Automatic variational inference

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Introduction

Black box variational inference (BBVI)

Demonstration of BBVI

Refinements of BBVI
- Rao–Blackwellization
- Control variates

Automatic differentiation variational inference (ADVI)

Outline

Introduction

Black box variational inference (BBVI)

Demonstration of BBVI

Refinements of BBVI

- Rao–Blackwellization
- Control variates

Automatic differentiation variational inference (ADVI)

Introduction

- The classic VI updates for a given model require a fair amount of math to derive, which is difficult for nonexperts.
- Automatic VI techniques are much easier to use, since minimal derivations are required.
- Classic VI also has limited flexibility, since:
  - factorized approximations have limited accuracy, and
  - the target distribution has to be conducive to classic VI.
- Automatic VI enables more flexible approximations that can be more accurate and more generally applicable.
- We'll consider two automatic VI techniques:
  - Black box VI (Ranganath et al., 2014)
  - Automatic differentiation VI (Kucukelbir et al., 2015)

Introduction: Stochastic optimization

- Most automatic VI techniques employ stochastic optimization.

- Suppose we want to minimize a differentiable function $f(a)$.

- Initialize a. Iterate for $t = 1, 2, \ldots$ until convergence:
  1. Let $G_t(a)$ be a random function such that $\mathrm{E} G_t(a) = \nabla f(a)$.
  2. Update $a \leftarrow a - \rho_t G_t(a)$.

- This can be viewed as a gradient descent algorithm with noisy approximations to the gradient.

- This converges to a local minimum of f if the step size $\rho_t > 0$ satisfies the Robbins–Munro conditions:
  1. $\sum_{t=1}^{\infty} \rho_t = \infty$ and
  2. $\sum_{t=1}^{\infty} \rho_t^2 < \infty$.
- The step size ρ_t is also referred to as the "learning rate".

Outline

Introduction

Black box variational inference (BBVI)

Demonstration of BBVI

Refinements of BBVI

- Rao–Blackwellization
- Control variates

Automatic differentiation variational inference (ADVI)

Black box variational inference (BBVI)

- Recall that in classic VI, we seek a $q \in \mathcal{Q}$ that minimizes the KL divergence from the target distribution $\pi(\theta)$:

$$q^{\text{opt}} \in \underset{q \in \mathcal{Q}}{\operatorname{argmin}} D(q \| \pi).$$

- Recall that $D(q\|\pi) = \displaystyle\int q(\theta) \log \frac{q(\theta)}{\pi(\theta)} d\theta$.

- The basic idea of black box VI (Ranganath et al., 2014) is to employ stochastic optimization to minimize $D(q\|\pi)$.

- A parametric family $\mathcal{Q} = \{q_a(\theta) : a \in \mathcal{A}\}$ is used to facilitate optimization.

- BBVI uses a certain expression for the gradient (wrt a) of the KL divergence $D(q_a\|\pi)$, which we informally derive now.

Black box variational inference

- First, observe that

$$\nabla_a \log q_a = (\nabla_a q_a)/q_a, \text{ and thus, } \nabla_a q_a = q_a(\nabla_a \log q_a).$$

- Hence, by the product rule,

$$\nabla_a\Big(q_a \log \frac{q_a}{\pi}\Big) = q_a(\nabla_a \log q_a)\Big(\log \frac{q_a}{\pi}\Big) + \nabla_a q_a.$$

- Note that $\displaystyle\int \nabla_a q_a = \nabla_a \int q_a = \nabla_a 1 = 0.$

- Hence, the gradient (wrt a) of the KL divergence is:

$$\begin{aligned}\nabla_a D(q_a \| \pi) &= \int \nabla_a\Big(q_a \log \frac{q_a}{\pi}\Big)\\ &= \int q_a(\nabla_a \log q_a)\Big(\log \frac{q_a}{\pi}\Big) + \int \nabla_a q_a\\ &= \mathrm{E}_{q_a}(\nabla_a \log q_a)(\log q_a - \log \pi).\end{aligned}$$

Black box VI: Naive algorithm

- Thus, the gradient of the KL divergence can be written as

$$\nabla_a D(q_a \| \pi) = \mathrm{E}_{q_a}\Big(\big(\nabla_a \log q_a(\theta)\big)\big(\log q_a(\theta) - \log \pi(\theta)\big)\Big).$$

- In its simplest form, BBVI uses a Monte Carlo approximation to this gradient: let $\theta_1, \ldots, \theta_N \sim q_a$ and define

$$G_t(a) := \frac{1}{N} \sum_{i=1}^{N} \big(\nabla_a \log q_a(\theta_i)\big)\big(\log q_a(\theta_i) - \log \pi(\theta_i)\big).$$

- Naive version of black box VI algorithm:
  - Initialize a.
  - Iterate for $t = 1, 2, \ldots$ until convergence:
    1. Sample $\theta_1, \ldots, \theta_N \sim q_a$ i.i.d.
    2. Update $a \leftarrow a - \rho_t G_t(a)$ where $G_t(a)$ is defined as above.
- Since a changes at each step, it is necessary to redraw the samples θ_i at each step.

BBVI: Normalization constants and the ELBO

- It turns out that we only need to be able to compute π up to a normalization constant.
- The reason is that if $\pi = \tilde{\pi}/z$ then

$$D(q\|\pi) = D(q\|\tilde{\pi}/z) = \log(z) + \int q \log(q/\tilde{\pi}).$$

- $\log(z)$ is a constant that doesn't depend on q.
- BBVI can be applied to minimize $\int q \log(q/\tilde{\pi})$ since it doesn't require the target to integrate to 1.

BBVI: Normalization constants and the ELBO

- Incidentally, this relates to the so-called "evidence lower bound", or ELBO for short.

- Rearranging the equation above and using $D(q\|\pi) \geq 0$,

$$\log(z) = D(q\|\pi) - \int q \log(q/\tilde{\pi}) \geq - \int q \log(q/\tilde{\pi}).$$

- The RHS is called the *evidence lower bound* (ELBO).

- Since $\log(z)$ doesn't depend on q, minimizing $D(q\|\pi)$ with respect to q is equivalent to maximizing the ELBO.

- When π is the posterior of Bayesian model,

$$\pi(\theta) = p(\theta|x) = p(x, \theta)/p(x),$$

  we can choose $\tilde{\pi}(\theta) = p(x, \theta)$, and in this case, $z = p(x)$ is the marginal likelihood, a.k.a., the model evidence.

Black box VI: Advantages

- Advantages
  - Allows flexible choice of approximating family $\{q_a(\theta)\}$.
  - Allows wider range of target distributions π, compared to classic VI.
  - Does not require any model-specific derivations (that is, derivations specific to π).
  - Only need to be able to compute π up to a normalization constant.
  - Deriving and computing $\nabla_a \log q_a(\theta)$ is usually easy.
  - Convergence guarantee from stochastic optimization theory.
  - Uses logs everywhere, which is good for numerical stability.

Black box VI: Disadvantages

- Disadvantages
  - Main issue: The Monte Carlo approximation is often too noisy, making the stochastic optimization rather inefficient.
  - Finding a good sequence of step sizes ρ_t can be tricky.
  - Gradient descent can be somewhat slow, and stochastic gradient requires even more iterations to converge.
- When the variance of the Monte Carlo approximation is higher, the step sizes ρ_t need to be smaller, slowing the rate of convergence.

Outline

Introduction

Black box variational inference (BBVI)

Demonstration of BBVI

Refinements of BBVI
- Rao–Blackwellization
- Control variates

Automatic differentiation variational inference (ADVI)

Black box VI: Demo on kidney disease data

- Before discussing how to resolve this issue, we look at BBVI on an example.

- Longitudinal data from 976 patients (803 train, 173 test) with chronic kidney disease.

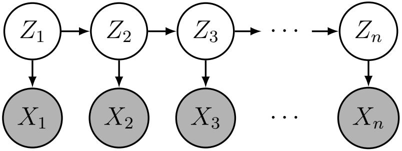
- During each visit, some combination of 17 measurements were taken.

- Additionally, the times between patient visits were highly irregular.

- The goal is to come up with a low-dimensional summarization of patients' data.

Kidney disease demo: Model

- A hidden Markov model with continuous latent states and factor analysis-like emissions is used for each patient.



- For each patient $p = 1, \ldots, P$, for each visit $v = 1, \ldots, V_p$, the model is as follows:
  - $Z_{p,v,k} | Z_{p,v-1,k}$ is Gamma distributed with mean $Z_{p,v-1,k}$ and variance σ_z^2 for each latent dimension $k = 1, \ldots, K$.
  - $X_{p,v} = W Z_{p,v} + \eta_p + \varepsilon_{p,v}$ where $X_{p,v} \in \mathbb{R}^M$ and $W \in \mathbb{R}^{M \times K}$.

- Normal priors are placed on the entries of W, η_p, and $\varepsilon_{p,v}$.

Kidney disease demo: Comments

- This is a rather complex model, making it an interesting example to consider.
- Neither classic VI nor Gibbs sampling can be used for this model, since the required updates are not closed form.
- However, BBVI is relatively straightforward to apply.
- For comparison, a Metropolis–Hastings-within-Gibbs algorithm was also run.
- To assess performance, the posterior predictive log-likelihood on the test set was computed as a function of computation time.

Kidney disease demo: Empirical results

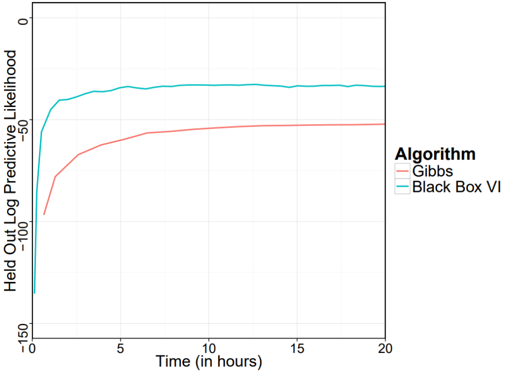


Figure 1: Comparison between Metropolis-Hastings within Gibbs and Black Box Variational Inference. In the x axis is time and in the y axis is the predictive likelihood of the test set. Black Box Variational Inference reaches better predictive likelihoods faster than Gibbs sampling. The Gibbs sampler's progress slows considerably after 5 hours.

(figure and caption from Ranganath et al., 2014)

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/4vZ3vASAjxCSt32AA

(Three people per room, randomly assigned. 10 minutes.)

Kidney disease demo: Empirical results

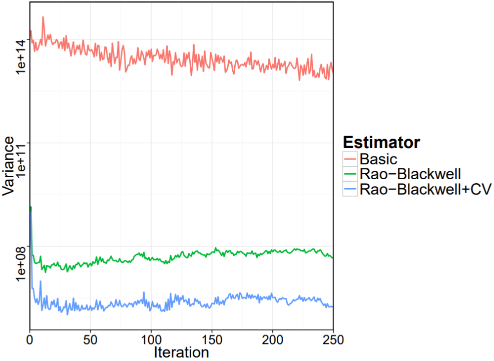


Figure 2: Variance comparison for the first component of a random patient on the following estimators: Eq. 3, the Rao-Blackwellized estimator Eq. 6, and the Rao-Blackwellized control variate estimator Eq. 10. We find that Rao-Blackwellizing the naive estimator reduces the variance by several orders of magnitude from the naive estimator. Adding control variates reduces the variance even further.

(figure and caption from Ranganath et al., 2014)

Outline

Introduction

Black box variational inference (BBVI)

Demonstration of BBVI

Refinements of BBVI

- Rao–Blackwellization
- Control variates

Automatic differentiation variational inference (ADVI)

Refinements of BBVI

- The naive BBVI algorithm works in principle, but tends to be too slow in practice due to its high Monte Carlo variance.

- Convergence speed can be improved by reducing the variance.

- Two generally applicable variance-reduction techniques are:
  1. Rao–Blackwellization, and
  2. control variates.

- Fortunately, when applied to BBVI, these variance reduction methods do not require model-specific computations.

- Other refinements include:
  1. adaptive choice of step size ρ_t, and
  2. stochastic VI based on random subsets of the data.

Rao–Blackwellization

- Rao–Blackwellization is a general technique for reducing the variance of an estimator.

- Suppose X is an unbiased estimator of some quantity of interest, and Y is some other random variable.
- Then $\mathrm{E}(X|Y)$ is also an unbiased estimator, and its variance is less or equal to that of X.

- This is because by the law of total expectation,

  $$\mathrm{E}(\mathrm{E}(X|Y)) = \mathrm{E}(X)$$

  and by the law of total variance,

  $$\mathrm{Var}(X) = \mathrm{Var}(\mathrm{E}(X|Y)) + \mathrm{E}(\mathrm{Var}(X|Y)) \geq \mathrm{Var}(\mathrm{E}(X|Y)).$$

- In BBVI, we consider a case in which $X = h(\theta_1, \theta_2)$ and $\theta_1 \perp\!\!\!\perp \theta_2$, so $\mathrm{E}(h(\theta_1, \theta_2)|\theta_1) = \int h(\theta_1, \theta_2)p(\theta_2)d\theta_2$.

Rao–Blackwellization for Black Box VI (1/3)

- Rao–Blackwellization for BBVI relies on a factorization of q similar to classic VI, however, the form of this factorization is somewhat less restrictive than classic VI.

- Suppose the approximating distributions factor as

$$q(\theta|a) = q_1(\theta_1|a_1) \cdots q_m(\theta_m|a_m)$$

  for some decomposition $\theta = (\theta_1, \ldots, \theta_m)$.

- Incidentally, in VI, this type of factorization is referred to as a "mean field" or "structured mean field" approximation.

Rao–Blackwellization for Black Box VI (2/3)

- Let θ_{-j} denote all the components of θ except θ_j.

- We factor π into parts that do and do not depend on θ_j.

- Specifically, factor π as

$$\pi(\theta) = \pi_{S_j}(\theta_{S_j})\pi_{-j}(\theta_{-j}),$$

  where $S_j \subseteq \{1, \ldots, m\}$ and $\pi_{-j}(\theta_{-j})$ does not depend on θ_j.

- Then $\log q - \log \pi = \log q_j - \log \pi_{S_j} + (\text{const wrt } \theta_j)$.

- Since $\mathrm{E}_{q_j}(\nabla_{a_j} \log q_j) = 0$, we have

$$\begin{aligned} &\mathrm{E}_q\big((\nabla_{a_j} \log q_j)(\text{const wrt } \theta_j)\big) \\ &= \mathrm{E}_{q_j}(\nabla_{a_j} \log q_j)\, \mathrm{E}_{q_{-j}}(\text{const wrt } \theta_j) = 0. \end{aligned}$$

Rao–Blackwellization for Black Box VI (3/3)

- Consider the gradient wrt a_j (instead of wrt all of a).
- Like before, the gradient of the KL divergence wrt a_j is

$$\begin{aligned}\nabla_{a_j} D(q_a\|\pi) &= \mathrm{E}_q\big((\nabla_{a_j}\log q)(\log q - \log\pi)\big)\\ &= \mathrm{E}_q\Big((\nabla_{a_j}\log q_j)(\log q_j - \log\pi_{S_j} + (\text{const wrt } \theta_j))\Big)\\ &= \mathrm{E}_{q_{S_j}}\big((\nabla_{a_j}\log q_j)(\log q_j - \log\pi_{S_j})\big) + 0.\end{aligned}$$

- Therefore, we can use a Monte Carlo approximation

$$\nabla_{a_j} D(q_a\|\pi) \approx \frac{1}{N}\sum_{i=1}^{N}\big(\nabla_{a_j}\log q_j(\theta_{j,i})\big)\big(\log q_j(\theta_{j,i}) - \log\pi_{S_j}(\theta_{S_j,i})\big)$$

  where $\theta_{S_j,1},\ldots,\theta_{S_j,N}$ are i.i.d. from q_{S_j} under the current value of a.

- This is a Rao–Blackwellized estimate since all the components of θ outside of S_j are implicitly integrated out.

Control variates

- The method of control variates is another general purpose technique for reducing the variance of an estimator.

- Suppose X is an unbiased estimator of some quantity of interest.

- Let Y be any random variable with finite variance, and define a new estimator

$$\tilde{X} = X - c\,(Y - \mathrm{E}(Y)).$$

- Then $\tilde{X}$ is also unbiased, and the variance of $\tilde{X}$ is

$$\mathrm{Var}(\tilde{X}) = \mathrm{Var}(X) + c^2\mathrm{Var}(Y) - 2c\,\mathrm{Cov}(X, Y)$$

  by straightforward calculations.

- Setting the derivative wrt c to 0 and solving, we find that $\mathrm{Var}(\tilde{X})$ is minimized when $c = \mathrm{Cov}(X, Y)/\mathrm{Var}(Y)$.

Control variates

- Plugging in the optimal value of c, we get

$$\mathrm{Var}(\tilde{X}) = \mathrm{Var}(X) - \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(Y)}.$$

- Thus, when using the optimal c, the variance is reduced whenever $\mathrm{Cov}(X,Y) \neq 0$.

- In practice, we often need to estimate the optimal c using samples.

- Usually, one chooses Y such that $\mathrm{E}(Y)$ is known or easy to compute.

Control variates for Black Box VI

- Recall that Rao–Blackwellized BBVI uses a Monte Carlo approximation of $\nabla_{a_j} D(q_a \| \pi)$.

- To reduce the variance of this Monte Carlo approximation, the BBVI paper proposes to use $Y = \nabla_{a_j} \log q_j(\theta_j | a_j)$ as a control variate, where $\theta_j \sim q_j(\cdot | a_j)$.

- This is convenient since then $\mathrm{E}(Y) = 0$ and the optimal c can easily be estimated from the samples used to construct the Monte Carlo approximation itself.

- This leads to a slightly modified update step in the BBVI stochastic optimization algorithm.

- In addition to Rao–Blackwellization, control variates can significantly improve performance.

Kidney disease demo: Empirical results

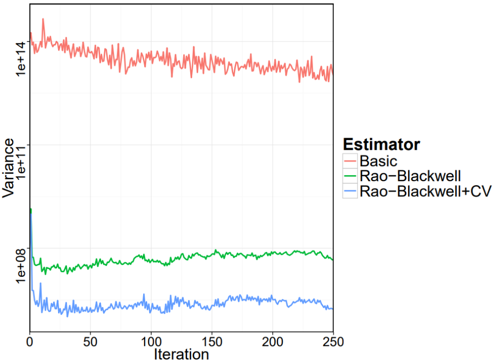


Figure 2: Variance comparison for the first component of a random patient on the following estimators: Eq. 3, the Rao-Blackwellized estimator Eq. 6, and the Rao-Blackwellized control variate estimator Eq. 10. We find that Rao-Blackwellizing the naive estimator reduces the variance by several orders of magnitude from the naive estimator. Adding control variates reduces the variance even further.

(figure and caption from Ranganath et al., 2014)

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/3b6mhpE2Fc1WyD5B9

(Three people per room, randomly assigned. 10 minutes.)

Outline

Introduction

Black box variational inference (BBVI)

Demonstration of BBVI

Refinements of BBVI

- Rao–Blackwellization
- Control variates

Automatic differentiation variational inference (ADVI)

Automatic differentiation variational inference (ADVI)

- ADVI (Kucukelbir et al., 2015) is a variant of BBVI that can be used within the Stan language.

- ADVI employs automatic differentiation to compute gradients, so that no mathematical derivations are required on the part of the user.

- In ADVI, the approximating family $\{q_a\}$ consists of multivariate Gaussians with diagonal covariance.

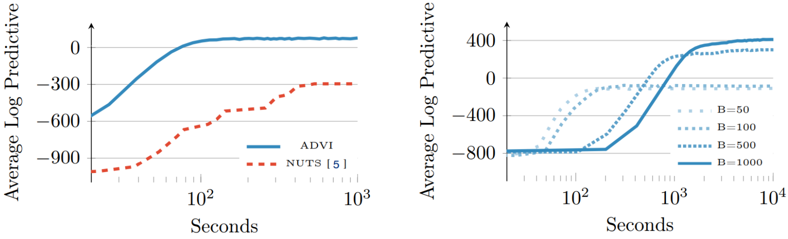
Automatic differentiation variational inference (ADVI)

- In ADVI, all parameters must be continuous, and any parameters with restricted domain are mapped to $\mathbb{R}$.

- While limited in terms of approximation accuracy, Gaussians are analytically convenient.

- In ADVI, the variance of the stochastic gradient steps is reduced since the $\int q \log q$ term in the KL divergence can be computed analytically when q is a Gaussian.

ADVI demo: Gaussian mixture model on image dataset



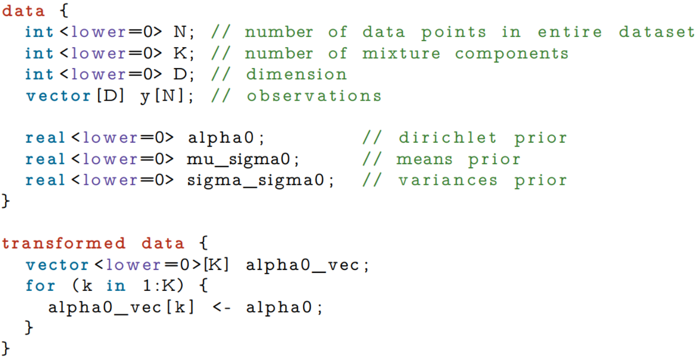
(a) Subset of 1000 images

(b) Full dataset of 250 000 images

Figure 1: Held-out predictive accuracy results | GMM of the imageCLEF image histogram dataset. **(a)** ADVI outperforms the NUTS, the default sampling method in Stan [5]. **(b)** ADVI scales to large datasets by subsampling minibatches of size B from the dataset at each iteration [3]. We present more details in Section 3.3 and Appendix J.

(figure and caption from Kucukelbir et al., 2015)

ADVI demo: Stan code for GMM on image dataset (1/2)



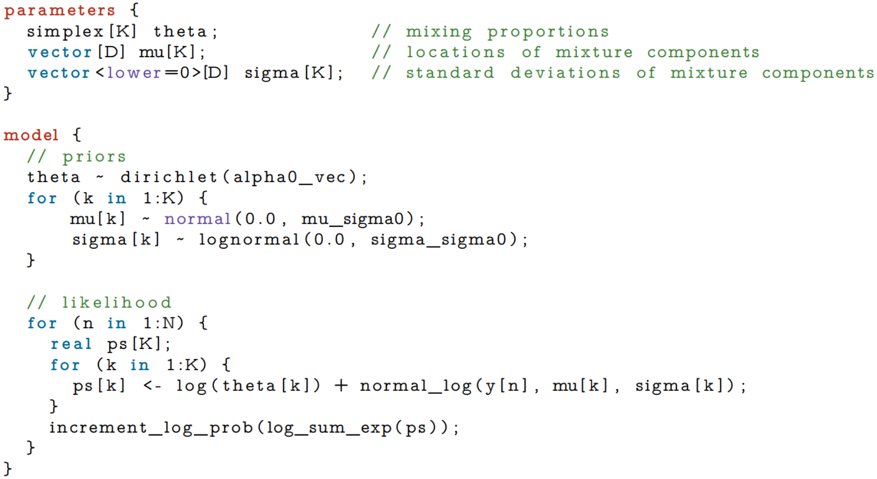
```
data {
  int<lower=0> N; // number of data points in entire dataset
  int<lower=0> K; // number of mixture components
  int<lower=0> D; // dimension
  vector[D] y[N]; // observations

  real<lower=0> alpha0;          // dirichlet prior
  real<lower=0> mu_sigma0;       // means prior
  real<lower=0> sigma_sigma0;    // variances prior
}

transformed data {
  vector<lower=0>[K] alpha0_vec;
  for (k in 1:K) {
    alpha0_vec[k] <- alpha0;
  }
}
```

(figure from Kucukelbir et al., 2015)

ADVI demo: Stan code for GMM on image dataset (2/2)



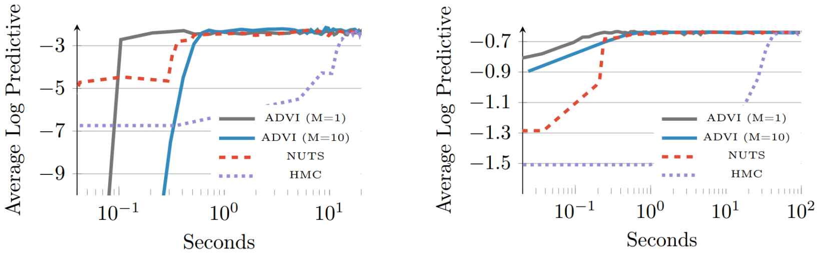
```
parameters {
  simplex[K] theta;              // mixing proportions
  vector[D] mu[K];               // locations of mixture components
  vector<lower=0>[D] sigma[K];   // standard deviations of mixture components
}

model {
  // priors
  theta ~ dirichlet(alpha0_vec);
  for (k in 1:K) {
      mu[k] ~ normal(0.0, mu_sigma0);
      sigma[k] ~ lognormal(0.0, sigma_sigma0);
  }

  // likelihood
  for (n in 1:N) {
    real ps[K];
    for (k in 1:K) {
      ps[k] <- log(theta[k]) + normal_log(y[n], mu[k], sigma[k]);
    }
    increment_log_prob(log_sum_exp(ps));
  }
}
```

(figure from Kucukelbir et al., 2015)

ADVI demo: Hierarchical GLMs

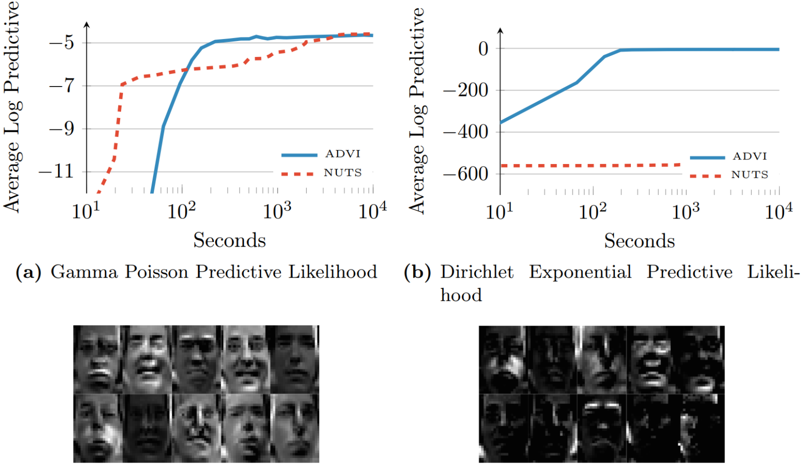


(a) Linear Regression with ARD

(b) Hierarchical Logistic Regression

(figure from Kucukelbir et al., 2015)

ADVI demo: Hierarchical GLMs



(a) Gamma Poisson Predictive Likelihood

(b) Dirichlet Exponential Predictive Likelihood

(c) Gamma Poisson Factors

(d) Dirichlet Exponential Factors

Figure 5: Non-negative matrix factorization of the Frey Faces dataset.

(figure from Kucukelbir et al., 2015)

References and supplements

- Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. In Artificial Intelligence and Statistics (pp. 814-822).
- Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. (2015). Automatic variational inference in Stan. In Advances in Neural Information Processing Systems (pp. 568-576).