Black box variational inference methods

2014, 2015, 2015+

Dustin Tran
Department of Statistics
Harvard University

Joint work with:
Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, David M. Blei, Edoardo M. Airoldi

A few works

1. Black box variational inference (BBVI) [RGB14]
2. Automatic differentiation variational inference (ADVI) [Kuc+15]
3. Variational inference with copulas [TBA15]
4. Variational Gaussian process (VARIATIONAL GP)

Extensions/relations to:

1. variational autoencoder
2. probabilistic programming
3. reinforcement learning

Setup

- Given: Dataset $\mathbf{x}$; Likelihood $p(\mathbf{x} \mid \mathbf{z})$; Prior $p(\mathbf{z})$
- Goal: Compute posterior $p(\mathbf{z} \mid \mathbf{x})$

Setup

- Given: Dataset $\mathbf{x}$; Likelihood $p(\mathbf{x} \mid \mathbf{z})$; Prior $p(\mathbf{z})$
- Goal: Compute posterior $p(\mathbf{z} \mid \mathbf{x})$
- Variational inference: approximate the posterior with variational distribution $q(\mathbf{z}\,;\,\lambda)$ and solve

$$\min_{\lambda \in \Lambda} \mathrm{KL}(q \parallel p) = \min_{\lambda \in \Lambda} \int q(\mathbf{z}\,;\,\lambda) \frac{q(\mathbf{z}\,;\,\lambda)}{p(\mathbf{z} \mid \mathbf{x})} \, \mathrm{d}\mathbf{z}$$

Equivalent to maximizing the Evidence Lower Bound (ELBO)

$$\lambda^* = \arg\max_{\lambda \in \Lambda} \mathcal{L}(\lambda) \quad \text{s.t.} \quad \operatorname{supp}(q(\mathbf{z}\,;\,\lambda)) \subseteq \operatorname{supp}(p(\mathbf{z} \mid \mathbf{x}))$$

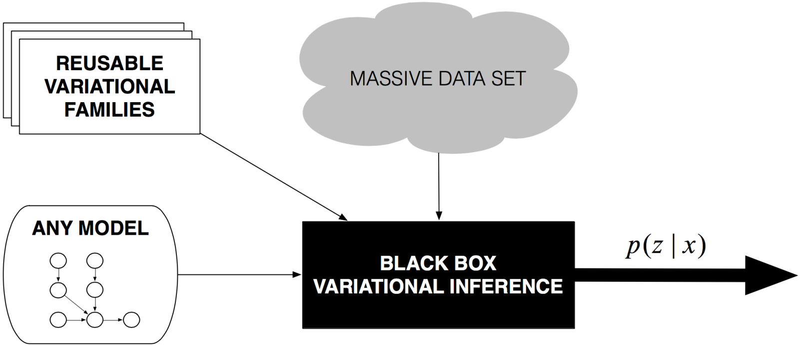
$$\mathcal{L}(\lambda) = \mathbb{E}_{q(\mathbf{z})}\big[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}\,;\,\lambda)\big]$$

Use the *REINFORCE* gradient

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{q(\mathbf{z})}\big[\nabla \log q(\mathbf{z}\,;\,\lambda)(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}\,;\,\lambda))\big]$$

Algorithm 1 Black Box Variational Inference

Input: data x, joint distribution p, mean field variational family q.
Initialize λ randomly, $t = 1$.
repeat
  // **Draw** S **samples from** q
  for $s = 1$ **to** S **do**
    $z[s] \sim q$
  end for
  $\rho = t$th value of a Robbins Monro sequence
  $\lambda = \lambda + \rho \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z[s] \mid \lambda)(\log p(x, z[s]) - \log q(z[s] \mid \lambda))$
  $t = t + 1$
until change of λ is less than 0.01.


REUSABLE VARIATIONAL FAMILIES
MASSIVE DATA SET
ANY MODEL
BLACK BOX VARIATIONAL INFERENCE
$p(z \mid x)$

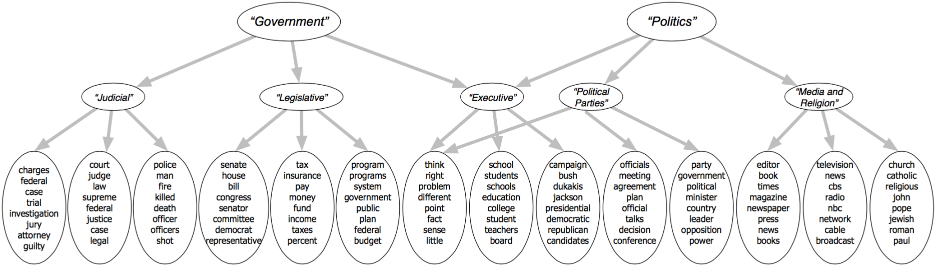


Figure 1: A fraction of the three layer topic hierarchy on 166K *The New York Times* articles. The top words are shown for each topic. The arrows represent hierarchical groupings.

New challenges

1. *Expressivity*: Specification of q
2. *Scalability*: Variance of stochastic approximation
3. *Generality*: Infinite-dimensional, continuous-discrete variable models[1]
4. *Robustness*: Initialization; numerical stability
5. *Diagnostics*: Convergence; nonconvex optima

[1] probabilistic programs most generally

Biggest challenges

- *Theory (asymptotics)*: consistency, efficiency, and asymptotic normality of variational MAP estimate
- *Bayesian analysis*: hypothesis testing; model criticism; causal inference

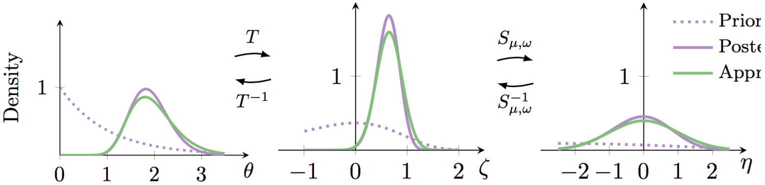
Automatic differentiation variational inference (2015)

- Implicit Gaussian approximation
- Parameter transformations
- Automatic differentiation
- Massive software engineering task, now available in **Stan**[2]

[2] http://mc-stan.org

New challenges

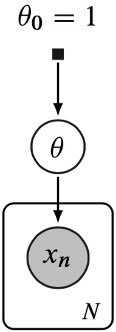
1. *Expressivity*: Specification of q [ADVI]
2. *Scalability*: Variance of stochastic approximation [ADVI]
3. *Generality*: Infinite-dimensional, continuous-discrete variable models
4. *Robustness*: Initialization; numerical stability [ADVI]
5. *Diagnostics*: Convergence; nonconvex optima [ADVI]



(a) Latent variable space

(b) Real coordinate space

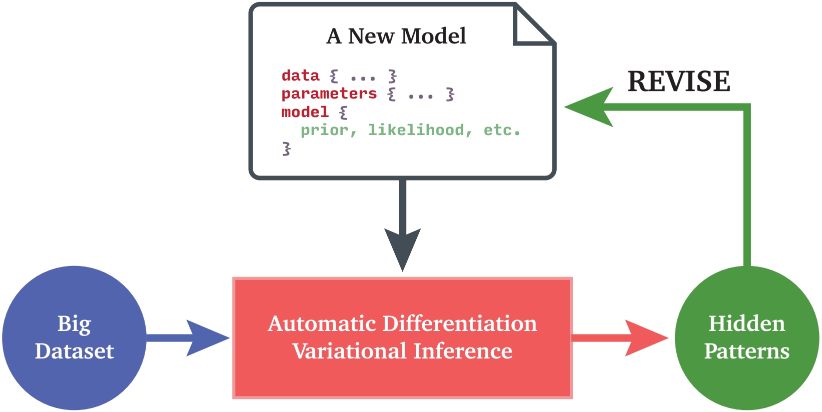
(c) Standardized space

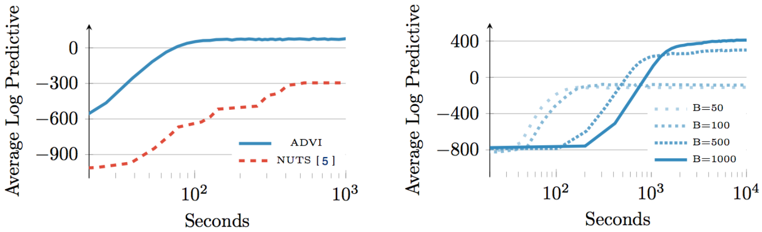



```
data {
  int N;     // number of observations
  int x[N]; // discrete-valued observations
}
parameters {
  // latent variable, must be positive
  real<lower=0> theta;
}
model {
  // non-conjugate prior for latent variable
  theta ~ exponential(1.0);

  // likelihood
  for (n in 1:N)
    increment_log_prob(poisson_log(x[n], theta));
}
```

Figure 1: Specifying a simple nonconjugate probability model in Stan.


A New Model
data { ... }
parameters { ... }
model {
prior, likelihood, etc.
}
REVISE
Big Dataset
Automatic Differentiation Variational Inference
Hidden Patterns



(a) Subset of 1000 images

(b) Full dataset of 250 000 images

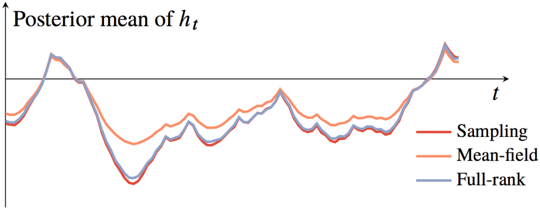
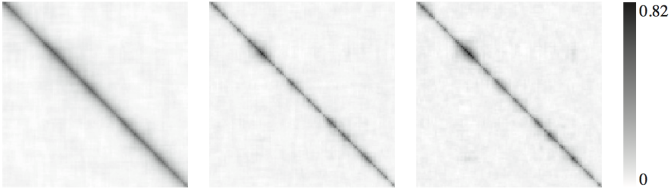


Figure 6: Comparison of posterior mean estimates of volatility.



(a) Mean-field **(b)** Full-rank **(c)** Sampling

Figure 7: Comparison of empirical posterior covariance matrices.

Variational inference with copulas (2015)

New challenges

1. *Expressivity*: Specification of q [copulas]
2. *Scalability*: Variance of stochastic approximation
3. *Generality*: Infinite-dimensional, continuous-discrete variable models
4. *Robustness*: Initialization; numerical stability
5. *Diagnostics*: Convergence; nonconvex optima [copulas]

Variational inference with copulas (2015)

Goal: Compute posterior $p(\mathbf{z} \mid \mathbf{x})$

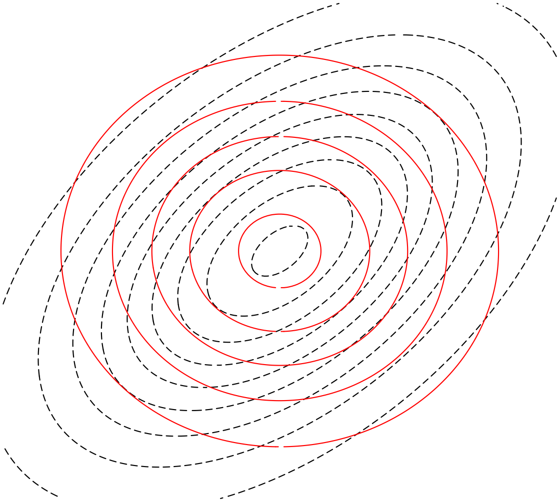
$$q(z \mid \lambda) = \underbrace{\prod_{i=1}^{d} q(z_i \mid \lambda_i)}_{\text{mean-field}}$$

Variational inference with copulas (2015)

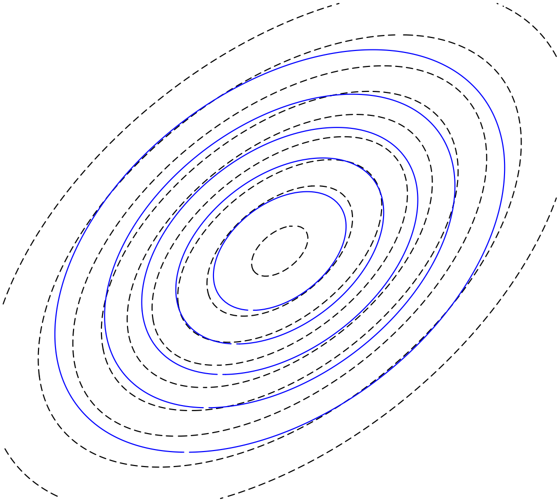
Goal: Compute posterior $p(\mathbf{z} \mid \mathbf{x})$

$$q(z \mid \lambda, \eta) = \underbrace{\left[\prod_{i=1}^{d} q(z_i \mid \lambda_i)\right]}_{\text{mean-field}} \underbrace{c(Q(z_1 \mid \lambda_1), \ldots, Q(z_d \mid \lambda_d); \eta)}_{\text{copula}}$$

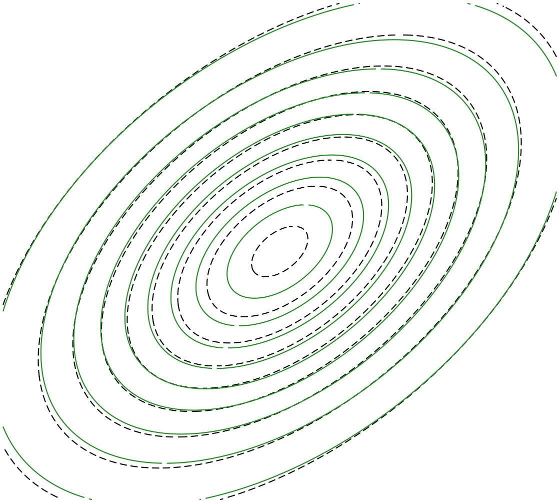
Iteration 1: Fit $\lambda \mid \eta = \text{Unif}$



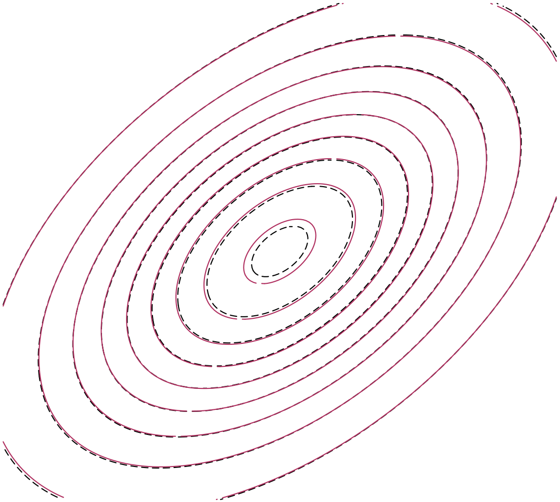
Iteration 2: Fit $\eta \mid \lambda$



Iteration 3: Fit $\lambda \mid \eta$



Iteration 4: Fit $\eta \mid \lambda$



$$q(z \mid \lambda, \eta) = \underbrace{\left[\prod_{i=1}^{d} q(z_i \mid \lambda_i)\right]}_{\text{mean-field}} \underbrace{c(Q(z_1 \mid \lambda_1), \ldots, Q(z_d \mid \lambda_d); \eta)}_{\text{copula}}$$

How to specify the copula?

$$q(z \mid \lambda, \eta) = \underbrace{\left[\prod_{i=1}^{d} q(z_i \mid \lambda_i)\right]}_{\text{mean-field}} \underbrace{c(Q(z_1 \mid \lambda_1), \ldots, Q(z_d \mid \lambda_d); \eta)}_{\text{copula}}$$

How to specify the copula?

1. Factorize c into a product of two-dimensional conditional copulas (*vine*)

$$\begin{aligned} c(u_1, u_2, u_3, u_4, u_5) = {} & c(u_1, u_5)c(u_1, u_3)c(u_2, u_3)c(u_3, u_4) \cdot \\ & c(u_1, u_2 | u_3)c(u_3, u_5 | u_1)c(u_1, u_4 | u_3) \cdot \\ & c(u_2, u_4 | u_1, u_3)c(u_2, u_5 | u_1, u_3) \cdot \\ & c(u_4, u_5 | u_1, u_2, u_3) \end{aligned}$$

$$q(z \mid \lambda, \eta) = \underbrace{\left[\prod_{i=1}^{d} q(z_i \mid \lambda_i)\right]}_{\text{mean-field}} \underbrace{c(Q(z_1 \mid \lambda_1), \ldots, Q(z_d \mid \lambda_d); \eta)}_{\text{copula}}$$

How to specify the copula?

1. Factorize c into a product of two-dimensional conditional copulas (*vine*)

$$\begin{aligned} c(u_1, u_2, u_3, u_4, u_5) = {} & c(u_1, u_5)c(u_1, u_3)c(u_2, u_3)c(u_3, u_4) \cdot \\ & c(u_1, u_2|u_3)c(u_3, u_5|u_1)c(u_1, u_4|u_3) \cdot \\ & c(u_2, u_4|u_1, u_3)c(u_2, u_5|u_1, u_3) \cdot \\ & c(u_4, u_5|u_1, u_2, u_3) \end{aligned}$$

2. Perform model selection to choose a parameteric form, each of which are equipped with parameters η_i

Variational parameters: (λ, η)

Algorithm 1 Copula variational inference (CVI)

Input: data x, joint distribution p, variational family q
Initialize λ randomly, η so that c is uniform
repeat
  // Fix η, maximize over λ
  Initialize $t = 1$
  repeat
    $z \sim q(z \mid \lambda, \eta)$
    $\lambda = \lambda + \rho_t(\nabla_\lambda \log q(z \mid \lambda, \eta) \cdot (\log p(x, z) - \log q(z \mid \lambda, \eta)))$
    $t = t + 1$
  until change of λ is less than a small ϵ_1
  // Fix λ, maximize over η
  Initialize $t = 1$
  repeat
    $z \sim q(z \mid \lambda, \eta)$
    $\eta = \eta + \rho_t(\nabla_\eta \log q(z \mid \lambda, \eta) \cdot (\log p(x, z) - \log q(z \mid \lambda, \eta)))$
    $t = t + 1$
  until change of η is less than a small ϵ_2
until change of λ and η is less than a small ϵ_1 and ϵ_2 respectively

Variational inference methods	Predictive Likelihood
Mean-field	-383.2
LRVB	-330.5
CVI (2 iterations)	-303.2
CVI (5 iterations)	-80.2
CVI (converged)	-50.5

Table 1: A comparison of mean-field, LRVB, and CVI on the latent space model, where each CVI iteration either refits the mean-field or the copula. CVI converges in roughly 10 iterations and already significantly outperforms both mean-field and LRVB upon fitting the copula once (two iterations).

Alp Kucukelbir et al. “Automatic Variational Inference in Stan”. In: *Neural Information Processing Systems (forthcoming)*. 2015.

Rajesh Ranganath, Sean Gerrish, and David M Blei. “Black Box Variational Inference”. In: *Artificial Intelligence and Statistics*. 2014.

Dustin Tran, David M. Blei, and Edoardo M. Airoldi. “Variational inference with copula augmentation”. In: *Neural Information Processing Systems (forthcoming)*. 2015.

More details at dustintran.com