Bayesian modelling

Variational inference

Léo Belzile

Last compiled Monday Mar 24, 2025

Variational inference

Laplace approximation provides a heuristic for large-sample approximations, but it fails to characterize well $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

We consider rather a setting where we approximate p by another distribution g which we wish to be close. The terminology **variational** is synonym for optimization.

Kullback–Leibler divergence

The Kullback–Leibler divergence between densities $f_t(\cdot)$ and $g(\cdot;\boldsymbol{\psi})$, is

$$
\begin{aligned}
\mathsf{KL}(f_t \parallel g) &= \int \log\left(\frac{f_t(\boldsymbol{x})}{g(\boldsymbol{x};\boldsymbol{\psi})}\right) f_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int \log f_t(\boldsymbol{x}) f_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int \log g(\boldsymbol{x};\boldsymbol{\psi}) f_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \mathsf{E}_{f_t}\{\log f_t(\boldsymbol{X})\} - \mathsf{E}_{f_t}\{\log g(\boldsymbol{X};\boldsymbol{\psi})\}
\end{aligned}
$$

The negative entropy does not depend on $g(\cdot)$.

Model misspecification

- The divergence is strictly positive unless $g(\cdot;\boldsymbol{\psi}) \equiv f_t(\cdot)$.
- The divergence is not symmetric.

The Kullback–Leibler divergence notion is central to study of model misspecification.

- if we fit $g(\cdot)$ when data arise from f_t, the maximum likelihood estimator of the parameters $\widehat{\boldsymbol{\psi}}$ will be the value of the parameter that minimizes the Kullback–Leibler divergence $\mathsf{KL}(f_t \parallel g)$.

Marginal likelihood

Consider now the problem of approximating the marginal likelihood, sometimes called the evidence,

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

where we only have the joint $p(\boldsymbol{y}, \boldsymbol{\theta})$ is the product of the likelihood times the prior.

Reformulation of the marginal density

Consider $g(\boldsymbol{\theta}; \boldsymbol{\psi})$ with $\boldsymbol{\psi} \in \mathbb{R}^J$ an approximating density function whose integral is one over $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ and whose support is part of that of $\mathrm{supp}(g) \subseteq \mathrm{supp}(p) = \boldsymbol{\Theta}$:

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{g(\boldsymbol{\theta}; \boldsymbol{\psi})} g(\boldsymbol{\theta}; \boldsymbol{\psi}) \mathrm{d}\boldsymbol{\theta}.$$

Bounding the marginal likelihood

For $h(x)$ a convex function, **Jensen’s inequality** implies that

$$h\{\mathsf{E}(X)\} \leq \mathsf{E}\{h(X)\},$$

and applying this with $h(x) = -\log(x)$, we get

$$-\log p(\boldsymbol{y}) \leq -\log\left(\frac{p(\boldsymbol{y},\boldsymbol{\theta})}{g(\boldsymbol{\theta};\boldsymbol{\psi})}\right)g(\boldsymbol{\theta};\boldsymbol{\psi})\mathrm{d}\boldsymbol{\theta}.$$

Evidence lower bound

We can thus consider the model that minimizes the **reverse Kullback–Leibler divergence**

$$g(\boldsymbol{\theta}; \widehat{\boldsymbol{\psi}}) = \mathrm{argmin}_{\boldsymbol{\psi}} \mathsf{KL}\{g(\boldsymbol{\theta}; \boldsymbol{\psi}) \parallel p(\boldsymbol{\theta}, \boldsymbol{y})\}.$$

Consider the reformulation

$$\begin{aligned}\mathsf{KL}\{g(\boldsymbol{\theta}; \boldsymbol{\psi}) \parallel p(\boldsymbol{\theta}, \boldsymbol{y})\} = \mathsf{E}_g\{\log g(\boldsymbol{\theta})\} - \mathsf{E}_g\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\} \\ + \log p(\boldsymbol{y}).\end{aligned}$$

Evidence lower bound

Instead of minimizing the Kullback–Leibler divergence, we can equivalently maximize the so-called **evidence lower bound** (ELBO)

$$f\mathsf{ELBO}(g) = \mathsf{E}_g\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\} - \mathsf{E}_g\{\log g(\boldsymbol{\theta})\}$$

The ELBO is a lower bound for the marginal likelihood because a Kullback–Leibler divergence is non-negative and

$$\log p(\boldsymbol{y}) = \mathsf{ELBO}(g) + \mathsf{KL}\{g(\boldsymbol{\theta}; \boldsymbol{\psi}) \parallel p(\boldsymbol{\theta}, \boldsymbol{y})\}.$$

Use of ELBO

The idea is that we will approximate the density $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx g(\boldsymbol{\theta}; \widehat{\boldsymbol{\psi}})$.

- the ELBO can be used for model comparison (but we compare bounds...)
- we can sample from q as before.

Heuristics of ELBO

Maximize the evidence, subject to a regularization term:

$$\mathsf{ELBO}(g) = \mathsf{E}_g\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\} - \mathsf{E}_g\{\log g(\boldsymbol{\theta})\}$$

The ELBO is an objective function comprising:

- the first term will be maximized by taking a distribution placing mass near the MAP of $p(\boldsymbol{y}, \boldsymbol{\theta})$,
- the second term can be viewed as a penalty that favours high entropy of the approximating family (higher for distributions which are diffuse).

Laplace vs variational approximation

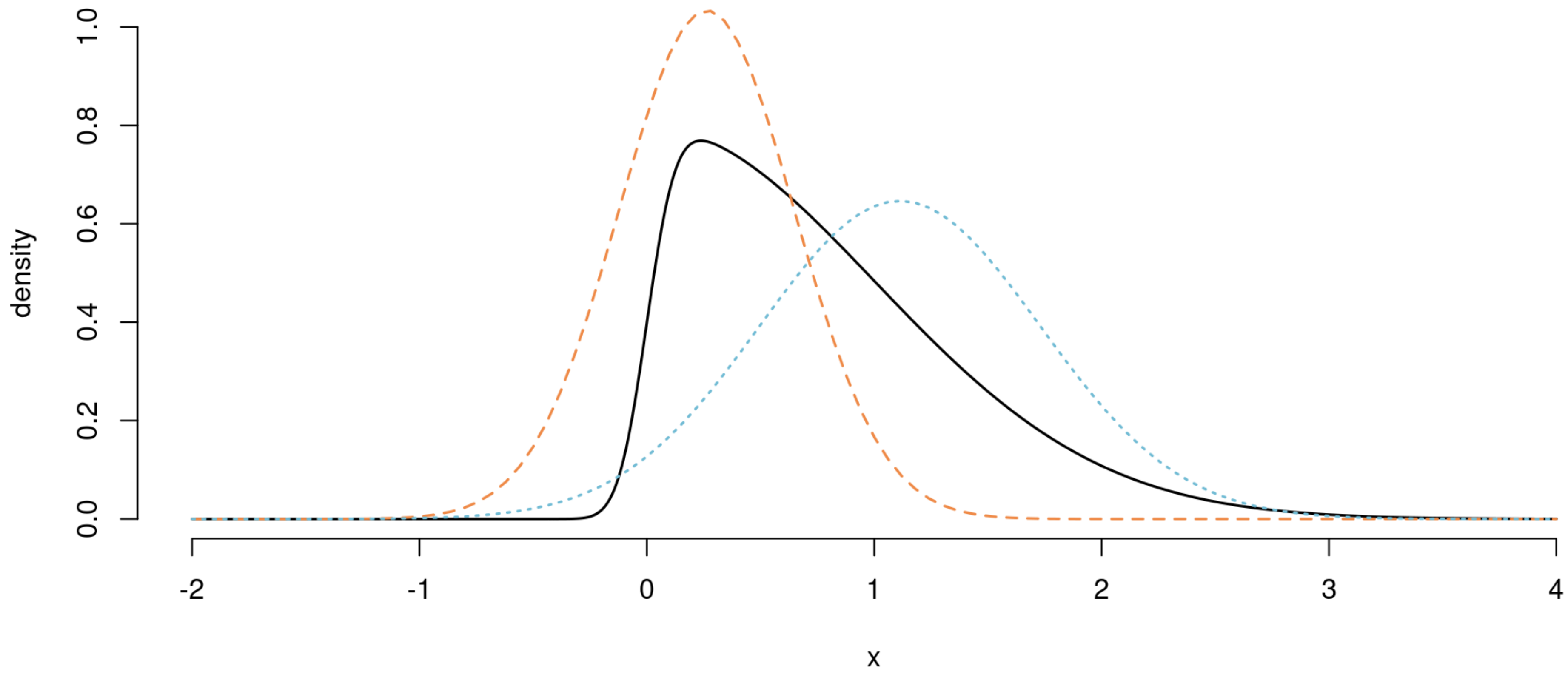


Figure 1: Skewed density with the Laplace approximation (dashed orange) and variational Gaussian approximation (dotted blue).

Choice of approximating density

In practice, the quality of the approximation depends on the choice of $g(\cdot;\boldsymbol{\psi})$.

- We typically want matching support.
- The approximation will be affected by the correlation between posterior components $\boldsymbol{\theta} \mid \boldsymbol{y}$
- Derivations can also be done for $(\boldsymbol{U}, \boldsymbol{\theta})$, where $\boldsymbol{U}$ are latent variables from a data augmentation scheme.

Factorization

We can consider densities $g(;\boldsymbol{\psi})$ that factorize into blocks with parameters $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M$, where

$$g(\boldsymbol{\theta};\boldsymbol{\psi}) = \prod_{j=1}^{M} g_j(\boldsymbol{\theta}_j;\boldsymbol{\psi}_j)$$

If we assume that each of the J parameters $\theta_1, \ldots, \theta_J$ are independent, then we obtain a **mean-field** approximation.

Optimal form of approximating density

$$\begin{aligned}\mathsf{ELBO}(g) &\overset{\boldsymbol{\theta}_i}{\propto} \int \mathsf{E}_{-i}\left\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\right\} g_i(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i \\ &\quad - \int \log\{g_i(\boldsymbol{\theta}_i)\} g_i(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i\end{aligned}$$

The choice of g_i that maximizes the ELBO is

$$g_i^{\star}(\boldsymbol{\theta}_i) \propto \exp\left[\mathsf{E}_{-i}\left\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\right\}\right].$$

Often, we look at the kernel of $g_j^{\star}$ to deduce the normalizing constant.

Coordinate-ascent variational inference (CAVI)

- We can maximize $g_j^\star$ in turn for each $j = 1, \ldots, M$ treating the other parameters as fixed.
- This scheme is guaranteed to monotonically increase the ELBO until convergence to a local maximum.
- Convergence: monitor ELBO and stop when the change is lower then some present numerical tolerance.
- The approximation may have multiple local optima: perform random initializations and keep the best one.

Example of CAVI mean-field for Gaussian target

We consider the example from Section 2.2.2 of Ormerod & Wand (2010) for approximation of a Gaussian distribution, with

$$\begin{aligned} Y_i &\sim \textsf{Gauss}(\mu, \tau^{-1}), \qquad i = 1, \ldots, n; \\ \mu &\sim \textsf{Gauss}(\mu_0, \tau_0^{-1}) \\ \tau &\sim \textsf{gamma}(a_0, b_0). \end{aligned}$$

This is an example where the full posterior is available in closed-form, so we can compare our approximation with the truth.

Variational approximation to Gaussian — mean

We assume a factorization of the variational approximation $g_\mu(\mu)g_\tau(\tau)$; the factor for g_μ is proportional to

$$\log g^\star_\mu(\mu) \propto -\frac{\mathsf{E}_\tau(\tau)}{2}\sum_{i=1}^n (y_i-\mu)^2 - \frac{\tau_0}{2}(\mu-\mu_0)^2$$

which is quadratic in μ and thus must be Gaussian with precision $\tau_n = \tau_0 + n\tau$ and mean $\tau_n^{-1}\{\tau_0\mu_0 + \mathsf{E}_\tau(\tau)n\overline{y}\}$

Variational approximation to Gaussian — precision

The optimal precision factor satisfies

$$\begin{aligned}\ln g_{\tau}^{\star}(\tau) &\propto (a_0 - 1 + n/2)\log \tau \\ &\qquad -\tau\left[b_0 + \frac{1}{2}\mathsf{E}_{\mu}\left\{\sum_{i=1}^{n}(y_i-\mu)^2\right\}\right].\end{aligned}$$

This is of the same form as $p(\tau \mid \mu, \boldsymbol{y})$, namely a gamma with shape $a_n = a_0 + n/2$ and rate b_n.

Rate of the gamma for g_τ

It is helpful to rewrite the expected value as

$$\mathsf{E}_\mu\left\{\sum_{i=1}^{n}(y_i-\mu)^2\right\}=\sum_{i=1}^{n}\{y_i-\mathsf{E}_\mu(\mu)\}^2+n\mathsf{Var}_\mu(\mu),$$

so that it depends on the parameters of the distribution of μ directly.

CAVI for Gaussian

The algorithm cycles through the following updates until convergence:

- $\mathsf{Va}_{\mu}(\mu) = \{\tau_0 + n\mathsf{E}_{\tau}(\tau)\}^{-1}$,
- $\mathsf{E}_{\mu}(\mu) = \mathsf{Va}_{\mu}(\mu)(\tau_0\mu_0 + \mathsf{E}_{\tau}(\tau)n\overline{y})$,
- $\mathsf{E}_{\tau}(\tau) = a_n/b_n$ where b_n is a function of $\mathsf{E}_{\mu}(\mu)$ and $\mathsf{Var}_{\mu}(\mu)\}$.

We only compute the ELBO at the end of each cycle.

Monitoring convergence

The derivation of the ELBO is straightforward but tedious; we only need to monitor

$$-\frac{\tau_0}{2}\mathsf{E}_{\mu}\{(\mu-\mu_0)^2\}-\frac{\log\tau_n}{2}-a_n\log b_n$$

for convergence, although other normalizing constants would be necessary if we wanted to approximate the marginal likelihood.

Stochastic optimization

We consider alternative numeric schemes which rely on stochastic optimization (Hoffman et al., 2013).

The key idea behind these methods is that

- we can use gradient-based algorithms,
- and approximate the expectations with respect to g by drawing samples from it

Also allows for minibatch (random subset) selection to reduce computational costs in large samples

Black-box variational inference

Ranganath et al. (2014) shows that the gradient of the ELBO reduces to

$$\frac{\partial}{\partial \boldsymbol{\psi}} \mathsf{ELBO}(g) = \mathsf{E}_g \left\{ \frac{\partial \log g(\boldsymbol{\theta}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \times \log \left(\frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{g(\boldsymbol{\theta}; \boldsymbol{\psi})} \right) \right\}$$

using the change rule, differentiation under the integral sign (dominated convergence theorem) and the identity

$$\frac{\partial \log g(\boldsymbol{\theta}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} g(\boldsymbol{\theta}; \boldsymbol{\psi}) = \frac{\partial g(\boldsymbol{\theta}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$$

Black-box variational inference in practice

- Note that the gradient simplifies for g_i in exponential families (covariance of sufficient statistic with $\log(p/g)$).
- The gradient estimator is particularly noisy, so Ranganath et al. (2014) provide two methods to reduce the variance of this expression using control variates and Rao–Blackwellization.

Automatic differentiation variational inference

Kucukelbir et al. (2017) proposes a stochastic gradient algorithm, but with two main innovations.

- The first is the general use of Gaussian approximating densities for factorized density, with parameter transformations to map from the support of $T : \boldsymbol{\Theta} \mapsto \mathbb{R}^p$ via $T(\boldsymbol{\theta}) = \boldsymbol{\zeta}$.
- The second is to use the resulting **location-scale** family to obtain an alternative form of the gradient.

Gaussian full-rank approximation

Consider an approximation $g(\zeta; \boldsymbol{\psi})$ where $\boldsymbol{\psi}$ consists of

- mean parameters $\boldsymbol{\mu}$ and
- covariance $\boldsymbol{\Sigma}$, parametrized through a Cholesky decomposition

The full approximation is of course more flexible when the transformed parameters ζ are correlated, but is more expensive to compute than the mean-field approximation.

Change of variable

The change of variable introduces a Jacobian term $\mathbf{J}_{T^{-1}}(\boldsymbol{\zeta})$ for the approximation to the density $p(\boldsymbol{\theta}, \boldsymbol{y})$, where

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{\zeta}, \boldsymbol{y})\, |\mathbf{J}_{T^{-1}}(\boldsymbol{\zeta})|$$

Gaussian entropy

The entropy of the multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix, is

$$\mathbb{H}(\boldsymbol{L}) = -\mathsf{E}_g(\log g) = \frac{D + D\log(2\pi) + \log\left|\mathbf{L}\mathbf{L}^\top\right|}{2},$$

and only depends on $\mathbf{\Sigma}$.

ELBO with Gaussian approximation

Since the Gaussian is a location-scale family, we can rewrite the model in terms of a standardized Gaussian variable $\boldsymbol{Z} \sim \mathsf{Gauss}_p(\mathbf{0}_p, \mathbf{I}_p)$ where $\boldsymbol{\zeta} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{Z}$ (this transformation has unit Jacobian).

The ELBO with the transformation becomes

$$\mathsf{E}_{\boldsymbol{Z}}\left[\log p\{\boldsymbol{y}, T^{-1}(\boldsymbol{\zeta})\} + \log \left|\mathbf{J}_{T^{-1}}(\boldsymbol{\zeta})\right|\right] + \mathbb{H}(\boldsymbol{L}).$$

Chain rule

If $\boldsymbol{\theta} = T^{-1}(\zeta)$ and $\boldsymbol{\zeta} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$, we have for $\boldsymbol{\psi}$ equal to either $\boldsymbol{\mu}$ or $\mathbf{L}$ that

$$\begin{aligned}
&\frac{\partial}{\partial \psi} \log p(\boldsymbol{y}, \boldsymbol{\theta}) \\
&\quad = \frac{\partial \log p(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \times \frac{\partial T^{-1}(\boldsymbol{\zeta})}{\partial \zeta} \times \frac{\partial(\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z})}{\partial \boldsymbol{\psi}}
\end{aligned}$$

Gradients for ADVI

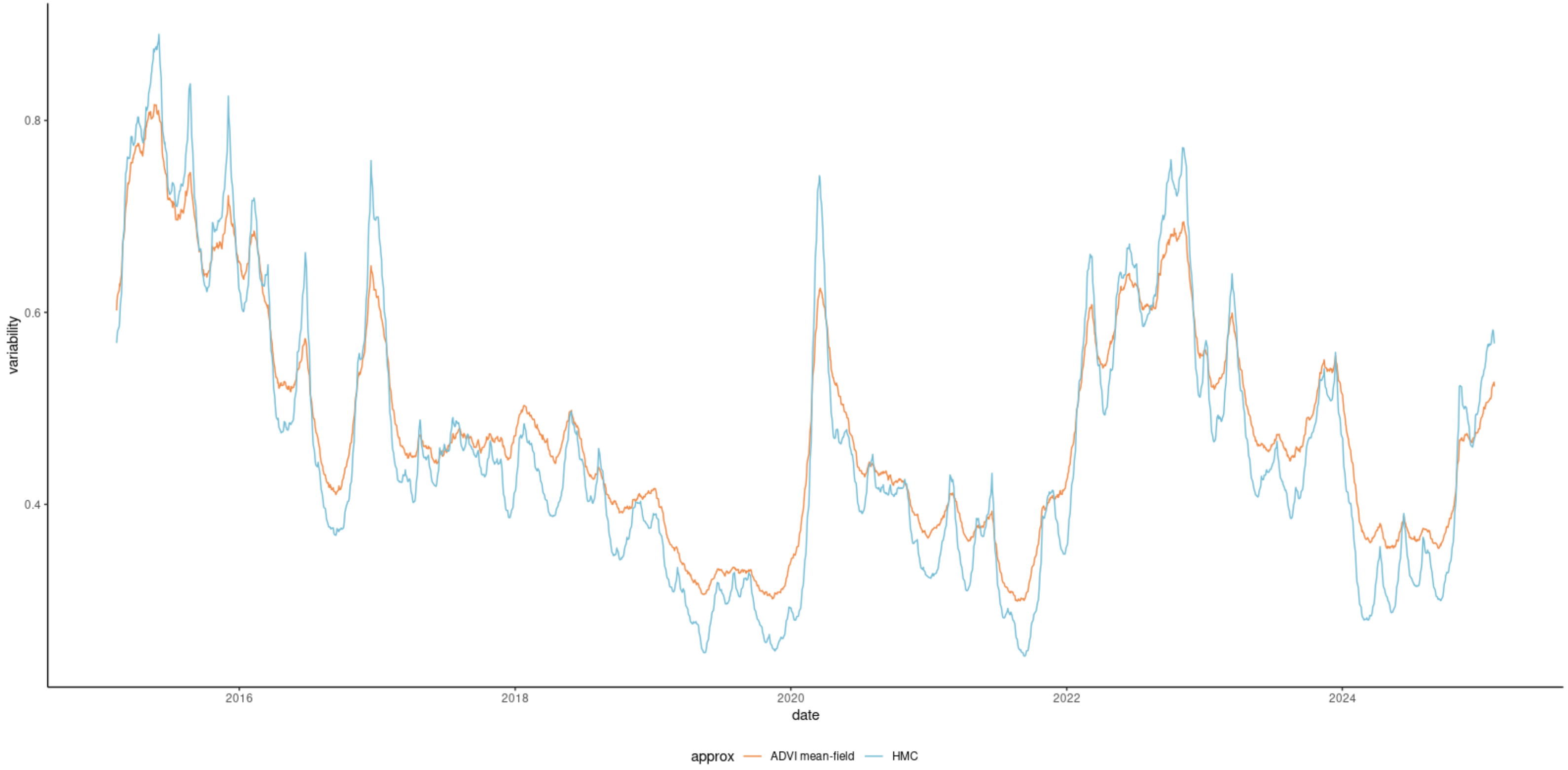
The gradients of the ELBO with respect to the mean and variance are

$$\begin{aligned}\nabla_{\boldsymbol{\mu}} &= \mathsf{E}_{\boldsymbol{Z}}\left\{\frac{\partial \log p(\boldsymbol{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial T^{-1}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} + \frac{\partial \log |\mathbf{J}_{T^{-1}}(\boldsymbol{\zeta})|}{\partial \boldsymbol{\zeta}}\right\}\\ \nabla_{\mathbf{L}} &= \mathsf{E}_{\boldsymbol{Z}}\left[\left\{\frac{\partial \log p(\boldsymbol{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial T^{-1}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} + \frac{\partial \log |\mathbf{J}_{T^{-1}}(\boldsymbol{\zeta})|}{\partial \boldsymbol{\zeta}}\right\}\boldsymbol{Z}^\top\right] + \mathbf{L}^{-\top}.\end{aligned}$$

and we can approximate the expectation by drawing standard Gaussian samples $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_B$.

Quality of approximation

Consider the stochastic volatility model.



Fitting HMC-NUTS to the exchange rate data takes 156 seconds for 10K iterations, vs 2 seconds for the mean-field approximation.

References

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(40), 1303–1347. http://jmlr.org/papers/v14/hoffman13a.html

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*(14), 1–45. http://jmlr.org/papers/v18/16-107.html

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*(2), 140–153. https://doi.org/10.1198/tast.2010.09058

Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. In S. Kaski & J. Corander (Eds.), *Proceedings of the seventeenth international conference on artificial intelligence and statistics* (Vol. 33, pp. 814–822). Pmlr. https://proceedings.mlr.press/v33/ranganath14.html