ML for Econometricians
Variational Inference

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- **Variational Inference**
- **Mean-field variational inference**
- **Stochastic Variational Inference**
- **Black-Box Variational Inference**

Variational Inference (VI)

- Let $\theta = (\theta_1, ..., \theta_p)^T$.
- Aim: **approximate posterior** $p(\theta|y)$ with a simpler distribution $q(\theta)$.
- **Normal approximation**: $q(\theta) = N\left[\tilde{\theta}, -H^{-1}(\tilde{\theta})\right]$, where $\tilde{\theta}$ is the mode, and $H^{-1}(\tilde{\theta})$ Hessian of log posterior at $\tilde{\theta}$.
- Normal approximation **turns an inference problem into an optimization problem**. VI does the same.
- **Variational Inference**: find $q(\theta) \in \mathcal{Q}$ that is closest to the posterior $p(\theta|\mathbf{x})$ in Kullback-Leibler sense

$$q^{\star}(\theta) = \underset{q(\theta) \in \mathcal{Q}}{\arg\min}\, \mathrm{KL}\left[q(\theta) || p(\theta|\mathbf{x})\right]$$

where

$$\mathrm{KL}\left[q(\theta) || p(\theta|\mathbf{x})\right] = \mathbb{E}_{q(\theta)} \log\left(\frac{q(\theta)}{p(\theta|\mathbf{x})}\right)$$

Variational Inference (VI)

- But ...

$$\mathrm{KL}\left[q(\theta)||p(\theta|\mathbf{x})\right] = \mathbb{E}_{q(\theta)} \log q(\theta) - \mathbb{E}_{q(\theta)} \log p(\theta|\mathbf{x})$$

  is intractable since $\mathbb{E}_{q(\theta)} \log p(\theta|\mathbf{x}) = \mathbb{E}_{q(\theta)} \log p(\mathbf{x},\theta) - \log p(\mathbf{x})$.

- But $\log p(x)$ is just a constant, so we can instead maximize **Evidence Lower BOund (ELBO)**:

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\theta)} \log p(\mathbf{x},\theta) - \mathbb{E}_{q(\theta)} \log q(\theta)$$

- ELBO is a lower bound for the (log) evidence (marginal likelihood) since $\mathrm{KL} \geq 0$ and

$$\log p(\mathbf{x}) = \mathrm{KL}\left[q(\theta)||p(\theta|\mathbf{x})\right] + \mathrm{ELBO}(q).$$

- Forward KL vs Backward KL.

Mean field approximation

- VI turns inference into a optimization problem

$$q^{\star}(\theta) = \underset{q(\theta)\in\mathcal{Q}}{\arg\max}\ \mathrm{ELBO}(q)$$

- Large $\mathcal{Q} \Rightarrow$ accurate approximation, hard optimization.
- Small $\mathcal{Q} \Rightarrow$ crude approximation, easy optimization.
- **Mean-field Variational Inference** (**MFVI**) uses the factorization

$$q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$$

- **No specific functional forms are assumed** for the $q_i(\theta)$.
- **Optimal densities** can be shown to satisfy [1]

$$q_i(\theta) \propto \exp\left(E_{-\theta_i} \ln p(\mathbf{y}, \theta)\right)$$

where $E_{-\theta_i}(\cdot)$ is the expectation with respect to $\prod_{i\neq j} q_j(\theta_j)$.
- Equivalent form with full conditionals

$$q_i(\theta) \propto \exp\left(E_{-\theta_i} \ln p(\theta_i|\theta_{-i}, \mathbf{y})\right)$$

Coordinate Ascent Variational Inference

Algorithm 1: Coordinate Ascent Variational Inference (CAVI)

Input: A model $p(\mathbf{x}, \theta)$, a data set $\mathbf{x}$, initial variational factors $q_j(\theta_j)$

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

  for $j \in \{1, ..., p\}$ **do**

    Set $q_j(\theta_j) \propto \exp\{\mathbb{E}_{-j} p(\theta_j | \theta_{-j}, \mathbf{x})\}$

  end

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

end

Output: A mean-field variational density $q(\theta) = \prod_{j=1}^{p} q_j(\theta_j)$

- Optimal $q_i(\theta)$ often turn out to be parametric (normal, gamma etc).
- Updates become just an **updating of variational hyperparameters**.
- See [2] for details for exponential families.

Mean-field VI - Bivariate Normal

- Bivariate normal model:

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \overset{iid}{\sim} N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \rho I\right],\ i = 1, ..., n$$

  where ρ is known. Indep normal prior.

- **Prior**:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \overset{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 0.25^2 I\right].$$

- Exact posterior by well known formulas:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} |\mathbf{x}_1, .., \mathbf{x}_n \overset{iid}{\sim} N\left(\tilde{\mu}, \tilde{\Omega}^{-1}\right),$$

  where

$$\tilde{\Omega} = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix}.$$

Mean-field VI - Bivariate Normal

- Exact posterior by well known formulas:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} |\mathbf{x}_1, .., \mathbf{x}_n \overset{iid}{\sim} N\left(\mathbf{m}, \Omega^{-1}\right),$$

where

$$\Omega = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix}.$$

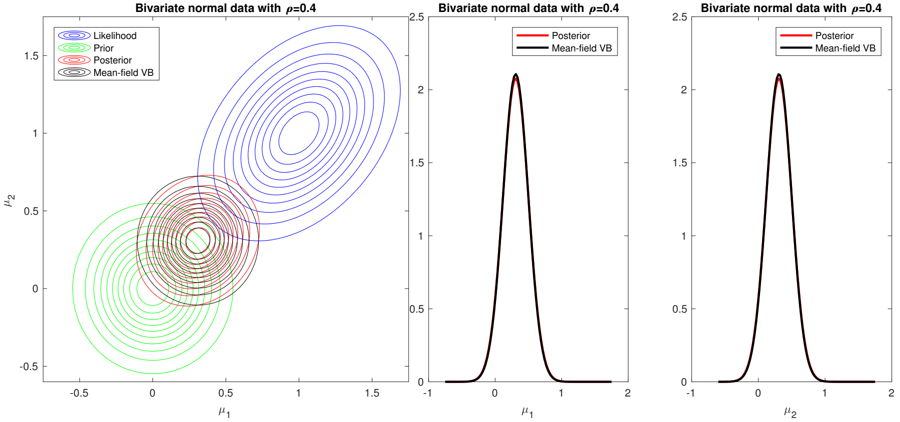
- **Mean-field approximation**: $q(\mu_1, \mu_2) = q_1(\mu_1) q_2(\mu_2)$.
- Find optimal mean-field VI solution by iterating on $\tilde{\mu}_1$ and $\tilde{\mu}_2$:

$$q_1^\star(\mu_1) \propto \exp\left[E_{q_2(\mu_2)} \ln p(\mu_1|\mu_2, \mathbf{x})\right] = N\left(\tilde{\mu}_1, \frac{1}{\omega_1^2}\right)$$

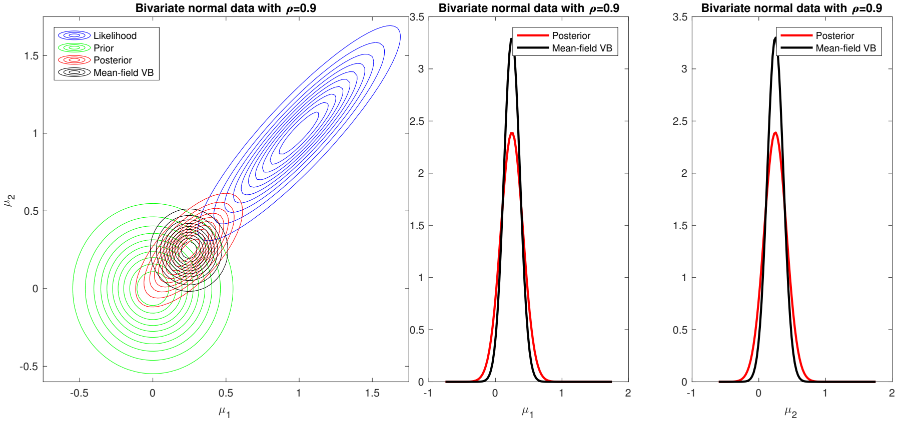
$$q_2^\star(\mu_2) \propto \exp\left[E_{q_1(\mu_1)} \ln p(\mu_2|\mu_1, \mathbf{x})\right] = N\left(\tilde{\mu}_2, \frac{1}{\omega_2^2}\right)$$

where $\tilde{\mu}_1 = m_1 - \frac{\omega_{12}}{\omega_1^2}(\tilde{\mu}_2 - m_2)$ and $\tilde{\mu}_2 = m_2 - \frac{\omega_{12}}{\omega_2^2}(\tilde{\mu}_1 - m_1)$.

Mean-field VI - Normal data with $\rho = 0.4$



MEAN-FIELD VI - NORMAL DATA WITH $\rho = 0.9$



Mean-field VI - Univariate normal data with unknown variance

- **Model**: $X_i|\theta,\sigma^2 \overset{iid}{\sim} N(\theta,\sigma^2)$.
- **Prior**: $\theta \sim N(\mu_0,\tau_0^2)$ **independent** of $\sigma^2 \sim Inv-\chi^2(\nu_0,\sigma_0^2)$.
- **Mean-field approximation**: $q(\theta,\sigma^2) = q_\theta(\theta)\cdot q_{\sigma^2}(\sigma^2)$.
- Optimal densities

$$q^*_\theta(\theta) \propto \exp\left[E_{q(\sigma^2)} \ln p(\theta,\sigma^2,\mathbf{x})\right]$$
$$q^*_{\sigma^2}(\sigma^2) \propto \exp\left[E_{q(\theta)} \ln p(\theta,\sigma^2,\mathbf{x})\right]$$

Normal model - VI algorithm

- **Variational density for σ^2**

$$\sigma^2 \sim Inv - \chi^2\left(\tilde{\nu}_n, \tilde{\sigma}_n^2\right)$$

where $\tilde{\nu}_n = \nu_0 + n$ and $\tilde{\sigma}_n = \frac{\nu_0\sigma_0^2 + \sum_{i=1}^n (x_i - \tilde{\mu}_n)^2 + n \cdot \tilde{\tau}_n^2}{\nu_0 + n}$

- **Variational density for θ**

$$\theta \sim N\left(\tilde{\mu}_n, \tilde{\tau}_n^2\right)$$

where

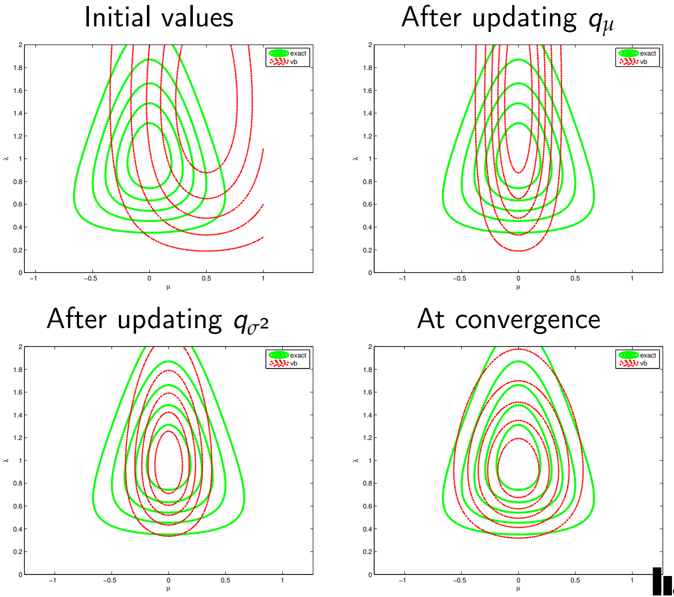
$$\tilde{\tau}_n^2 = \frac{1}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

$$\tilde{\mu}_n = \tilde{w}\bar{x} + (1 - \tilde{w})\mu_0,$$

where

$$\tilde{w} = \frac{\frac{n}{\tilde{\sigma}_n^2}}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

Normal example from Murphy [3] ($\lambda = 1/\sigma^2$)



Probit regression

- **Model**:
$$\Pr(y_i = 1 | \mathbf{x}_i) = \Phi(\mathbf{x}_i^T \beta)$$
- **Prior**: $\beta \sim N(0, \Sigma_\beta)$. For example: $\Sigma_\beta = \tau^2 I$.
- **Latent variable formulation** with $u = (u_1, ..., u_n)'$
$$\mathbf{u} | \beta \sim N(\mathbf{X}\beta, 1)$$
and
$$y_i = \begin{cases} 0 & \text{if } u_i \leq 0 \\ 1 & \text{if } u_i > 0 \end{cases}$$
- Factorized **variational approximation**
$$q(\mathbf{u}, \beta) = q_{\mathbf{u}}(\mathbf{u}) q_\beta(\beta)$$

VI FOR PROBIT REGRESSION [1]

- VI posterior

$$\beta \sim N\left(\tilde{\mu}_\beta, \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\right)$$

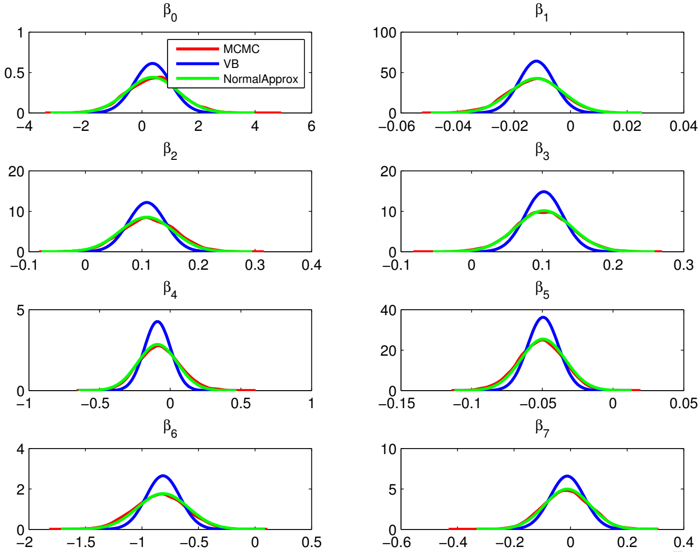
where

$$\tilde{\mu}_\beta = \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\mathbf{X}^T\tilde{\mu}_{\mathbf{u}}$$

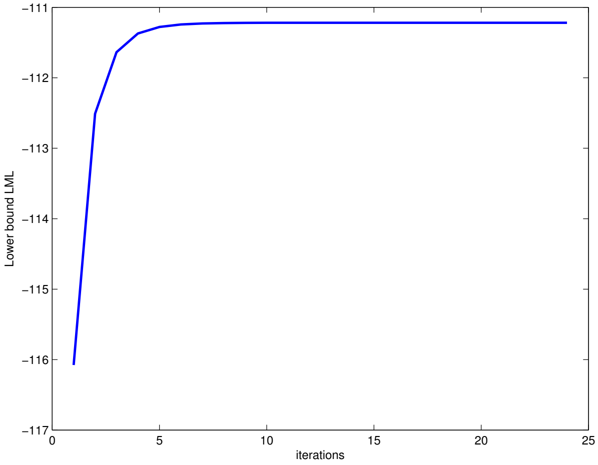
and

$$\tilde{\mu}_{\mathbf{u}} = \mathbf{X}\tilde{\mu}_\beta + \frac{\phi\left(\mathbf{X}\tilde{\mu}_\beta\right)}{\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right)^{\mathbf{y}}\left[\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right) - \mathbf{1}_n\right]^{\mathbf{1}_n - \mathbf{y}}}.$$

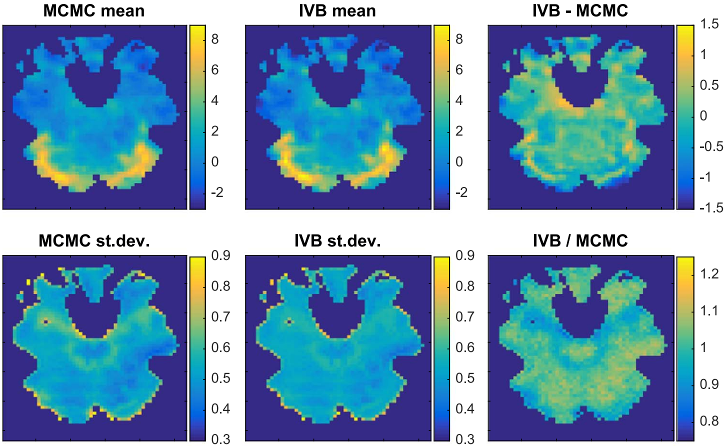
PROBIT EXAMPLE (N=200 OBSERVATIONS)



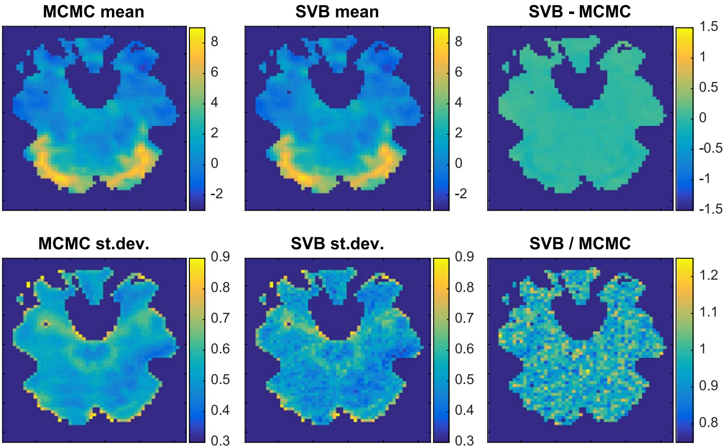
Probit example - ELBO



Spatial neuroimaging - MCMC vs Mean-field [4]



SPATIAL NEUROIMAGING - MCMC VS VI [4]



VI for latent variable models [2]

- Example: **Normal mixture model**

$$p(x_i) = \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)$$

$$z_i \sim \text{Categorical}(\pi_1, ...\pi_K) \quad \textit{and} \quad x_i | z_i \overset{indep}{\sim} N(\mu_{z_i}, \sigma_{z_i}^2)$$

- **Global variables** (parameters): $\theta = (\mu_{1:K}, \sigma_{1:K}, \pi_{1:K})^T$
- **Local parameters** (latent variables): $\mathbf{z} = (z_1, ..., z_n)^T$.
- CAVI:
  - (for $i = 1, ...n$) Update $q_{\varphi_i}(z_i)$ wrt local variational parameters φ_i
  - Update $q_\lambda(\theta)$ wrt global variational parameter λ
- Explicit updates for exponential family models.
- Slow for large n.

CAVI FOR MIXTURE OF NORMALS [2]

Algorithm 2: CAVI for a Gaussian mixture model

Input: Data $x_{1:n}$, number of components K, prior variance of component means σ^2
Output: Variational densities $q(\mu_k; m_k, s_k^2)$ (Gaussian) and $q(z_i; \varphi_i)$ (K-categorical)
Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s_{1:K}^2$, and $\boldsymbol{\varphi} = \varphi_{1:n}$
while *the ELBO has not converged* **do**
  for $i \in \{1, \dots, n\}$ **do**
    Set $\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2\}$
  end
  for $k \in \{1, \dots, K\}$ **do**

$$\text{Set } m_k \longleftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

$$\text{Set } s_k^2 \longleftarrow \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

  end
  Compute ELBO$(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$
end
return $q(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

Stochastic Variational Inference (SVI) [5]

- Conditionally conjugate models with (local) latent variables.
- Aim: variational approximation for global variables θ.
- Coordinate ascent is replaced by **gradient-based optimization**.
- **CAVI + Stochastic optimization** with noisy gradients
- Unbiased estimate of the gradient from a **minibatch** of observations.
- Only need to update the variational for latents corresponding to observations in the minibatch. Fast.

Black Box Variational Inference (BBVI)

- Mean-field makes $\mathcal{Q}$ conveniently small, but the approximation can be poor. Especially for the variance.
- Recent push to use more expressive $q \in \mathcal{Q}$ approximations. Harder optimization. $\mathbb{E}_q$ often intractable.
- BBVI also called doubly-stochastic:
  1. BBVI approximates $\mathbb{E}_q$ by **Monte Carlo integration**.
  2. Gradient-based **stochastic optimization** with unbiased estimates of the gradients.

Black Box Variational Inference (BBVI)[6]

- Assume: $q_{\mu,\mathbf{C}}(\theta) = N(\theta|\mu, \mathbf{C}\mathbf{C}^T)$. Optimize over $\mu \in \mathbb{R}^p$ and $\mathbf{C}$, a lower-triangular positive definite matrix.
- More generally, assume VI approximation of $p(\theta|\mathbf{x}) \approx q_{\mu,\mathbf{C}}(\theta)$ from correlating a standard random vector $\mathbf{z}$ with density $\phi(\mathbf{z})$

$$\theta = \mu + \mathbf{C}\mathbf{z}.$$

- Example: if $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$, then $q_{\mu,\mathbf{C}}(\theta) = N(\theta|\mu, \mathbf{C}\mathbf{C}^T)$.
- Generally, by the change-of-variable formula

$$q_{\mu,\mathbf{C}}(\theta) = \frac{1}{|\mathbf{C}|}\phi\left(\mathbf{C}^{-1}(\theta - \mu)\right)$$

- Using the change of variable formula:

$$\begin{aligned}\mathrm{ELBO}(\mu, \mathbf{C}) &= \mathbb{E}_{q(\theta)} \log p(\mathbf{x}, \theta) - \mathbb{E}_{q(\theta)} \log q(\theta) \\ &= \mathbb{E}_{\phi(\mathbf{z})} \log p(\mathbf{x}, \mu + \mathbf{C}\mathbf{z}) + \log|\mathbf{C}| + H_\phi,\end{aligned}$$

where $\log \mathbf{C} = \sum_{j=1}^{p} \log C_{jj}$ and H_ϕ is the entropy of $\phi(\mathbf{z})$, which is constant wrt μ and $\mathbf{C}$.

Black Box Variational Inference (BBVI)[6]

- From previous slide:

$$\mathrm{ELBO}(\mu, \mathbf{C}) = \mathbb{E}_{\phi(\mathbf{z})} \log p(\mathbf{x}, \mu + \mathbf{C}\mathbf{z}) + \log|\mathbf{C}| + H_\phi.$$

- Gradient-based optimization. **Swapping order of ∇ and $\mathbb{E}$.**

$$\nabla_\mu \mathrm{ELBO}(\mu, \mathbf{C}) = \mathbb{E}_{\phi(\mathbf{z})} \left[\nabla_\mu \log p(\mathbf{x}, \mu + \mathbf{C}\mathbf{z})\right]$$
$$\nabla_{\mathbf{C}} \mathrm{ELBO}(\mu, \mathbf{C}) = \mathbb{E}_{\phi(\mathbf{z})} \left[\nabla_{\mathbf{C}} \log p(\mathbf{x}, \mu + \mathbf{C}\mathbf{z})\right] + \Delta_{\mathbf{C}}$$

where $\Delta_{\mathbf{C}} \equiv \mathrm{diag}(1/c_{11}, ..., 1/c_{pp})$ and $\nabla_{\mathbf{C}} \log p(\mathbf{x}, \mu + \mathbf{C}\mathbf{z})$ a lower triangular matrix with partial derivatives.

- We can change back to θ to get alternative expressions:

$$\nabla_\mu \mathrm{ELBO}(\mu, \mathbf{C}) = \mathbb{E}_{q(\theta|\mu,\mathbf{C})} \left[\nabla_\theta \log p(\mathbf{x}, \theta)\right]$$
$$\nabla_{\mathbf{C}} \mathrm{ELBO}(\mu, \mathbf{C}) = \mathbb{E}_{q(\theta|\mu,\mathbf{C})} \left[\nabla_\theta \log p(\mathbf{x}, \theta)(\theta - \mu)^T \mathbf{C}^{-T}\right] + \Delta_{\mathbf{C}}$$

where $\nabla_\theta \log p(\mathbf{x}, \theta)\mathbf{z}^T$ is the lower triangular part after outer product.

Doubly Stochastic Variational Inference [6]

Algorithm 2: Doubly Stochastic Variational Inference (DSVI)

Input: A gradient function $\nabla_\theta \log p(\mathbf{x}, \theta)$, a simulator $\phi(\mathbf{z})$, a data set $\mathbf{x}$, initial values μ^0, $\mathbf{C}^0$, learning rate $\{\rho_t\}_{t\geq 1}$.

Set $t = 0$

while *convergence criterion not met* **do**

$t = t + 1$

$\mathbf{z} \sim \phi(\mathbf{z})$

$\theta^{(t-1)} = \mathbf{C}^{(t-1)}\mathbf{z} + \mu^{(t-1)}$

$\mu^{(t)} = \mu^{(t-1)} + \rho_t \nabla_\theta \log p(\mathbf{x}, \theta^{(t-1)})$

$\mathbf{C}^{(t)} = \mathbf{C}^{(t-1)} + \rho_t \left(\nabla_\theta \log p(\mathbf{x}, \theta^{(t-1)}) \times \mathbf{z}^T + \Delta_{\mathbf{C}^{(t-1)}}\right)$

end

Output: Optimal variational parameters $\mu^\star$ and $\mathbf{C}^\star$.

Extensions of BBVI

- Replacing covariance matrix $\mathbf{CC}^T$ in $q_{\mu,\mathbf{C}}(\theta) = N(\theta|\mu, \mathbf{CC}^T)$ by **sparse precision matrix** [7]. Very useful for models with sparse conditional dependence structure. Spatial. State-space.
- Modeling the variational covariance matrix by **factor structure**. [8]
- **Variational autoencoders**. $q_{\mu,\mathbf{C}}(\theta) = N(\theta|\mu(\mathbf{x}), \Sigma(\mathbf{x}))$, where mean and covariance is a deep neural network. [9]
- **Beyond Gaussian VI**:
  - Copula VI [10]
  - Gaussian processes VI [11]
  - Normalizing flows [12]
  - Variational Sequential Monte Carlo **[13]**
  - ...

J. Ormerod and M. Wand, "Explaining variational approximations," *The American Statistician*, vol. 64, no. 2, pp. 140–153, 2010.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, no. just-accepted, 2017.

K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

P. Sidén, A. Eklund, D. Bolin, and M. Villani, "Fast bayesian whole-brain fmri analysis with spatial 3d priors," *NeuroImage*, vol. 146, pp. 211–225, 2017.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

- M. Titsias and M. Lázaro-Gredilla, "Doubly stochastic variational bayes for non-conjugate inference," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979, 2014.
- L. S. Tan and D. J. Nott, "Gaussian variational approximation with sparse precision matrices," *Statistics and Computing*, pp. 1–17, 2017.
- V. M.-H. Ong, D. J. Nott, and M. S. Smith, "Gaussian variational approximation with a factor covariance structure," *arXiv preprint arXiv:1701.03208*, 2017.
- D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- D. Tran, D. Blei, and E. M. Airoldi, "Copula variational inference," in *Advances in Neural Information Processing Systems*, pp. 3564–3572, 2015.
- D. Tran, R. Ranganath, and D. M. Blei, "The variational gaussian process," *arXiv preprint arXiv:1511.06499*, 2015.

D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *arXiv preprint arXiv:1505.05770*, 2015.

C. A. Naesseth, S. W. Linderman, R. Ranganath, and D. M. Blei, "Variational sequential monte carlo," *arXiv preprint arXiv:1705.11140*, 2017.