THE UNIVERSITY *of* EDINBURGH

Lecture 2: Variational Inference

Dr Sara Wade
sara.wade@ed.ac.uk

20-23 May, 2024

Outline

Outline

Introduction

Bayesian statistics frames all inference about unknown quantities as a calculation involving the posterior density.

Modern Bayesian methods are characterized by complex models and difficult-to-compute posteriors $\rightarrow$ approximating posteriors is a core problem.

MCMC is the dominant paradigm and an indispensable tool for approximating the posterior.

However, there are problems for which we cannot easily use MCMC and need an approximate posterior faster (e.g. for large data sets or complex models/multimodal posteriors).

Setup

Consider the joint density of latent variables $\mathbf{z} = (z_1, \ldots, z_D)$ and observations $\boldsymbol{x} = (x_1, \ldots x_N)$:

$$p(\mathbf{z}, \mathbf{x}) = \underbrace{p(\mathbf{z})}_{\text{prior}} \underbrace{p(\mathbf{x} \mid \mathbf{z})}_{\text{likelihood}}.$$

Inference in a Bayesian model amounts to computing:

$$\underbrace{p(\mathbf{z}|\mathbf{x})}_{\text{posterior}} = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})},$$

where $p(\mathbf{x}) = \int p(\mathbf{z}, \mathbf{x}) d\mathbf{x}$ is the marginal likelihood (or evidence).

MCMC produces (dependent) draws $\mathbf{z}^{(t)}$, which are asymptotically exact samples from the posterior.

Variational inference (VI): uses optimization rather than sampling; faster but lose asymptotic guarantees.

Outline

Variational Inference

Step 1 posit a family of approximate densities $\mathcal{Q}$; which is a set of densities over $\mathbf{z}$.

Step 2 find $q^* \in \mathcal{Q}$ that minimizes the Kullback-Leibler (KL) divergence to the exact posterior $p(\mathbf{z} \mid \boldsymbol{x})$:

$$q^*(\mathbf{z}) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}\left(q(\mathbf{z}) || p(\mathbf{z} \mid \boldsymbol{x})\right).$$

$\rightarrow$ VI turns the inference problem into an optimization problem; the key is to choose $\mathcal{Q}$ to be flexible enough to closely approximate the posterior and simple enough for efficient optimization.

Evidence Lower Bound (ELBO)

Computing the KL divergence is not tractable because it requires computing the evidence:

$$\begin{aligned}\mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z} \mid \boldsymbol{x})\right) &= \mathrm{E}\left[\log q(\mathbf{z})\right] - \mathrm{E}\left[\log p(\mathbf{z} \mid \boldsymbol{x})\right] \\ &= -\int \log\left(\frac{p(\mathbf{z}, \boldsymbol{x})}{q(\mathbf{z})p(\boldsymbol{x})}\right) q(\mathbf{z})\, d\mathbf{z} \\ &= \mathrm{E}\left[\log q(\mathbf{z})\right] - \mathrm{E}\left[\log p(\mathbf{z}, \boldsymbol{x})\right] + \log p(\boldsymbol{x}).\end{aligned}$$

However, we can equivalently optimize an alternative objective that is equivalent to the KL up to a constant:

$$\mathrm{ELBO}\left(q\right) = \mathrm{E}\left[\log p(\mathbf{z}, \boldsymbol{x})\right] - \mathrm{E}\left[\log q(\mathbf{z})\right],$$

which is called the evidence lower bound (ELBO), as

$$\log p(\boldsymbol{x}) = \mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z} \mid \boldsymbol{x})\right) + \mathrm{ELBO}\left(q\right) \geq \mathrm{ELBO}\left(q\right).$$

Evidence Lower Bound (ELBO)

We can rewrite the ELBO as:

$$\begin{aligned}\text{ELBO}\,(q) &= \mathrm{E}\left[\log p(\mathbf{z})\right] + \mathrm{E}\left[\log p(\boldsymbol{x} \mid \mathbf{z})\right] - \mathrm{E}\left[\log q(\mathbf{z})\right] \\ &= \underbrace{\mathrm{E}\left[\log p(\boldsymbol{x} \mid \mathbf{z})\right]}_{\text{expected log likelihood}} - \underbrace{\mathrm{KL}\left(q(\mathbf{z}) || p(\mathbf{z})\right)}_{\text{KL between prior and posterior}} .\end{aligned}$$

$\rightarrow$ balances between likelihood and prior.

Note: sometimes the ELBO is used for model selection and comparison, under the assumption that it provides a good approximation of the evidence but caution is advised.

Variational Family

There are two main ways in which q can be restricted to a class of tractable distributions:

- Parametric forms: specify a parametric form $q_\gamma(\mathbf{z})$, with parameters γ, e.g. $q_\gamma(\mathbf{z})$ is Gaussian with variational parameters γ containing the mean and variance:

$$q_{\gamma^*}(\mathbf{z}) = \underset{\gamma}{\operatorname{argmin}} \, \mathrm{KL}\left(q_\gamma(\mathbf{z}) || p(\mathbf{z} \mid \boldsymbol{x})\right).$$

- Mean-field: assume that the variational posterior factorizes[1] over (blocks) of latent variables:

$$q(\mathbf{z}) = \prod_{d=1}^{D} q(z_d).$$

[1] Expansions include *structured variational inference*, which keeps some dependencies (Saul and Jordan (1999)) and *mixtures of variational densities*, which allow latent variables within the variational family (Bishop et al (1998)).

Mean-field approximation

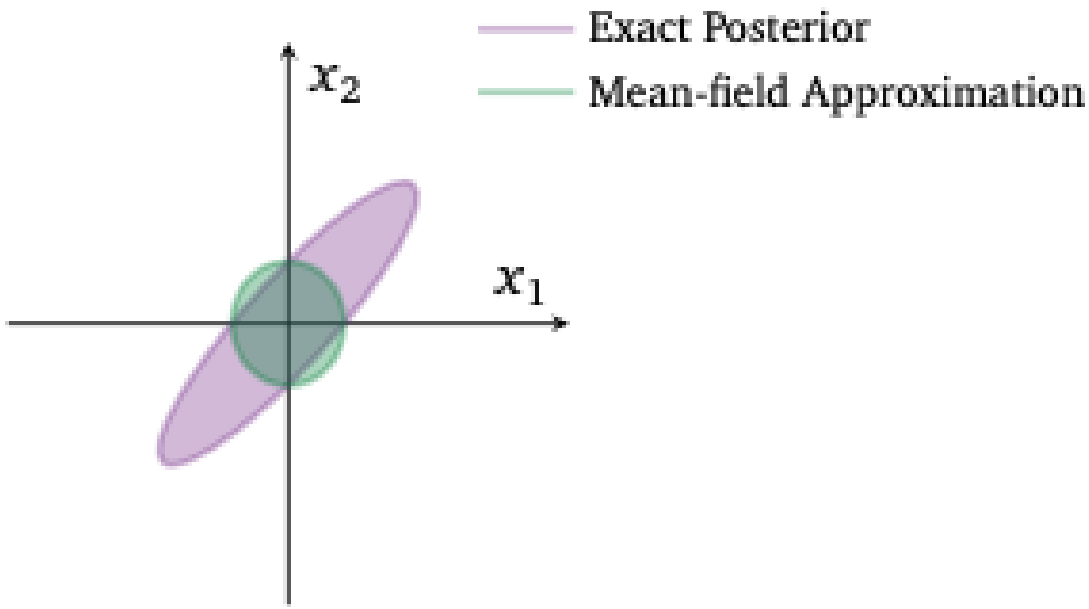

Figure: From Blei et al (2018). Visualizing the mean-field approximation to a two-dimensional Gaussian posterior. The optimal mean-field approximation is a product of two Gaussians 🤔; it has the same mean but the covariance is decoupled resulting in underestimation of the variance.

Mean-field approximation

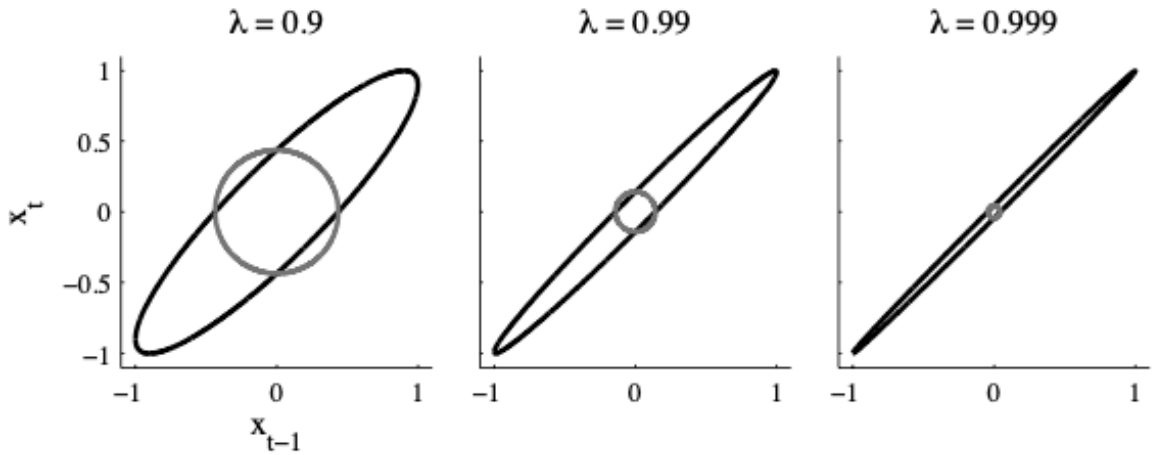

Figure: From Turner & Sahani (2008). Comparing the true prior-predictive to the mean-field approximation for a simple dynamic linear model. The marginal variance of the mean-field approximation is tiny for typical values of λ in time-series.

Coordinate ascent variational inference (CAVI)

Coordinate ascent variational inference (CAVI) (Bishop, 2006) is one of the most commonly used algorithms for solving the mean-field variational optimization problem; it iteratively optimizes each factor, while holding all others fixed.

For the mean-field family, the optimal choice is:

$$q_d(z_d) \propto \exp\left[\mathrm{E}_{\mathbf{z}_{-d}} \log\{p(z_d \mid \mathbf{z}_{-d}, \mathbf{x})\}\right],$$

where $\mathbf{z}_{-d} = (z_1, \ldots, z_{d-1}, z_{d+1}, \ldots, z_D)$ and the expectation is taken with respect to $\prod_{d' \neq d} q_{d'}(z_{d'})$.

Coordinate ascent variational inference (CAVI)

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$

Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Initialize: Variational factors $q_j(z_j)$

while *the* ELBO *has not converged* **do**

  for $j \in \{1, \ldots, m\}$ **do**

    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$

  end

  Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$

end

return $q(\mathbf{z})$

Figure: From Blei et al (2018). CAVI is closely related to Gibbs sampling; instead of sampling, CAVI takes the expected log to set each variable's variational factor.

Note: Collapsed variational inference is developed from similiar reasoning to collapsed Gibbs sampling (Sung et al (2008); Hensman et al (2012)).

Coordinate ascent variational inference (CAVI)

Derivation: the ELBO of the dth factor is:

$$\begin{aligned}
&\text{ELBO}\left(q_d\right) = \mathrm{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] - \mathrm{E}_{z_d}\left[\log q_d(\mathbf{z}_d)\right] + \text{const} \\
&= \mathrm{E}_{z_d}\left[\mathrm{E}_{\mathbf{z}_{-d}}\left[\log p(z_d, \mathbf{z}_{-d}, \mathbf{x})\right]\right] - \mathrm{E}_{z_d}\left[\log q_d(\mathbf{z}_d)\right] + \text{const} \\
&= \mathrm{E}_{z_d}\left[\log\left(\exp\left(\mathrm{E}_{\mathbf{z}_{-d}}\left[\log p(z_d, \mathbf{z}_{-d}, \mathbf{x})\right]\right)\right)\right] - \mathrm{E}_{z_d}\left[\log q_d(\mathbf{z}_d)\right] + \text{const}.
\end{aligned}$$

The first term is the negative KL divergence, which is maximized by setting:

$$q_d(z_d) \propto \exp\left[\mathrm{E}_{\mathbf{z}_{-d}} \log\{p(z_d \mid \mathbf{z}_{-d}, \mathbf{x})\}\right].$$

Practicalities

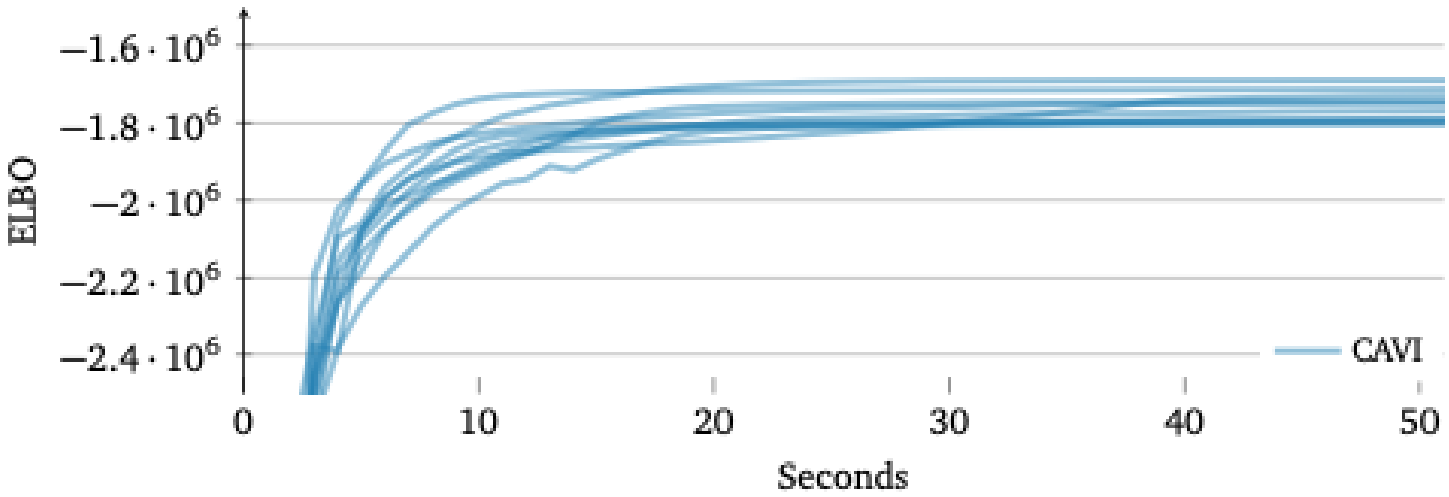

Figure: From Blei et al (2018).

- CAVI only guarantees convergence to a local optima and can be sensitive to initialization. In practice, multiple intializations are suggested.
- Convergence can be assessed by monitoring the ELBO, and stopping early once the change in ELBO has fallen below some threshold.

Outline

Example: Variational Bayesian (VB) Lasso

Recall the hierarchical representation of the Bayesian lasso:

$$\begin{aligned} y_n \mid \mathbf{x}_n, \boldsymbol{w}, \sigma^2 &\sim \mathrm{N}(y_n \mid \boldsymbol{w}^T \mathbf{x}_n, \sigma^2), \\ \boldsymbol{w} \mid \sigma^2, \boldsymbol{\tau} &\sim \mathrm{N}(0, \sigma^2 \mathbf{D}), \\ \tau_d &\sim \mathrm{Gamma}(1, 1/(2s^2)), \\ \sigma^2 &\sim \mathrm{IG}(\alpha_\sigma, \beta_\sigma), \end{aligned}$$

where $\mathbf{D} = \mathrm{diag}(\tau_1, \ldots, \tau_D)$.

Assume the mean-field variational family:

$$q(\boldsymbol{w}, \sigma^2, \boldsymbol{\tau}) = q(\boldsymbol{w}) q(\sigma^2) q(\boldsymbol{\tau}).$$

VB Lasso

Similar to Gibbs sampling, VB lasso[2] cycles through the steps:

Step 1 VB posterior of $\boldsymbol{w}$:

$$q(\boldsymbol{w}) \propto \exp\left[\mathrm{E}_{\boldsymbol{\tau},\sigma^2} \log\{\underbrace{p(\boldsymbol{w} \mid \boldsymbol{\tau}, \sigma^2, \mathcal{D})}_{N_D(\mathbf{w}|\mathbf{S}\mathbf{X}^T\mathbf{y},\sigma^2\mathbf{S})}\}\right],$$

$$\rightarrow q(\boldsymbol{w}) = \mathrm{N}_D(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w),$$

where $\widehat{\mathbf{D}}^{-1} = \mathrm{diag}(\mathrm{E}[1/\tau_1], \ldots, \mathrm{E}[1/\tau_D])$,

$$\boldsymbol{\Sigma}_w = \frac{1}{\mathrm{E}[1/\sigma^2]}\left(\widehat{\mathbf{D}}^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}, \quad \text{and} \quad \boldsymbol{\mu}_w = \left(\widehat{\mathbf{D}}^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

Note to update the other parameters, we require $\mathrm{E}[\boldsymbol{w}] = \boldsymbol{\mu}_w$ and $\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w\boldsymbol{\mu}_w^T$.

[2] Armagan (2009); Law & Zankin (2021)

VB Lasso

Step 2 VB posterior of σ^2:

$$q(\sigma^2) \propto \exp\left[\mathrm{E}_{\boldsymbol{\tau},\mathbf{w}} \log\{\underbrace{p(\sigma^2 \mid \boldsymbol{\tau}, \mathbf{w}, \mathcal{D})}_{\mathrm{IG}(\widehat{\alpha}_\sigma, \widehat{\beta}_\sigma)}\}\right],$$

$$\rightarrow q(\sigma^2) = \mathrm{IG}(\widehat{\alpha}_\sigma, \tilde{\beta}_\sigma),$$

where $\widehat{\alpha}_\sigma = \alpha_\sigma + N/2 + D/2$ and

$$\tilde{\beta}_\sigma = \beta_\sigma + 1/2\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathrm{E}[\boldsymbol{w}] + \mathrm{Tr}\left(\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^T][\mathbf{X}^T\mathbf{X} + \widehat{\mathbf{D}}^{-1}]\right)\right).$$

Note to update the other parameters, we require $\mathrm{E}[1/\sigma^2] = \widehat{\alpha}_\sigma/\widehat{\beta}_\sigma$.

VB Lasso

Step 3 VB posterior of $\boldsymbol{\tau}$:

$$q(\boldsymbol{\tau}) \propto \exp\left[\mathrm{E}_{\sigma^2,\mathbf{w}} \log\{ \prod_{d=1}^{D} \underbrace{p(\tau_d \mid \sigma^2, \mathbf{w}, \mathcal{D})}_{1/\mathrm{InvGaus}\left(\sqrt{\frac{\sigma^2}{s^2 w_d^2}}, 1/s^2\right)} \} \right],$$

$$\rightarrow q(1/\boldsymbol{\tau}) = \prod_{d=1}^{D} \mathrm{InvGaus}(\widehat{\mu}_d, \widehat{\lambda}),$$

where $\widehat{\lambda} = 1/s^2$ and

$$\widehat{\mu}_d = \left(s\sqrt{\mathrm{E}[1/\sigma^2]\mathrm{E}[w_d^2]} \right)^{-1}.$$

Note to update the other parameters, we require $\mathrm{E}[1/\tau_d] = \widehat{\mu}_d$.

VB Lasso

🤔 Computing the ELBO:

$$
\begin{aligned}
\text{ELBO} &= \mathrm{E}[\log p(\mathbf{y} \mid \boldsymbol{w}, \sigma^2)] + \mathrm{E}[\log p(\boldsymbol{w} \mid \boldsymbol{\tau}, \sigma^2)] + \mathrm{E}[\log p(\boldsymbol{\tau})] \\
&\quad + \mathrm{E}[\log p(\sigma^2)] - \mathrm{E}[\log q(\boldsymbol{w})] + \mathrm{E}[\log q(\boldsymbol{\tau})] + \mathrm{E}[\log q(\sigma^2)] \\
&= \text{const} - \frac{1}{2}\mathrm{E}[1/\sigma^2]\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathrm{E}[\boldsymbol{w}] + \mathrm{Tr}\left(\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^T][\mathbf{X}^T\mathbf{X} + \widehat{\mathbf{D}}^{-1}]\right)\right) \\
&\quad + \frac{1}{2}\log|\boldsymbol{\Sigma}_w| - D\log(s) - \sum_{d=1}^{D}\frac{1}{2s^2\widehat{\mu}_d} - \beta_\sigma\mathrm{E}[1/\sigma^2] - \widehat{\alpha}_\sigma\log(\widehat{\beta}_\sigma).
\end{aligned}
$$

VB Lasso

What about s?

- 🤔 Hierarchical VB Lasso: extend to allow a variational posterior on s.
- Empirical VB Lasso: set s to maximize the ELBO; this is known as variational expectation maximization (VEM).

VEM: set hyperparameters $\boldsymbol{\theta}$ to maximize the ELBO

$$\boldsymbol{\theta}^* = \underset{\theta}{\operatorname{argmax}} \, \mathrm{E}\left[\log p(\mathbf{z}, \mathbf{x} \mid \theta)\right] - \mathrm{E}\left[\log q(\mathbf{z})\right],$$

and are taken as approximate maximum marginal likelihood values (🤔 we can also consider as assuming the variational posterior is a point mass).

Empirical VB Lasso

Maximizing the ELBO wrt s equivalently maximizes the objective:

$$-D\log(s) - \sum_{d=1}^{D} \frac{1}{2s^2\widehat{\mu}_d}.$$

🤔 Show that

$$s = \sqrt{\frac{1}{D}\sum_{d=1}^{D} \frac{1}{\widehat{\mu}_d}}.$$

Outline

Full Conditionals in the Exponential Family

For a general model $p(\mathbf{z}, \mathbf{x})$, suppose that each full conditional is in the exponential family:

$$p(z_d \mid \mathbf{z}_{-d}, \mathbf{x}) = h(z_d) \exp(\eta_d(\mathbf{z}_{-d}, \mathbf{x})^T z_d - \alpha(\eta_d(\mathbf{z}_{-d}, \mathbf{x})),$$

where z_d is it own sufficient statistic, $h(\cdot)$ is a base measure, and $\alpha(\cdot)$ is the log normalizer.

Under the mean-field assumption, the coordinate update for z_d is

$$\begin{aligned} q(z_d) &\propto \exp\left(\mathrm{E}\left[\log p(z_d \mid \mathbf{z}_{-d}, \mathbf{x})\right]\right) \\ &\propto h(z_d) \exp(\mathrm{E}\left[\eta_d(\mathbf{z}_{-d}, \mathbf{x})\right]^T z_d). \end{aligned}$$

$\Rightarrow$ This provides a general recipe for deriving CAVI algorithms for *conditional conjugate models*.

Stochastic Variational Inference

CAVI may not scale to massive data, particularly if iterating through the entire data at each iteration is required.

Gradient-based optimization is an alternative to coordinate ascent that climbs the ELBO by computing and following its gradient at each iteration.

Stochastic variational inference[3] combines natural gradients with stochastic optimization.

[3] Hoffman et al (2013)

Stochastic Variational Inference

Basic idea: consider the model

$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{n=1}^{N} p(z_n, x_n \mid \beta),$$

where β is a *global* parameter and z_n are *local* parameters.

Consider the mean field variational posterior $q_{\boldsymbol{\lambda}}(\beta) \prod_{n=1}^{N} q_{\boldsymbol{\gamma}_n}(z_n)$.

Using gradient-based optimization, update $\boldsymbol{\lambda}$ by setting:

$$\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \epsilon_t g(\boldsymbol{\lambda}_{t-1}),$$

where $g(\boldsymbol{\lambda})$ is the *natural gradient* of the ELBO $\Rightarrow$ But computing the gradient has the same cost as CAVI!

Stochastic Variational Inference

Stochastic optimization: use noisy but cheap-to-compute, unbiased estimates of the gradient.

For example, in the case of the conditionally conjugate model, the natural gradient of the ELBO can be written as:

$$g(\boldsymbol{\lambda}) = \boldsymbol{\alpha} + N \left(1/N \sum_{n=1}^{N} \mathrm{E}_{\boldsymbol{\gamma}_n}[f(z_n, x_n)], 1 \right)^T - \boldsymbol{\lambda}.$$

We can construct a noisy, cheap, unbiased natural gradient:

$$\widehat{g}(\boldsymbol{\lambda}) = \boldsymbol{\alpha} + N \left(\mathrm{E}_{\boldsymbol{\gamma}_n}[f(z_n, x_n)], 1 \right)^T - \boldsymbol{\lambda},$$

with $n \sim \mathrm{Unif}(1, \ldots, N)$ (extends to minibatches).

Beyond conjugacy

The non-conjugate setting typically requires parametric assumptions on the form of $q_{\boldsymbol{\gamma}}(\mathbf{z})$ (e.g. Gaussian, neural network) and/or tailored variational bounds and methods to approximate difficult to compute variational approximations.

Stochastic variational inference can also be useful in this setting by employing Monte Carlo estimates of the gradients.

This is employed in automatic differentiation variational inference (ADVI)[4] in `Stan`, which assumes a Gaussian variational posterior in the *unconstrained variational parameter space*.

[4] Kucukelbir et al (2017)

Beyond conjugacy

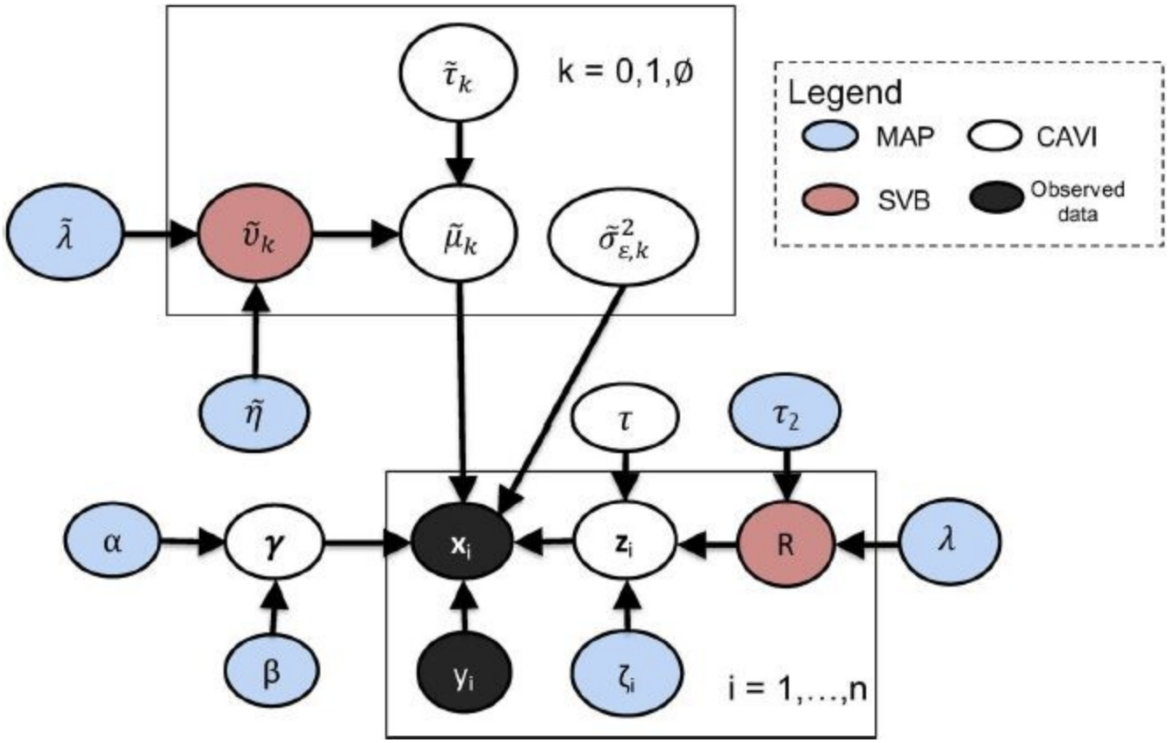

Figure: From Yu et al (2022). Different types of updates/inference can be combined.

Summary

- VI is an alternative to MCMC that provides faster, approximate inference.
- Applications: stemmed from computer science to many areas (biology, neuroscience, computer vision, ...)
- Theory: some results for different model-variational family combinations, particularly for the variational mean (e.g. Ormerod et al (2014); Ray and Szabo (2020)), however uncertainty estimation is challenging.
- Other research areas explore other measures, e.g. α-divergence, and combine variational inference with other schemes (e.g. MCMC).

Further Reading

See various references within, including:

- *Variational inference: A review for statisticians*, Blei et al (2018).
- *Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies*, Neville et al (2014).
- *Yes, but Did It Work?: Evaluating Variational Inference*, Yao et al (2018).