \text{Dir}_K(\cdot)$
5: $\quad q(\boldsymbol{\mu}|\boldsymbol{\Psi}) \leftarrow \text{N}_d(\cdot, \cdot)$
6: $\quad q(\boldsymbol{\Psi}) \leftarrow \text{Wis}_d(\cdot, \cdot)$
7: $\quad \mathcal{L}(q) \leftarrow \text{E}_q[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi})] - \text{E}_q[\log q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi})]$
8: **end while**
9: **return** $\tilde{q}(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \tilde{q}(\mathbf{z}) \tilde{q}(\boldsymbol{\pi}) \tilde{q}(\boldsymbol{\mu}|\boldsymbol{\Psi}) \tilde{q}(\boldsymbol{\Psi})$

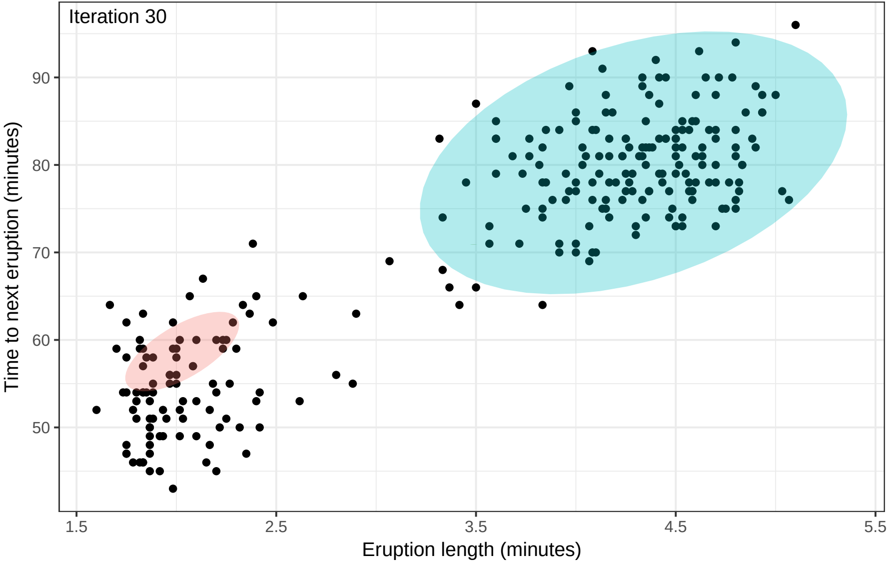
Variational inference for GMM (cont.)



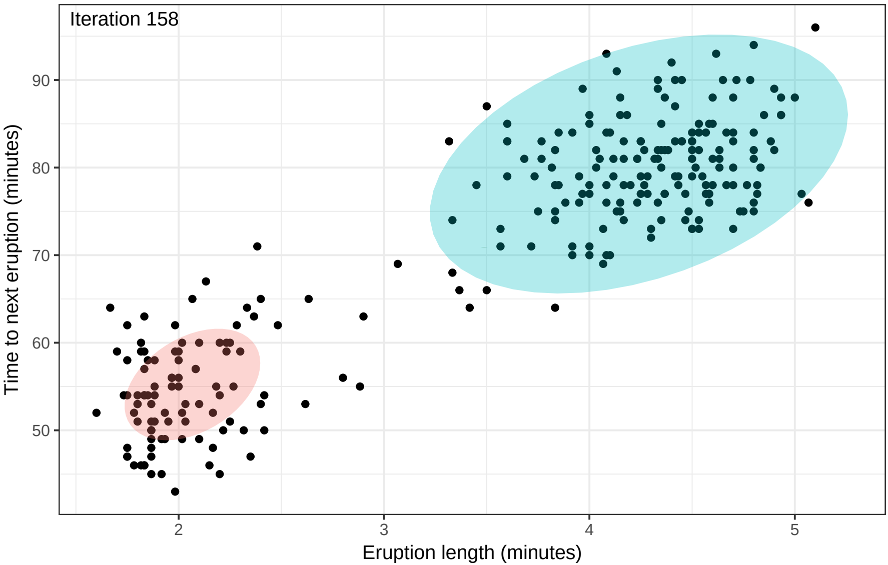
Variational inference for GMM (cont.)



Variational inference for GMM (cont.)



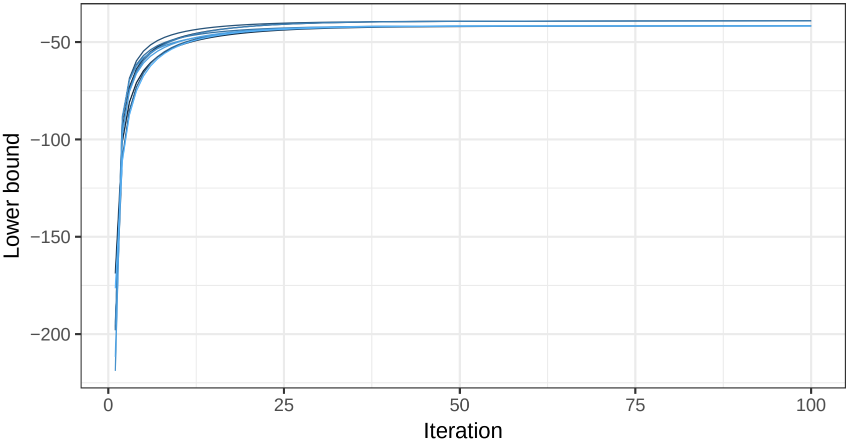
Variational inference for GMM (cont.)



Final thoughts on variational GMM

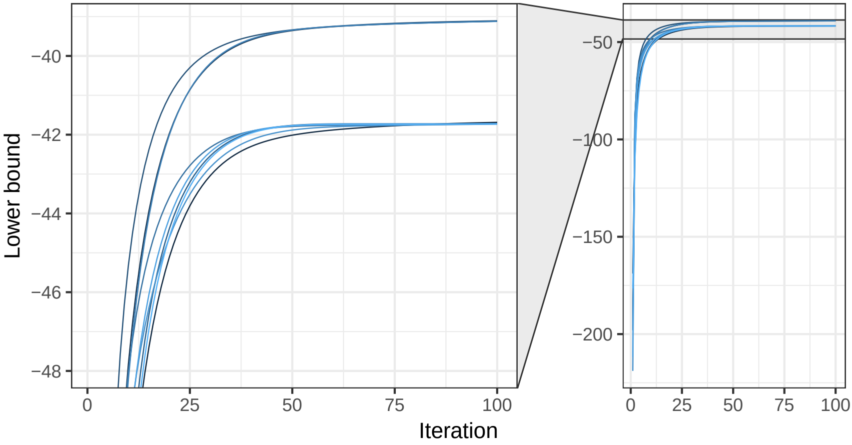
- Similar algorithm to the EM, and therefore similar computational time.
- Can extend to mixture of bernoullis a.k.a. latent class analysis.
- **PROS**:
  - Automatic selection of number of mixture components.
  - Less pathological special cases compared to EM solutions because regularised by prior information.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

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:

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

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

Zero-forcing vs Zero-avoiding

- Back to the KL divergence:

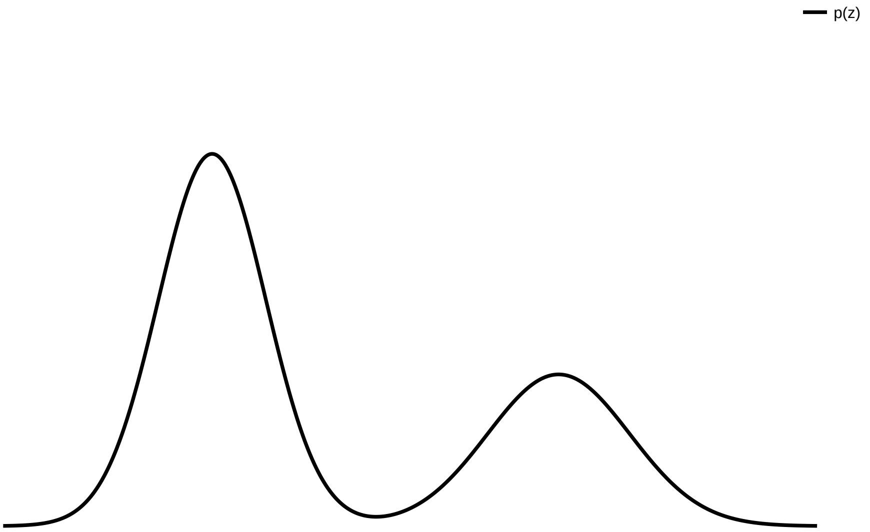
$$\mathrm{KL}(q\|p) = \int \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

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

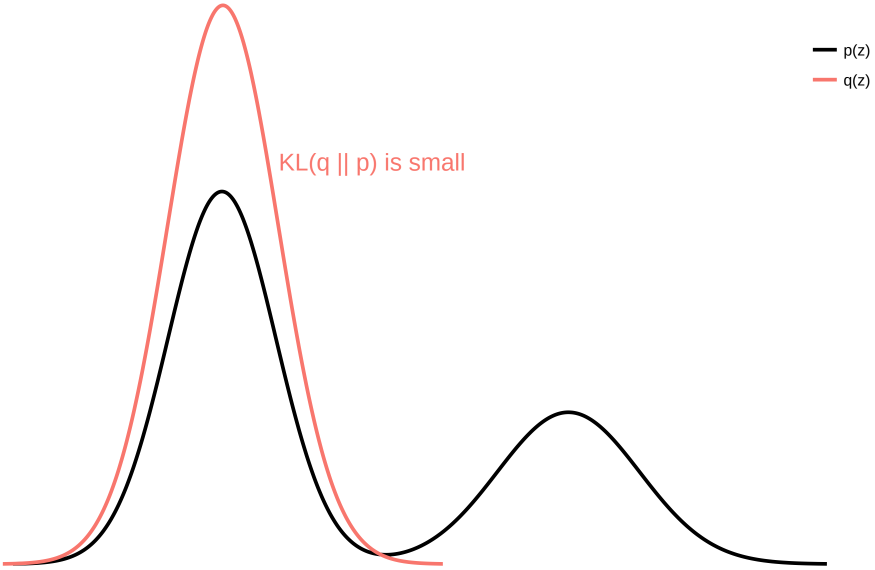
$$\mathrm{KL}(p\|q) = \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{y})} p(\mathbf{z}|\mathbf{y}) \, \mathrm{d}\mathbf{z}.$$

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

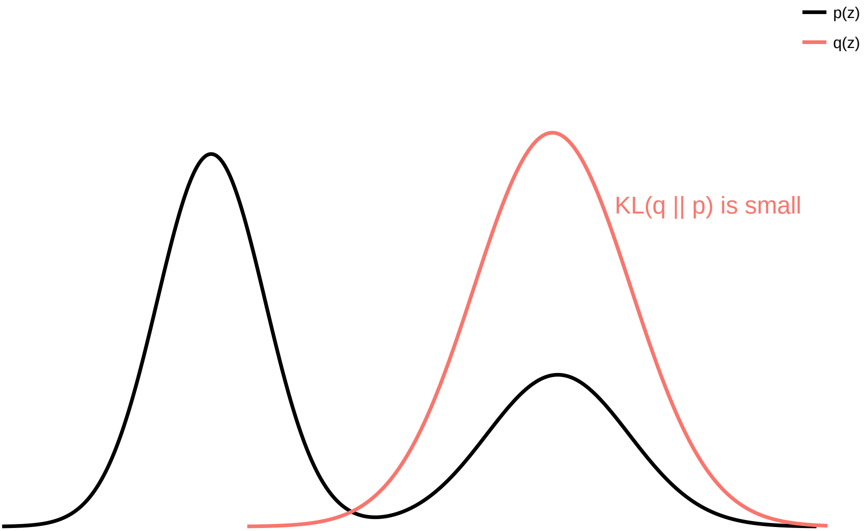
Zero-forcing vs Zero-avoiding (cont.)



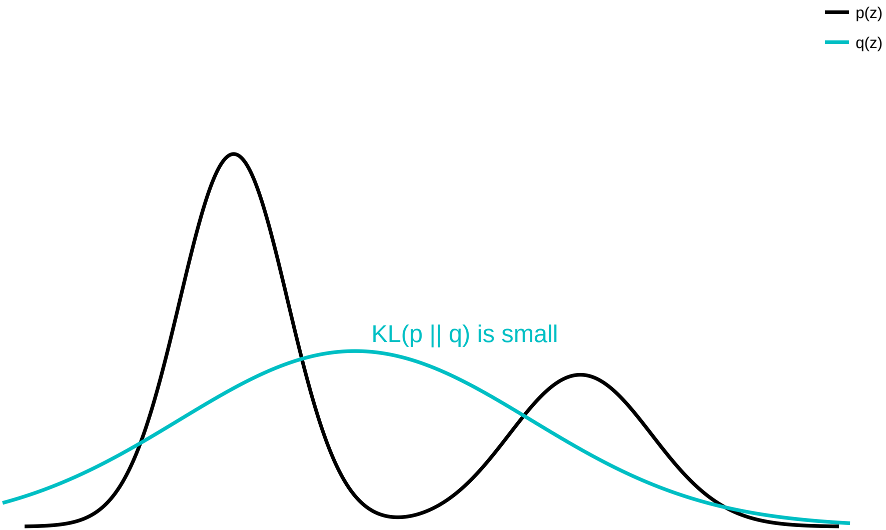
Zero-forcing vs Zero-avoiding (cont.)



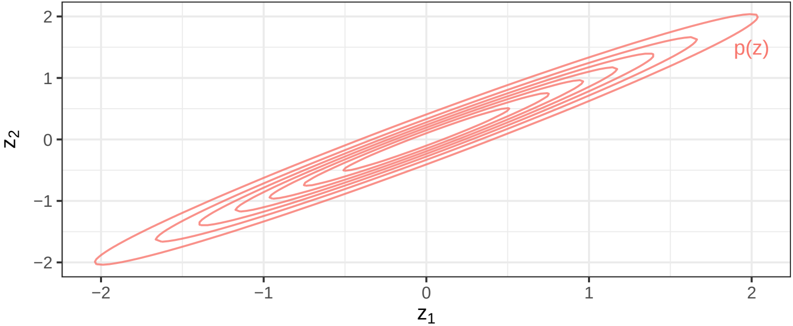
Zero-forcing vs Zero-avoiding (cont.)



Zero-forcing vs Zero-avoiding (cont.)

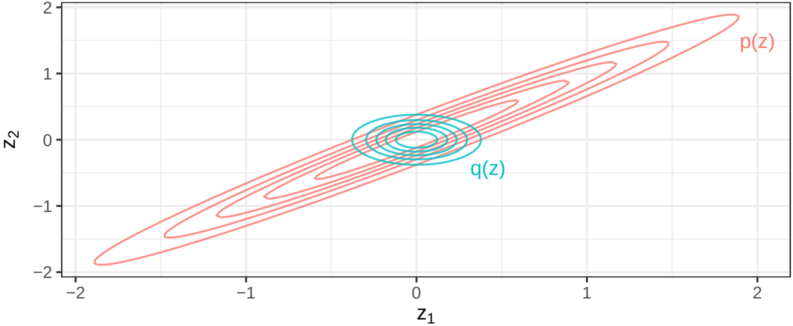


Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.

Distortion of higher order moments

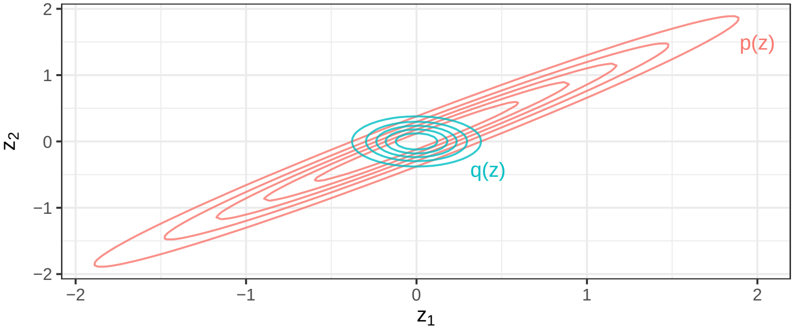


- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$ yields

$$\tilde{q}_1(z_1) = \mathrm{N}(z_1|\mu_1, \psi_{11}^{-1}) \quad \text{and} \quad \tilde{q}_2(z_2) = \mathrm{N}(z_2|\mu_2, \psi_{22}^{-1})$$

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

Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$ yields

$$\tilde{q}_1(z_1) = \mathrm{N}(z_1|\mu_1, \psi_{11}^{-1}) \text{ and } \tilde{q}_2(z_2) = \mathrm{N}(z_2|\mu_2, \psi_{22}^{-1})$$

  and by definition, $\mathrm{Cov}(z_1, z_2) = 0$ 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

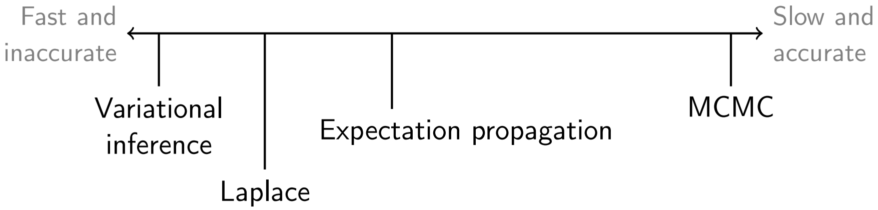
- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne, 2005).
- But not much can be said about the quality of approximation.

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?



End

Thank you!

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

References I

Beal, M. J. (2003). "Variational algorithms for approximate Bayesian inference". PhD thesis. Gatsby Computational Neuroscience Unit, University College London.

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*, 112.518, pp. 859–877.

Gunawardana, A. and W. Byrne (2005). "Convergence theorems for generalized alternating minimization procedures". *Journal of Machine Learning Research* 6, pp. 2049–2073.

References II

- 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.
- Jamil, H. (Oct. 2018). "Regression modelling using priors depending on Fisher information covariance kernels (I-priors)". PhD thesis. London School of Economics and Political Science.
- Zhao, H. and P. Marriott (2013). "Diagnostics for variational Bayes approximations". arXiv: 1309.5117.

The variational principle

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

$$\begin{array}{lll} \text{Functions} & p : \theta \mapsto \mathbb{R} & \text{(standard calculus)} \\ \text{Functionals} & \mathcal{H} : p \mapsto \mathbb{R} & \text{(variational calculus)} \end{array}$$

back

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.

back

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.

back

Comparison to the EM algorithm

- In addition to latent variables $\mathbf{z}$, typically there are unknown parameters θ to be estimated.
  - Frequentist estimation: θ is fixed
  - Bayesian estimation: $\theta \sim p(\theta)$ is random
- Consider θ fixed. Maximising the (marginal) log-likelihood directly

$$\arg\max_\theta \; \log \left\{ \int \overbrace{p(\mathbf{y}|\mathbf{z},\theta)p(\mathbf{z}|\theta)}^{p(\mathbf{y},\mathbf{z})} \, \mathrm{d}\mathbf{z} \right\}$$

  is difficult. However, if somehow the latent variables were known, then the problem may become easier.
- Given initial values $\theta^{(0)}$, the EM algorithm cycles through
  - **E-step**: Compute $Q(\theta|\theta^{(t)}) := \mathrm{E}_{\mathbf{z}}[\log p(\mathbf{y},\mathbf{z}|\theta) \mid \mathbf{y}, \theta^{(t)}]$
  - **M-step**: $\theta^{(t+1)} \leftarrow \arg\max_\theta Q(\theta|\theta^{(t)})$

  for $t = 1, 2, \ldots$ until convergence.

Comparison to the EM algorithm (cont.)

Variational inference/Bayes	(Variational) EM algorithm
GOAL: Posterior densities for $(\mathbf{w}, \theta)$	**GOAL**: ML/MAP estimates for θ
Variational approximation for latent variables and parameters $q(\mathbf{w}, \theta) \approx p(\mathbf{w}, \theta \vert \mathbf{y})$	Variational approximation for latent variables only $q(\mathbf{w}) \approx p(\mathbf{w} \vert \mathbf{y})$
Priors required on θ	Priors not necessary for θ
Derivation can be tedious	Derivation less tedious
Inference on θ through (approximate) posterior density $q(\theta)$	Asymptotic distribution of θ not well studied; standard errors for θ not easily obtained

back

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 $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} \, \mathrm{d}\mathbf{f}$. 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 $\mathbf{A} = -\mathrm{D}^2 Q(\mathbf{f})$ 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.

back

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 $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} \,\mathrm{d}\mathbf{f}$. 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 $\mathbf{A} = -\mathrm{D}^2 Q(\mathbf{f})$ being the negative Hessian of Q evaluated at $\tilde{\mathbf{f}}$.

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

back

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 $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} \, \mathrm{d}\mathbf{f}$. 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 $\mathbf{A} = -\mathrm{D}^2 Q(\mathbf{f})$ being the negative Hessian of Q evaluated at $\tilde{\mathbf{f}}$.

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

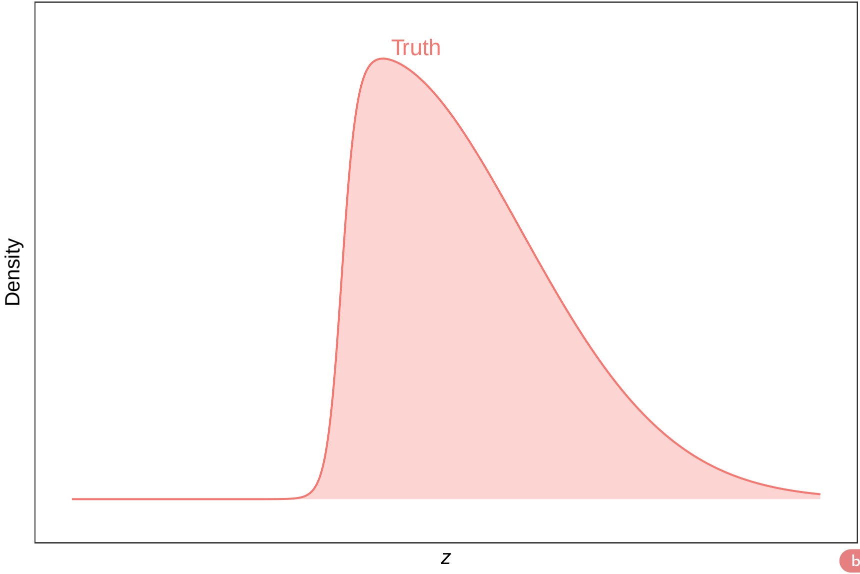
$$p(\mathbf{y}) \approx (2\pi)^{n/2} |\mathbf{A}|^{-1/2} p(\mathbf{y}|\tilde{\mathbf{f}}) p(\tilde{\mathbf{f}})$$

- 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.

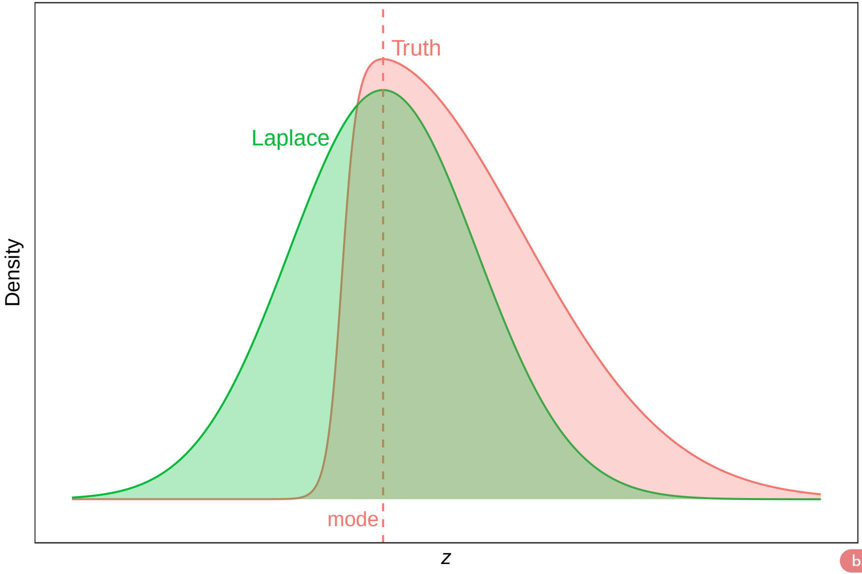
back

Comparison of approximations (density)



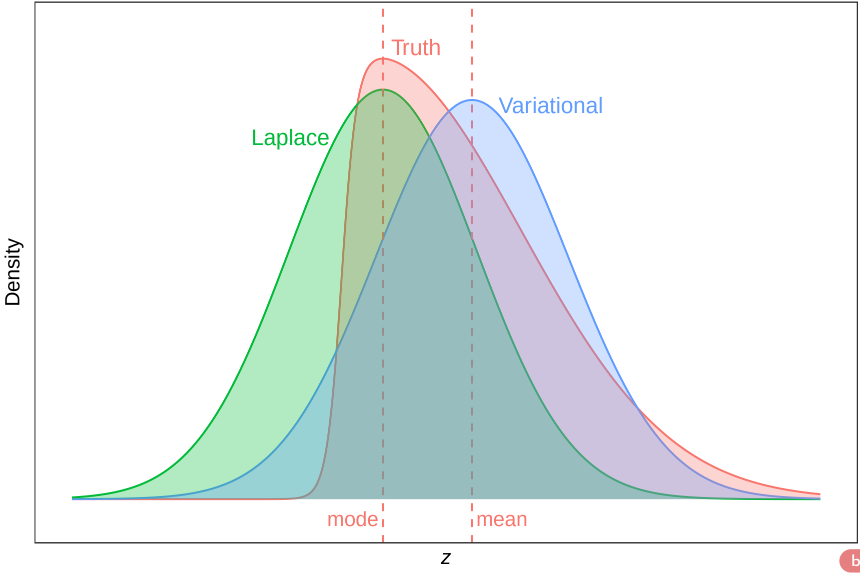
back

Comparison of approximations (density)



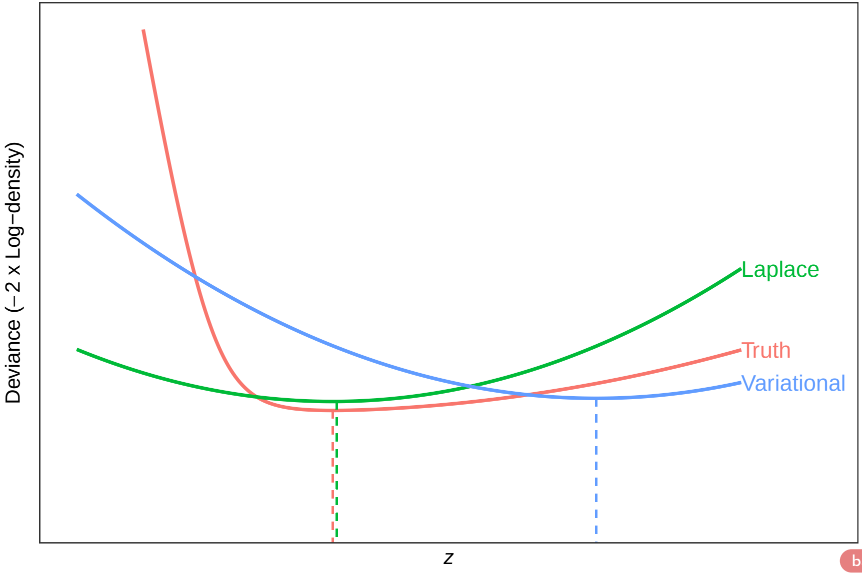
back

Comparison of approximations (density)



back

Comparison of approximations (deviance)



back

Variational solutions to Gaussian mixture model

Variational M-step

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

$$\log \rho_{ik} = \mathrm{E}[\log \pi_k] + \frac{1}{2} \mathrm{E}\left[\log |\boldsymbol{\Psi}_k| \right] - \frac{d}{2} \log 2\pi$$
$$- \frac{1}{2} \mathrm{E}\left[(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Psi}_k (\mathbf{x}_i - \boldsymbol{\mu}_k) \right]$$

Variational E-step

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

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

back

Variational solutions to Gaussian mixture model (cont.)

$$\tilde{\kappa}_k = \kappa_0 + \textstyle\sum_{i=1}^n r_{ik}$$

$$\tilde{\mathbf{m}}_k = (\kappa_0 \mathbf{m}_0 + \textstyle\sum_{i=1}^n r_{ik}\mathbf{x}_i)/\tilde{\kappa}_k$$

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + \sum_{i=1}^n r_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top$$

$$\bar{\mathbf{x}}_k = \sum_{i=1}^n r_{ik}\mathbf{x}_i \Big/ \sum_{i=1}^n r_{ik}$$

$$\nu_k = \nu_0 + \textstyle\sum_{i=1}^n r_{ik}$$

Also useful

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

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

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

back