# Session 1: introduction to variational inference

Camilla Lingjaerde and Hélène Ruffieux

MRC Biostatistics Unit, University of Cambridge, UK

Cosines + B4H masterclass on variational inference

November 7, 2022

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

# Overview

**The basics of variational inference**

- Motivation
- Main idea
- Brief history
- Mean-field variational inference
- Choice of divergence
- Relation to other inference approaches
- A flavour of some more recent trends
- Some open problems

**Practical**

- The Gaussian mixture model
- The linear regression model (variational and variational EM inference)

# Motivation

- Bayesian inference typically involves intractable integrals, e.g., in marginalisation:

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{y}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta};$$

- For instance, for the finite Gaussian mixture model:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2), & k &= 1, \ldots, K, \\ c_i &\sim \text{Categorical}(1/K, \ldots, 1/K), & i &= 1, \ldots, n, \\ y_i \mid c_i, \boldsymbol{\mu} &\sim \mathcal{N}(\mu_{c_i}, 1). \end{aligned}$$

Direct computation of the posterior is infeasible for large $n$:

$$p(\boldsymbol{\mu}, \boldsymbol{c} \mid \boldsymbol{y}) = \frac{\prod_{i=1}^{n} p(y_i \mid c_i, \boldsymbol{\mu}) p(c_i) \prod_{k=1}^{K} p(\mu_k)}{\int_{\boldsymbol{\mu}} \sum_{\boldsymbol{c}} \prod_{i=1}^{n} p(y_i \mid c_i, \boldsymbol{\mu}) p(c_i) \prod_{k=1}^{K} p(\mu_k) \mathrm{d}\boldsymbol{\mu}}.$$

# Motivation

- Bayesian computation relies on two main classes of approaches:
  (1) **Exact inference**: use Monte Carlo integration and sampling to approximate integrals;
  (2) **Approximate inference (e.g., variational inference)**: reframe Bayesian inference as an optimisation problem.
- We are increasingly confronted with "large *n*" and/or "large *p*" problems, where computational scalability is critical $\longrightarrow$ sampling methods can be impractical.

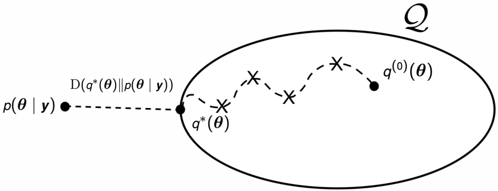
# Main idea

- Turn sampling into optimisation;
- Variational inference involves two ingredients:
  - a "restricted" variational family $\mathcal{Q}$ of "simpler" densities to approximate the posterior;
  - a measure of dissimilarity $\mathrm{D}$ between two probability distributions.

## General approach

(1) Propose a variational family $\mathcal{Q}$;

(2) Find $q(\cdot) \in \mathcal{Q}$ that is closest to $p(\cdot \mid \boldsymbol{y})$ in terms of the dissimilarity $\mathrm{D}$.



# Main idea (cont'd)

- If we let $\mathrm{D}$ be the reverse Kullback-Leibler $(\mathrm{KL})$ divergence (Kullback and Leibler, 1951), the "optimal" distribution is then

$$\arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q \| p),$$

where

$$\mathrm{KL}(q \| p) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_q \left\{ \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right\}.$$

- Properties:
  (1) $\mathrm{KL}(q \| p) \geq 0$ (non-negativity);
  (2) $\mathrm{KL}(q \| p) = 0$ iff $q = p$;
  (3) $\mathrm{KL}(q \| p) \neq \mathrm{KL}(p \| q)$.

# Main idea (cont'd)

...annoyingly the reverse $\mathrm{KL}$ divergence still depends on the marginal likelihood $p(\boldsymbol{y})$. Indeed,

$$
\begin{aligned}
\mathrm{KL}(q \| p) &= \mathbb{E}_q\{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q\{\log p(\boldsymbol{\theta} \mid \boldsymbol{y})\} \\
&= \mathbb{E}_q\{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} + \mathbb{E}_q\{\log p(\boldsymbol{y})\} \\
&= \mathbb{E}_q\{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} + \log p(\boldsymbol{y}) \int q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_q\{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} + \log p(\boldsymbol{y}).
\end{aligned}
$$

# Main idea (cont'd)

...annoyingly the reverse $\mathrm{KL}$ divergence still depends on the marginal likelihood $p(\boldsymbol{y})$. Indeed,

$$\begin{aligned}
\mathrm{KL}(q \| p) &= \mathbb{E}_q \{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q \{\log p(\boldsymbol{\theta} \mid \boldsymbol{y})\} \\
&= \mathbb{E}_q \{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q \{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} + \mathbb{E}_q \{\log p(\boldsymbol{y})\} \\
&= \mathbb{E}_q \{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q \{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} + \log p(\boldsymbol{y}) \int q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_q \{\log q(\boldsymbol{\theta})\} - \mathbb{E}_q \{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} + \log p(\boldsymbol{y}).
\end{aligned}$$

However, we now note that

$$\mathrm{KL}(q \| p) = \log p(\boldsymbol{y}) - \mathrm{ELBO}, \qquad \mathrm{ELBO} := \mathbb{E}_q \left\{ \log \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta})} \right\},$$

and, since $p(\boldsymbol{y})$ is constant w.r.t. to $\boldsymbol{\theta}$, minimising $\mathrm{KL}(q \| p)$ amounts to maximising $\mathrm{ELBO}$ – which is easier as ELBO *doesn't* involve $p(\boldsymbol{y})$.

# Main idea (cont'd)

- $\mathrm{ELBO}$ stands for Evidence Lower BOund, as it is a lower bound on the marginal log likelihood:

$$\log p(\boldsymbol{y}) = \mathrm{ELBO} + \mathrm{KL}(q \| p) \geq \mathrm{ELBO}.$$

# Main idea (cont'd)

- $\mathrm{ELBO}$ stands for Evidence Lower BOund, as it is a lower bound on the marginal log likelihood:

$$\log p(\boldsymbol{y}) = \mathrm{ELBO} + \mathrm{KL}(q \| p) \geq \mathrm{ELBO}.$$

- This can also be immediately seen from Jensen's inequality:

$$\log p(\boldsymbol{y}) = \log \int q(\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \geq \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} = \mathrm{ELBO}.$$

# Main idea (cont'd)

- $\mathrm{ELBO}$ stands for Evidence Lower BOund, as it is a lower bound on the marginal log likelihood:

$$\log p(\boldsymbol{y}) = \mathrm{ELBO} + \mathrm{KL}(q \| p) \geq \mathrm{ELBO}.$$

- This can also be immediately seen from Jensen's inequality:

$$\log p(\boldsymbol{y}) = \log \int q(\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \geq \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} = \mathrm{ELBO}.$$

- Heuristically, one might then use the $\mathrm{ELBO}$ as a way to select between models.

# Main idea (cont'd)

- $\mathrm{ELBO}$ stands for Evidence Lower BOund, as it is a lower bound on the marginal log likelihood:

$$\log p(\boldsymbol{y}) = \mathrm{ELBO} + \mathrm{KL}(q \| p) \geq \mathrm{ELBO}.$$

- This can also be immediately seen from Jensen's inequality:

$$\log p(\boldsymbol{y}) = \log \int q(\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \geq \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} = \mathrm{ELBO}.$$

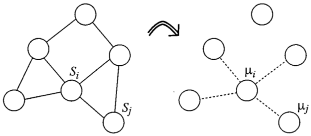
- Heuristically, one might then use the $\mathrm{ELBO}$ as a way to select between models.
- Optimising

$$\mathrm{ELBO} = \underbrace{\mathbb{E}_q \left\{ \log p(\boldsymbol{\theta}, \boldsymbol{y}) \right\}}_{\text{expected log joint}} \underbrace{- \mathbb{E}_q \left\{ q(\boldsymbol{\theta}) \right\}}_{\text{entropy}}$$

entails a trade-off between placing mass on the MAP estimate and regularising the solution.

# Brief history

- Solving problems for which exact inference is unfeasible has always been a challenge in statistics;
- Until 1999, the common approach used sampling methods such as MH, Gibbs and HMC;
- The ideas behind variational inference were developed in the field of statistical physics, where there was a pressing need for faster computation, in particular for graphical models;
- The concept first emerged in the 80s with Anderson and Peterson (1987), who developed a mean-field method to fit a neural network;
- In 1999, Jordan et al. proposed a generalised variational inference framework for probabilistic models, offering a novel approach for solving Bayesian problems.




[Jordan et al. (1999)]

## Choice of the variational family

- The optimal variational density $q(\boldsymbol{\theta})$ is the target posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ when the variational family $\mathcal{Q}$ is unrestricted;
- However restricting the variational family $\mathcal{Q}$ enhances the tractability of optimisation;
- Two common restrictions for $\mathcal{Q}$:
  (1) use some pre-specified parametric distribution, governed by a set of *variational parameters* $\boldsymbol{\eta}$, $q(\boldsymbol{\theta}; \boldsymbol{\eta})$ – e.g., a Gaussian distribution;
  (2) use the so-called *mean-field* approximation, which assumes posterior independence among the parameters: $q(\theta_1, \ldots, \theta_p) = \prod_{j=1}^{p} q_j(\theta_j)$.

# Mean-field variational inference

- The *mean-field* variational approximation (Anderson and Peterson, 1987) assumes a factorised distribution:

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j);$$

- Variational parameters under the mean-field assumption are obtained iteratively by coordinate ascent (Coordinate Ascent Variational Inference, CAVI; Jordan et al., 1999);
- Specifically, maximising the $\mathrm{ELBO}$ amounts in updating the variational factors $\{q_j(\cdot)\}_{j=1,\dots,p}$ in turn using

$$q_j(\theta_j) \propto \exp\left\{\mathbb{E}_{q_{-j}}\left[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})\right]\right\} \qquad \text{(optimal rule)},$$

  where $\boldsymbol{\theta}_{-j}$ denotes the parameter vector without component $\theta_j$, and $\mathbb{E}_{q_{-j}}(\cdot)$ is the expectation w.r.t. the factors $q_k(\cdot)$ over all $\theta_k$, $k \neq j$;
- We iteratively update the factors until convergence of either the variational factors or the $\mathrm{ELBO}$;
- Note the connection to Gibbs sampling, which involves successive draws from the full conditionals.

# Deriving the optimal solutions

- Using the chain rule and the fact that $q(\cdot)$ can be factorised, we can decompose the $\mathrm{ELBO}$:

$$\begin{aligned}\mathrm{ELBO} &= \mathbb{E}_q\left\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\right\} - \mathbb{E}_q\left\{\log q(\boldsymbol{\theta})\right\} \\ &= \log p(\boldsymbol{y}) + \sum_{j=1}^{p}\left[\mathbb{E}_q\left\{\log p(\theta_j \mid \boldsymbol{\theta}_{1:(j-1)}, \boldsymbol{y})\right\} - \mathbb{E}_{q_j}\left\{\log q_j(\theta_j)\right\}\right].\end{aligned}$$

- Considering the $\mathrm{ELBO}$ as function of $q_k(\theta_k)$, and employing the chain rule with $\theta_k$ as the last variable in the list, we get the objective function

$$\begin{aligned}\mathrm{ELBO}_k &= \mathbb{E}_q\left\{\log p(\theta_k \mid \boldsymbol{\theta}_{-k}, \boldsymbol{y})\right\} - \mathbb{E}_{q_k}\left\{\log q_k(\theta_k)\right\} + \text{const.} \\ &= \int q_k(\theta_k)\mathbb{E}_{q_{-k}}\left\{\log p(\theta_k \mid \boldsymbol{\theta}_{-k}, \boldsymbol{y})\right\} \mathrm{d}\theta_k - \int q_k(\theta_k)\log q(\theta_k)\mathrm{d}\theta_k + \text{const.},\end{aligned}$$

where the latter expression is derived using the law of total expectation.

# Deriving the optimal solutions

- Taking the derivative w.r.t. $q(\theta_k)$, we get:

$$\frac{\partial \mathrm{ELBO}_k}{\partial q_k(\theta_k)} = \mathbb{E}_{q_{-k}} \left\{ \log p(\theta_k \mid \boldsymbol{\theta}_{-k}, \boldsymbol{y}) \right\} - \log q_k(\theta_k) - 1.$$

- This (and Lagrange multipliers) leads to the coordinate ascent update for $q_k(\theta_k)$:

$$q_k(\theta_k) \propto \exp \left\{ \mathbb{E}_{q_{-k}} \left[ \log p(\theta_k \mid \boldsymbol{\theta}_{-k}, \boldsymbol{y}) \right] \right\},$$

  which is iteratively updated for $k = 1, \ldots, p$ in the CAVI algorithm.

- The resulting algorithm iteratively and monotonically maximises the $\mathrm{ELBO}$ (useful of sanity checks!), converging to a local maximum of the bound.

# Toy example: bivariate Gaussian

- We want to approximate a bivariate Gaussian distribution with a factorised mean-field approximation.
- Target distribution:

$$\boldsymbol{\theta} = (\theta_1, \theta_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \quad \boldsymbol{\mu} = (\mu_1, \mu_2), \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix},$$

  where $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are known. Note: no observed data $\boldsymbol{y}$ in this toy example.
- The variational density:

$$q(\boldsymbol{\theta}) = q_{\theta_1}(\theta_1) q_{\theta_2}(\theta_2).$$

- Using the optimal rule to find the form of the updates:

$$\begin{aligned}
\log q_{\theta_1}(\theta_1) &= \mathbb{E}_{q_{\theta_2}}\left[\log p(\theta_1 \mid \theta_2)\right] + \text{const.} = \mathbb{E}_{q_{\theta_2}}\left[\log p(\theta_1, \theta_2)\right] + \text{const.} \\
&= \mathbb{E}_{q_{\theta_2}}\left[-\frac{1}{2}(\theta_1 - \mu_1)^2 \lambda_{11} - (\theta_1 - \mu_1)\lambda_{12}(\theta_2 - \mu_2)\right] + \text{const.} \\
&= -\frac{1}{2}\theta_1^2 \lambda_{11} + \theta_1 \mu_1 \lambda_{11} - (\theta_1 - \mu_1)\lambda_{12}\left(\mathbb{E}_{q_{\theta_2}}[\theta_2] - \mu_2\right) + \text{const.}
\end{aligned}$$

# Toy example: bivariate Gaussian

- We recognise this as

$$q_{\theta_1}(\theta_1) \propto \mathcal{N}\left(m_1, \lambda_{11}^{-1}\right), \quad \text{with } m_1 = \mu_1 - \lambda_{11}^{-1}\lambda_{12}\left(\mathbb{E}_{q_{\theta_2}}[\theta_2] - \mu_2\right),$$

and similarly

$$q_{\theta_2}(\theta_2) \propto \mathcal{N}\left(m_2, \lambda_{22}^{-1}\right), \quad \text{with } m_2 = \mu_2 - \lambda_{22}^{-1}\lambda_{21}\left(\mathbb{E}_{q_{\theta_1}}[\theta_1] - \mu_1\right).$$

# Toy example: bivariate Gaussian

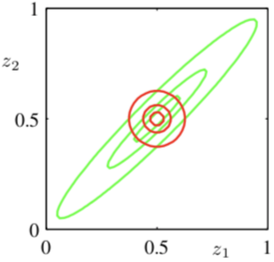
- We recognise this as

$$q_{\theta_1}(\theta_1) \propto \mathcal{N}\left(m_1, \lambda_{11}^{-1}\right), \quad \text{with } m_1 = \mu_1 - \lambda_{11}^{-1}\lambda_{12}\left(\mathbb{E}_{q_{\theta_2}}[\theta_2] - \mu_2\right),$$

and similarly

$$q_{\theta_2}(\theta_2) \propto \mathcal{N}\left(m_2, \lambda_{22}^{-1}\right), \quad \text{with } m_2 = \mu_2 - \lambda_{22}^{-1}\lambda_{21}\left(\mathbb{E}_{q_{\theta_1}}[\theta_1] - \mu_1\right).$$

- By starting with some initial $m_1$, and iteratively updating $m_1$ and $m_2$ until convergence, we obtain the factorised approximation.




[Bishop (2006)]

The resulting approximation:

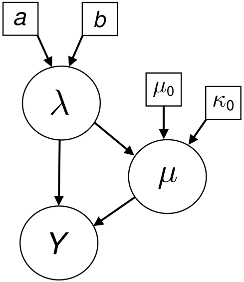
- captures the mean correctly,
- underestimates the variance,
- misses directionality.

# Example: univariate Gaussian

- We want to infer the posterior $p(\mu, \lambda \mid \boldsymbol{y})$ over the parameters $\boldsymbol{\theta} = (\mu, \lambda)$ for a univariate Gaussian, when we have $N$ observations $\boldsymbol{y} = (y_1, \ldots, y_N)$.
- Specify the generative model using a conjugate prior:

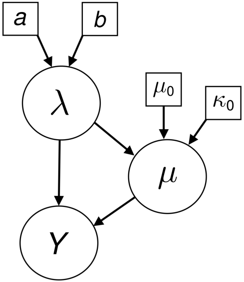
$$
\begin{aligned}
Y &\sim \mathcal{N}(\mu, \lambda^{-1}), \\
\mu &\sim \mathcal{N}(\mu_0, (\kappa_0 \lambda)^{-1}), \\
\lambda &\sim \mathrm{Gamma}(a, b),
\end{aligned}
$$

where $a$, $b$, $\kappa_0 > 0$ and $\mu_0$ are hyperparameters.



# Example: univariate Gaussian

- We want to infer the posterior $p(\mu, \lambda \mid \boldsymbol{y})$ over the parameters $\boldsymbol{\theta} = (\mu, \lambda)$ for a univariate Gaussian, when we have $N$ observations $\boldsymbol{y} = (y_1, \ldots, y_N)$.




- Specify the generative model using a conjugate prior:

$$
\begin{aligned}
Y &\sim \mathcal{N}(\mu, \lambda^{-1}), \\
\mu &\sim \mathcal{N}(\mu_0, (\kappa_0 \lambda)^{-1}), \\
\lambda &\sim \mathrm{Gamma}(a, b),
\end{aligned}
$$

  where $a$, $b$, $\kappa_0 > 0$ and $\mu_0$ are hyperparameters.

- The logarithm of the joint distribution:

$$
\begin{aligned}
\log p(\boldsymbol{y}, \boldsymbol{\theta}) = \log p(\boldsymbol{y}, \mu, \lambda) &= \log p(\boldsymbol{y} \mid \mu, \lambda) + \log p(\mu \mid \lambda) + \log p(\lambda) \\
&= \frac{N}{2} \log \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} (y_i - \mu)^2 + \frac{1}{2} \log(\kappa_0 \lambda) - \frac{\kappa_0 \lambda}{2} (\mu - \mu_0)^2 \\
&\quad + (a - 1) \log \lambda - b\lambda + \text{const}.
\end{aligned}
$$

# Example: univariate Gaussian

- The variational density:

$$q(\boldsymbol{\theta}) = q(\mu, \lambda) = q_\mu(\mu) q_\lambda(\lambda).$$

- Using the optimal rule to find the form of the updates:

$$\begin{aligned}
\log q_\mu(\mu) &= \mathbb{E}_{q_\lambda}\left[\log p(\mu \mid \lambda, \boldsymbol{y})\right] + \text{const.} = \mathbb{E}_{q_\lambda}\left[\log p(\mu, \lambda, \boldsymbol{y})\right] + \text{const.} \\
&= \mathbb{E}_{q_\lambda}\Big[\frac{N}{2}\log\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(y_i - \mu)^2 + \frac{1}{2}\log(\kappa_0\lambda) - \frac{\kappa_0\lambda}{2}(\mu - \mu_0)^2 \\
&\quad + (a-1)\log\lambda - b\lambda\Big] + \text{const.} \\
&= \mathbb{E}_{q_\lambda}\Big[-\frac{\lambda}{2}\sum_{i=1}^{N}(y_i - \mu)^2 - \frac{\kappa_0\lambda}{2}(\mu - \mu_0)^2\Big] + \text{const.} \\
&= -\frac{\mathbb{E}_{q_\lambda}[\lambda]}{2}\left(\sum_{i=1}^{N}(y_i - \mu)^2 + \kappa_0(\mu - \mu_0)^2\right) + \text{const.}
\end{aligned}$$

# Example: univariate Gaussian

- We observe that this is a quadratic function in $\mu$, implying that $q_\mu(\mu)$ is normally distributed. Completing the square, we see that the updates take the form:

$$\log q_\mu(\mu) = -\frac{(\kappa_0 + N)\mathbb{E}_{q_\lambda}[\lambda]}{2}\left(\mu - \frac{\kappa_0\mu_0 + \sum_{i=1}^N y_i}{\kappa_0 + N}\right)^2 + \text{const.}$$

which means that

$$q_\mu(\mu) \propto \mathcal{N}\left(\mu_N, \lambda_N^{-1}\right),$$

where

$$\mu_N = \frac{\kappa_0\mu_0 + \sum_{i=1}^N y_i}{\kappa_0 + N},$$
$$\lambda_N = (\kappa_0 + N)\mathbb{E}_{q_\lambda}[\lambda].$$

# Example: univariate Gaussian

- Doing the same for $\lambda$, we get

$$
\begin{aligned}
\log q_\lambda(\lambda) &= \mathbb{E}_{q_\mu}\left[\log p(\lambda \mid \mu, \boldsymbol{y})\right] + \text{const.} = \mathbb{E}_{q_\mu}\left[\log p(\lambda, \mu, \boldsymbol{y})\right] + \text{const.} \\
&= \mathbb{E}_{q_\mu}\Big[\frac{N}{2}\log\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(y_i-\mu)^2 + \frac{1}{2}\log(\kappa_0\lambda) - \frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2 \\
&\quad + (a-1)\log\lambda - b\lambda\Big] + \text{const.} \\
&= \left(a + \frac{N-1}{2} - 1\right)\log\lambda - \left(b - \frac{1}{2}\mathbb{E}_{q_\mu}\Big[\sum_{i=1}^{N}(y_i-\mu)^2 + \kappa_0(\mu-\mu_0)^2\Big]\right)\lambda + \text{const.}
\end{aligned}
$$

which we recognise as the logarithm of a Gamma distribution, yielding

$$q_\lambda(\lambda) \propto \text{Gamma}(a_n, b_N),$$

where

$$a_N = a + \frac{N-1}{2}, \qquad b_N = b + \frac{1}{2}\mathbb{E}_{q_\mu}\Big[\sum_{i=1}^{N}(y_i-\mu)^2 + \kappa_0(\mu-\mu_0)^2\Big].$$

# Example: univariate Gaussian

- Since we know the distributions of $q_\lambda(\lambda)$ and $q_\mu(\mu)$, we can easily find the expectations:

$$\mathbb{E}_{q_\mu}\left[\mu\right] = \mu_N, \qquad \mathbb{E}_{q_\mu}\left[\mu^2\right] = \frac{1}{\lambda_N} + \mu_N^2, \qquad \mathbb{E}_{q_\lambda}\left[\lambda\right] = \frac{a_N}{b_N},$$

which gives us the actual updates

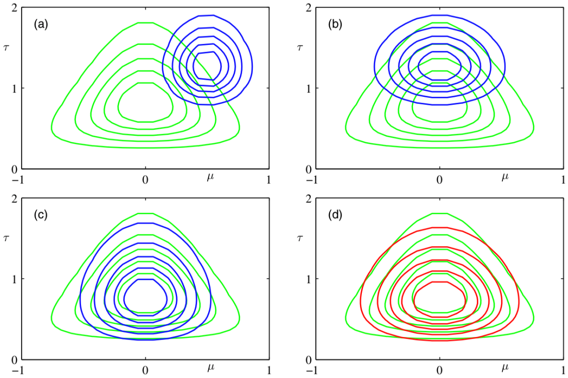
$$\mu_N = \frac{\kappa_0\mu_0 + \sum_{i=1}^N y_i}{\kappa_0 + N}, \qquad \lambda_N = \left(\kappa_0 + N\right)\frac{a_N}{b_N}, \qquad a_N = a + \frac{N-1}{2},$$

$$b_N = b + \frac{\kappa_0}{2}\left(\frac{1}{\lambda_N} + \mu_N^2 + \mu_0^2 - 2\mu_N\mu_0\right) + \frac{1}{2}\sum_{i=1}^{N}\left(y_i^2 + \frac{1}{\lambda_N} + \mu_N^2 - 2\mu_N y_i\right);$$

- By first computing $\mu_N$ and $a_N$ from the data, we can then iteratively update $\lambda_N$ and $b_N$ until convergence to obtain the parameters of $q_\mu(\mu)$ and $q_\lambda(\lambda)$;
- The $\mathrm{ELBO}$ is easily computed for each update of $\lambda_N$ and $b_N$, if we want to check it for convergence;
- We can then compute anything we want, such as the mean, variance, 95% credible intervals etc.

# Visualization of VI solution to univariate Gaussian

- Fitting the factorised approximation $q_\mu(\mu)q_\lambda(\lambda)$ (blue) to the true posterior $p(\mu, \lambda \mid \boldsymbol{y})$ (green).
- The iterative scheme continues until convergence to obtain the optimal factorised approximation (red).

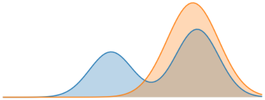



[Bishop (2006)]

# More on the KL divergence: asymmetry

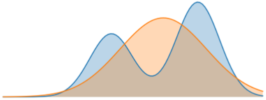
$$\arg\min_q \mathrm{KL}(q\|p) = \arg\min_q \int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x},$$

- Optimal $q$ avoids regions where $p$ is small;
- Produces a good local fit ("mode seeking");
  $\rightarrow$ pushes $q$ to underestimate the support of $p$.



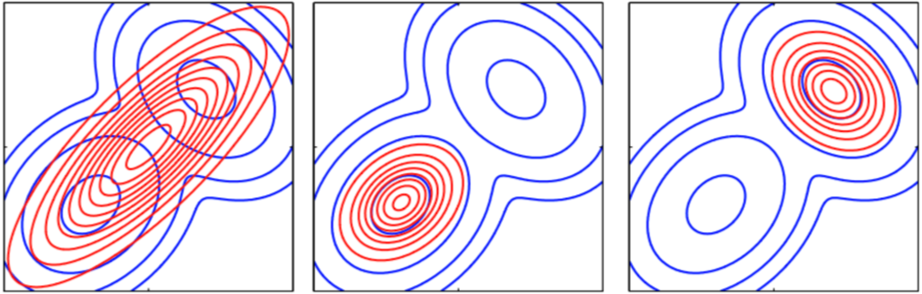
$$\arg\min_q \mathrm{KL}(p\|q) = \arg\min_q \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x},$$

- Optimal $q$ is nonzero where $p$ is nonzero (and does not care about regions where $p$ is small);
- Produces a global fit ("moment matching");
  $\rightarrow$ pushes $q$ to overestimate the support of $p$.



# Multivariate Gaussian distribution

- Blue: mixture of Gaussians $p(\boldsymbol{x})$;
- Red: optimal (unimodal) Gaussians $q(\boldsymbol{x})$;
- Global moment matching (left) versus mode seeking (middle and right).



(a) (b) (c)

[Bishop (2006)]

## Alternative divergences

- The $\mathrm{KL}$ divergence is a special case of $\alpha$-*divergences* (Rényi, 1961; Amari, 1985; Tsallis, 1988);
- Rényi's $\alpha$-divergence:

$$\mathrm{D}^{R}_{\alpha}\left(p \| q\right) = \frac{1}{\alpha - 1} \log \int p\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)^{\alpha} q\left(\boldsymbol{\theta}\right)^{1-\alpha} \mathrm{d}\boldsymbol{\theta}, \tag{1}$$

  for $\alpha \in \mathbb{R}_{+} \setminus \{1\}$ such that $\mathrm{D}^{R}_{\alpha}\left(p \| q\right) < +\infty$;
- Amari $\alpha$-divergence:

$$\mathrm{D}^{A}_{\alpha}\left(p \| q\right) = \frac{4}{1 - \alpha^{2}} \left(1 - \int p\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)^{\frac{1+\alpha}{2}} q\left(\boldsymbol{\theta}\right)^{\frac{1-\alpha}{2}} \mathrm{d}\boldsymbol{\theta}\right), \tag{2}$$

  for $\alpha \in \mathbb{R} \setminus \{\pm 1\}$ such that $\mathrm{D}^{A}_{\alpha}\left(p \| q\right) < +\infty$;
- Forward $\mathrm{KL}$: $\lim_{\alpha \to 1} \mathrm{D}^{R}_{\alpha}\left(p \| q\right) = \mathrm{KL}\left(p \| q\right), \quad \lim_{\alpha \to 1} \mathrm{D}^{A}_{\alpha}\left(p \| q\right) = \mathrm{KL}\left(p \| q\right);$
- Reverse $\mathrm{KL}$: $\lim_{\alpha \to -1} \mathrm{D}^{A}_{\alpha}\left(p \| q\right) = \mathrm{KL}\left(q \| p\right);$
- Choice of $\alpha$ leads to approximations with different behaviours but driven by practical considerations.

# Relation to other inference approaches

**Expectation Propagation (EP)** (Minka, 2013):

- minimises the forward $\mathrm{KL}$ divergence (moment-matching behaviour) over a family of tractable distributions;
- iterative algorithm leveraging factorisation structures in the posterior (convergence not guaranteed).

# Relation to other inference approaches

**Expectation Propagation (EP)** (Minka, 2013):

- minimises the forward $\mathrm{KL}$ divergence (moment-matching behaviour) over a family of tractable distributions;
- iterative algorithm leveraging factorisation structures in the posterior (convergence not guaranteed).

**Gibbs sampling** (Casella and George, 1992):

- iteratively samples from the *conditional posterior* of one variable, given all other latent variables and the observed data (exploiting conditional conjugacy);
- CAVI iteratively set each factor to *distribution of* $\theta_j \propto \exp\left\{\mathbb{E}\left[\log p(\text{conditional posterior})\right]\right\}$.

# Relation to other inference approaches

**Expectation Propagation (EP)** (Minka, 2013):

- minimises the forward $\mathrm{KL}$ divergence (moment-matching behaviour) over a family of tractable distributions;
- iterative algorithm leveraging factorisation structures in the posterior (convergence not guaranteed).

**Gibbs sampling** (Casella and George, 1992):

- iteratively samples from the *conditional posterior* of one variable, given all other latent variables and the observed data (exploiting conditional conjugacy);
- CAVI iteratively set each factor to *distribution of* $\theta_j \propto \exp\left\{\mathbb{E}\left[\log p(\text{conditional posterior})\right]\right\}$.

**Expectation Maximisation (EM)** (Dempster et al., 1977):

- alternates between taking the expectation of $\log p(\boldsymbol{\theta}, \boldsymbol{y})$ (E-step) and maximising it (M-step);
- the expected log joint distribution corresponds to the first term of the $\mathrm{ELBO}$ with the expectation taken with respect to $p(\cdot \mid \boldsymbol{y})$ instead of $q(\cdot)$.

## Couplings

- Variational inference can be coupled with other inference methods, such as the EM algorithm (VBEM) or MCMC methods (VBMC);
- For instance, VBEM (Blei et al., 2003) alternates optimisations w.r.t. $q(\cdot)$ and w.r.t. other model parameters $\boldsymbol{\eta}$ using

$$\mathrm{ELBO}\,(q;\boldsymbol{\eta}) := \mathbb{E}_q \log p(\boldsymbol{y},\boldsymbol{\theta} \mid \boldsymbol{\eta}) - \mathbb{E}_q \log q(\boldsymbol{\theta}),$$

where $q(\boldsymbol{\theta})$ is the variational density for $p\left(\boldsymbol{\theta} \mid \boldsymbol{y},\hat{\boldsymbol{\eta}}\right)$ for a current estimate $\hat{\boldsymbol{\eta}}$, i.e., it alternates between:

$$q^{(t)} = \arg\max_{q\in\mathcal{Q}} \mathrm{ELBO}\left(q;\boldsymbol{\eta}^{(t-1)}\right) \qquad \text{(E-step)},$$

using variational inference for obtaining $q^{(t)}$ at iteration $t$, and

$$\boldsymbol{\eta}^{(t)} = \arg\max_{\boldsymbol{\eta}} \mathrm{ELBO}\left(q^{(t)};\boldsymbol{\eta}\right) \qquad \text{(M-step)},$$

until convergence of $\boldsymbol{\eta}^{(t)}$.

# A flavour of some more recent trends

**Structured variational inference** (Ranganath et al., 2016):

- vanilla mean-field inference uses fully-factorised distributions: strong independence assumptions!
- structured variational inference maintains dependencies where possible.

# A flavour of some more recent trends

**Structured variational inference** (Ranganath et al., 2016):

- vanilla mean-field inference uses fully-factorised distributions: strong independence assumptions!
- structured variational inference maintains dependencies where possible.

**Stochastic variational inference** (Hoffman et al., 2013):

- scales variational inference to "large *n*" data;
- relies on stochastic optimisation (Robbins and Monro, 1951): replace the gradient with cheaper noisy estimates and guaranteed to converge to a local optimum.

# A flavour of some more recent trends

**Structured variational inference** (Ranganath et al., 2016):

- vanilla mean-field inference uses fully-factorised distributions: strong independence assumptions!
- structured variational inference maintains dependencies where possible.

**Stochastic variational inference** (Hoffman et al., 2013):

- scales variational inference to “large *n*” data;
- relies on stochastic optimisation (Robbins and Monro, 1951): replace the gradient with cheaper noisy estimates and guaranteed to converge to a local optimum.

**Black box variational inference** (Ranganath et al., 2014):

- produces generic inference, i.e., easily use variational inference with any model (no conditional conjugacy requirement);
- no mathematical work beyond specifying the model;
- uses noisy gradients and stochastic optimisation.

# Some open problems

**Theory:**

- has long seemed understudied, especially when contrasted with the theory on MCMC inference;
- mainly assumes specific models and variational families;

# Some open problems

**Theory:**

- has long seemed understudied, especially when contrasted with the theory on MCMC inference;
- mainly assumes specific models and variational families;

**Posterior variance underestimation & finite sample diagnostics:**

- can alleviate the variance underestimation issue? (Giordano et al., 2018)
- can we obtain reliable diagnostics (even in high-dimension) ?

  Pareto smoothed importance sampling (PSIS), variational simulation-based calibration diagnostic (VSBC) (Yao et al., 2018).

# Some open problems

**Theory:**

- has long seemed understudied, especially when contrasted with the theory on MCMC inference;
- mainly assumes specific models and variational families;

**Posterior variance underestimation & finite sample diagnostics:**

- can alleviate the variance underestimation issue? (Giordano et al., 2018)
- can we obtain reliable diagnostics (even in high-dimension) ?

  Pareto smoothed importance sampling (PSIS), variational simulation-based calibration diagnostic (VSBC) (Yao et al., 2018).

**Optimisation:**

- find better local optima?
- accelerate convergence?

# Exercise: Gaussian mixture model

We have the model

$$\begin{aligned}
\mu_k &\sim \mathcal{N}(0, \sigma^2), \quad k = 1, \dots, K, \\
c_i &\sim \text{Categorical}\big(1/K, \dots, 1/K\big), \quad i = 1, \dots, n, \\
Y_i \mid c_i, \boldsymbol{\mu} &\sim \mathcal{N}(\mu_{c_i}, 1),
\end{aligned}$$

where we assume $\sigma^2$ is known. Approximate the posterior

$$p(\boldsymbol{\mu}, \boldsymbol{c} \mid \boldsymbol{y}) \propto p(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{y}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(y_i | c_i, \boldsymbol{\mu}),$$

with the variational approximation

$$q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\mu_k) \prod_{i=1}^{n} q(c_i).$$

1. Derive $q(c_i) \propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c}_{-i}, \boldsymbol{\mu}}}\left[\log p(c_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu}, \boldsymbol{y})\right]\right\}$ for $i = 1, \dots, n$ and $q(\mu_k) \propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c}, \boldsymbol{\mu}_{-k}}}\left[\log p(\boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{y})\right]\right\}$ for $k = 1, \dots, K$ to obtain updates;
2. Derive the $\mathrm{ELBO} = \mathbb{E}_q\left[\log p(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{y})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\mu}, \boldsymbol{c})\right]$.

# Solution: Gaussian mixture model

We first derive

$$
\begin{aligned}
q(c_i) &\propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c}_{-i}},\boldsymbol{\mu}}\left[\log p(c_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu}, \boldsymbol{y})\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c}_{-i}},\boldsymbol{\mu}}\left[\log p(c_i) + \log(\boldsymbol{c}_{-i}) + p(\boldsymbol{\mu}) + \sum_{j=1}^{n} \log p(y_j \mid c_j, \boldsymbol{\mu})\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c}_{-i}},\boldsymbol{\mu}}\left[\log p(c_i) + \log p(y_i \mid c_i, \boldsymbol{\mu})\right]\right\} \propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c}_{-i}},\boldsymbol{\mu}}\left[\frac{1}{K} + \log p(y_i \mid \mu_{c_i})\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{\mu}}}\left[-\frac{1}{2}(y_i - \mu_{c_i})^2\right]\right\} \propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{\mu}}}\left[-\frac{1}{2}(y_i^2 - 2y_i\mu_{c_i} + \mu_{c_i}^2)\right]\right\} \propto \phi_{i,c_i},
\end{aligned}
$$

where

$$
\phi_{i,c_i} \propto \exp\left\{y_i m_{c_i} - \frac{1}{2}s_{c_i}^2 - \frac{1}{2}m_{c_i}^2\right\},\ m_{c_i} = \mathbb{E}_{q_{\mu_{c_i}}}\left[\mu_{c_i}\right],\ s_{c_i}^2 = \mathrm{Var}_{q_{\mu_{c_i}}}\left[\mu_{c_i}\right] = \mathbb{E}_{q_{\mu_{c_i}}}\left[\mu_{c_i}^2\right] - \mathbb{E}_{q_{\mu_{c_i}}}\left[\mu_{c_i}\right]^2.
$$

This gives us the distribution of the $i^{\text{th}}$ observations mixture, with $\sum_{k=1}^{K} \phi_{i,k} = 1$ for $i = 1, \ldots, n$.

# Solution: Gaussian mixture model

Then for $\mu_k$:

$$
\begin{aligned}
q(\mu_k) &\propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}}\left[\log p(\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{y})\right]\right\} \propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}}\left[\log p(\mu_k) + \sum_{i=1}^{n}\log p(y_i \mid c_i, \boldsymbol{\mu})\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{q_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}}\left[-\frac{1}{2\sigma^2}\mu_k^2 + \sum_{i=1}^{n}\mathbb{I}(c_i = k)\log p(y_i \mid \mu_k)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\mu_k^2 + \sum_{i=1}^{n}\phi_{i,k}\left[-\frac{1}{2}(y_i - \mu_k)^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\mu_k^2 - \frac{1}{2}\sum_{i=1}^{n}\phi_{i,k}\left[y_i^2 - 2y_i\mu_k + \mu_k^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma^2} + \sum_{i=1}^{n}\phi_{i,k}\right)\mu_k^2 - 2\sum_{i=1}^{n}\phi_{i,k}y_i\mu_k\right]\right\}
\end{aligned}
$$

# Solution: Gaussian mixture model

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}+\sum_{i=1}^{n}\phi_{i,k}\right)\left[\mu_k-\frac{\sum_{i=1}^{n}\phi_{i,k}y_i}{1/\sigma^2+\sum_{i=1}^{n}\phi_{i,k}}\right]^2\right\} \\
&\propto \mathcal{N}(m_k, s_k^2),
\end{aligned}
$$

with

$$
\begin{aligned}
m_k &= \frac{\sum_{i=1}^{n}\phi_{i,k}y_i}{1/\sigma^2+\sum_{i=1}^{n}\phi_{i,k}}, \\
s_k^2 &= \left(\frac{1}{\sigma^2}+\sum_{i=1}^{n}\phi_{i,k}\right)^{-1}.
\end{aligned}
$$

This gives us the updates for the $k^{\text{th}}$ component. Recalling that the $i^{\text{th}}$ observations mixture had the update $\phi_{i,c_i} \propto \exp\left\{y_i m_{c_i} - \frac{1}{2}s_{c_i}^2 - \frac{1}{2}m_{c_i}^2\right\}$, with $\sum_{k=1}^{K}\phi_{i,k} = 1$ for $i = 1, \ldots, n$, this gives us the complete CAVI updates, which we can iteratively compute to get to the local optimal and thus our inference.

# Solution: Gaussian mixture model

Finally, we derive the $\mathrm{ELBO}$:

$$\begin{aligned}
\mathrm{ELBO} &= \mathbb{E}_q\left[\log p(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{y})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\mu}, \boldsymbol{c})\right] \\
&= \mathbb{E}_q\left[\sum_{i=1}^{n} \log p(c_i) + \sum_{k=1}^{K} \log p(\mu_k) + \sum_{i=1}^{n} \log p(y_i \mid c_i, \boldsymbol{\mu})\right] \\
&\quad - \mathbb{E}_q\left[\sum_{i=1}^{n} \log q(c_i) + \sum_{k=1}^{K} \log q(\mu_k)\right] \\
&= \mathbb{E}_q\left[\sum_{i=1}^{n} \log \frac{1}{K} - \frac{1}{2\sigma^2}\sum_{k=1}^{K} \mu_k^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{I}(c_i = k)\left(y_i - \mu_k\right)^2\right] \\
&\quad - \sum_{i=1}^{n}\sum_{k=1}^{K} \phi_{i,k} \log \phi_{i,k} - \mathbb{E}_q\left[\sum_{k=1}^{K}\left[-\frac{1}{2}\log s_k^2 - \frac{1}{2s_k^2}\left(\mu_k - m_k\right)^2\right]\right] + \text{const.}
\end{aligned}$$

# Solution: Gaussian mixture model

$$
\begin{aligned}
&= -\frac{1}{2\sigma^2}\sum_{k=1}^{K}\left[s_k^2 + m_k^2\right] - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\phi_{i,k}\mathbb{E}_q\left[y_i^2 - 2y_i\mu_k + \mu_k^2\right] \\
&\quad - \sum_{i=1}^{n}\sum_{k=1}^{K}\phi_{i,k}\log\phi_{i,k} - \mathbb{E}_q\left[\sum_{k=1}^{K}\left[-\frac{1}{2}\log s_k^2 - \frac{1}{2}\right]\right] + \text{const.} \\
&= -\frac{1}{2\sigma^2}\sum_{k=1}^{K}\left[s_k^2 + m_k^2\right] + \sum_{i=1}^{n}\sum_{k=1}^{K}\phi_{i,k}\left[y_i m_k - \frac{1}{2}s_k^2 - \frac{1}{2}m_k^2\right] \\
&\quad - \sum_{i=1}^{n}\sum_{k=1}^{K}\phi_{i,k}\log\phi_{i,k} + \frac{1}{2}\sum_{k=1}^{K}\log s_k^2 + \text{const.}
\end{aligned}
$$

After each iteration, we compute the $\mathrm{ELBO}$ using the updates for $s_k^2$, $m_k$, $\phi_{i,k}$ to check for convergence.

# Further reading on the basics of variational inference

Bishop (2006): Pattern recognition and machine learning

Blei et al. (2017): Variational inference: a review for statisticians

Zhang et al. (2018): Advances in variational inference

Ganguly and Earp (2021): An introduction to variational inference

Practical

Solutions can be found at: `www.github.com/Camiling/B4H_Masterclass_VI`.

# 1) Gaussian mixture model

Recall the Gaussian mixture model from last session

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2), \quad k = 1, \ldots, K, \\ c_i &\sim \text{Categorical}(1/K, \ldots, 1/K), \quad i = 1, \ldots, n, \\ Y_i \mid c_i, \boldsymbol{\mu} &\sim \mathcal{N}(\mu_{c_i}, 1), \end{aligned}$$

for which we had the mean-field variational approximation factors

$$q(\mu_k) \propto \mathcal{N}(m_k, s_k^2),\ q(c_i) \propto \phi_{i,c_i} \propto \exp\left\{ y_i m_{c_i} - \frac{1}{2} s_{c_i}^2 - \frac{1}{2} m_{c_i}^2 \right\},$$

where

$$m_k = \frac{\sum_{i=1}^n \phi_{i,k} y_i}{1/\sigma^2 + \sum_{i=1}^n \phi_{i,k}}, \quad s_k^2 = \left( \frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{i,k} \right)^{-1}, \quad \sum_{k=1}^K \phi_{i,k} = 1,$$

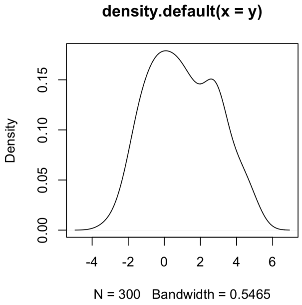
and the $\mathrm{ELBO}$ was derived to be

$$\mathrm{ELBO} = -\frac{1}{2\sigma^2} \sum_{k=1}^K \left[ s_k^2 + m_k^2 \right] + \sum_{i=1}^n \sum_{k=1}^K \phi_{i,k} \left[ y_i m_k - \frac{1}{2} s_k^2 - \frac{1}{2} m_k^2 \right] - \sum_{i=1}^n \sum_{k=1}^K \phi_{i,k} \log \phi_{i,k} + \frac{1}{2} \sum_{k=1}^K \log s_k^2 + \text{const.}$$

Implement the CAVI algorithm, and run it on simulated data with $K = 3$, $\boldsymbol{\mu} = (-1, 1, 3)$, $\sigma^2 = 1$, $n = 300$. Use the initialisation $\phi_{i,c_i} = 1/K$ for all $i = 1, \ldots, n$, and $m_1 = 1$, $m_2 = 2$, $m_3 = 3$ and $s_k^2 = 0.5$ for all $k = 1, \ldots, K$. Use the $\mathrm{ELBO}$ to assess convergence and estimate 95% credible intervals for $\mu_k$, $k = 1, \ldots, K$ from their estimated distribution. What happens if you instead initialise all $m_k$ with the same value?

# 1) Gaussian mixture model - generating the data

```r
set.seed(123)
K = 3
mu = c(-1,1,3)
sig.mu = 1
tau.mu = 1/sig.mu^2
sig2 = 1
n1 = 100
n = K*n1
y = rep(NA,n)
eps = 0.001
for(k in 1:K){
  y[(k-1)*n1+1:n1] = rnorm(n1,mu[k],sqrt(sig2))
}
plot(density(y))
```



# 1) Gaussian mixture model - computing VI approximation

```
phi = matrix(1/K,nrow=n,ncol=K); m = c(1,2,3); s2 = rep(0.5,K)
more = TRUE; Elbo = 0
while(more){
  for(i in 1:n){
    phi[i,] = exp(m*y[i]-0.5*s2-0.5*m^2)
    phi[i,] = phi[i,]/sum(phi[i,])
  }
  for(k in 1:K){
    m[k] = sum(phi[,k]*y)/(tau.mu+sum(phi[,k]))
    s2[k] = 1/(tau.mu+sum(phi[,k]))
  }
  elbo = -0.5*tau.mu*sum(s2+m^2)-sum(rowSums(phi*log(phi)))+0.5*sum(log(
    s2))
  for(k in 1:K){
    elbo = elbo + sum(phi[,k]*(y*m[k]-0.5*s2[k]-0.5*m[k]^2))
  }
  more = abs(tail(Elbo,n=1)-elbo)>eps
  Elbo = c(Elbo,elbo)
}
qnorm(c(0.025, 0.975), m[1], sqrt(s2[1])) # 95% CI for mu_1
```

# 2) Linear regression model

We have the model

$$
\begin{aligned}
y_i \mid \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \phi^{-1}), \quad i = 1, \dots, n, \\
\boldsymbol{\beta} \mid \kappa &\sim \mathcal{N}(0, \kappa^{-1} \boldsymbol{I}), \\
\kappa &\sim \text{Gamma}(a_0, b_0),
\end{aligned}
$$

where $\phi = 1/\sigma^2$ is the precision parameter, which we assume is known, $\boldsymbol{x}_i,\ i = 1, \dots, n$ are known covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ includes the intercept, and is unknown, and $\boldsymbol{I}$ is the identity matrix. Assume $a_0$ and $b_0$ are known. Find a variational approximation to the posterior

$$
p(\boldsymbol{\beta}, \kappa \mid \boldsymbol{y})
$$

on the form

$$
q(\boldsymbol{\beta}, \kappa) = q(\boldsymbol{\beta}) q(\kappa)
$$

and derive the CAVI updates. Implement the algorithm, and run on simulated data with one covariate $x_{i,1} \sim \mathcal{N}(0, 1)$, $n = 50$, $\phi = 0.5$, $\beta_0 = -1, \beta_1 = 2$, $a_0 = b_0 = 0.001$. Assess convergence by the variational factors, or derive the $\mathrm{ELBO}$ to assess convergence. Visualise the resulting bivariate Gaussian approximation for the intercept $\beta_0$ and coefficient $\beta_1$.

# 3) Linear regression with empirical Bayes estimation for hyperparameters

We assume the same model as in the previous exercise

$$\begin{aligned} y_i \mid \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \phi^{-1}), \quad i = 1, \ldots, n, \\ \boldsymbol{\beta} \mid \kappa &\sim \mathcal{N}(0, \kappa^{-1} \boldsymbol{I}), \\ \kappa &\sim \text{Gamma}(a_0, b_0), \end{aligned}$$

where $\boldsymbol{x}_i, \ i = 1, \ldots, n$ are known covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ includes the intercept, and is unknown, and $\boldsymbol{I}$ is the identity matrix. We assume $a_0$ and $b_0$ are known. However, we now assume the precision parameter $\phi = 1/\sigma^2$ is unknown, and must be estimated.

Instead of the fully variational approach, use VBEM to estimate the posterior by treating $\phi$ as a hyperparameter to update in the M-step. Implement the algorithm, and run on simulated data with one covariate $x_{i,1} \sim \mathcal{N}(0, 1)$, $n = 50$, $\beta_0 = -1, \beta_1 = 2$, $a_0 = b_0 = 0.001$. Visualise the resulting bivariate Gaussian approximation for the intercept $\beta_0$ and coefficient $\beta_1$, and compare to the one you obtained in exercise 2. Play around with different initial values for $\phi$ - is this choice important?

# References


S.-I. Amari. *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer-Verlag, New York, United States, 1985.

J. R. Anderson and C. Peterson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, United States, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B (methodological)*, 39:1–38, 1977.

A. Ganguly and S. W. F. Earp. An introduction to variational inference. *arXiv preprint arXiv:2108.13083*, 2021.

R. Giordano, T. Broderick, and M. I. Jordan. Covariances, robustness and variational bayes. *Journal of Machine Learning Research*, 19, 2018.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14:1303–1347, 2013.

M. I. Jordan, Z. Ghahramani, T. S Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

T. P. Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In S. Kaski and J. Corander, editors, *Artificial Intelligence and Statistics*, pages 814–822, Reykjavik, Iceland, 2014. Proceedings of Machine Learning Research.

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333. PMLR, 2016.

A. Rényi. On measures of entropy and information. Technical report, Hungarian Academy of Sciences Budapest Hungary, 1961.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

C. Tsallis. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR, 2018.

C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2008–2026, 2018.