A beginner's guide to variational inference

Haziq Jamil

Social Statistics
London School of Economics and Political Science

1 February 2018

Social Statistics Meeting

http://socialstats.haziqj.ml

Outline

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)}|\mathbf{z}_{-j},\mathbf{y}) = h(\mathbf{z}^{(j)}) \exp\big(\eta_j(\mathbf{z}_{-j},\mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)\big).$$

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)}|\mathbf{z}_{-j},\mathbf{y}) = h(\mathbf{z}^{(j)}) \exp\big(\eta_j(\mathbf{z}_{-j},\mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)\big).$$

- Then, from (??),

$$\begin{aligned}
\tilde{q}_j(\mathbf{z}^{(j)}) &\propto \exp\big(\,\mathrm{E}_{-j}[\log p(\mathbf{z}^{(j)}|\mathbf{z}_{-j},\mathbf{y})]\big) \\
&= \exp\Big(\log h(\mathbf{z}^{(j)}) + \mathrm{E}[\eta_j(\mathbf{z}_{-j},\mathbf{y})] \cdot \mathbf{z}^{(j)} - \mathrm{E}[A(\eta_j)]\Big) \\
&\propto h(\mathbf{z}^{(j)}) \exp\Big(\,\mathrm{E}[\eta_j(\mathbf{z}_{-j},\mathbf{y})] \cdot \mathbf{z}^{(j)}\Big)
\end{aligned}$$

is also in the same exponential family.

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)}|\mathbf{z}_{-j}, \mathbf{y}) = h(\mathbf{z}^{(j)}) \exp\big(\eta_j(\mathbf{z}_{-j}, \mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)\big).$$

- Then, from (??),

$$\begin{aligned} \tilde{q}_j(\mathbf{z}^{(j)}) &\propto \exp\big(\, \mathrm{E}_{-j}[\log p(\mathbf{z}^{(j)}|\mathbf{z}_{-j}, \mathbf{y})]\big) \\ &= \exp\Big(\log h(\mathbf{z}^{(j)}) + \mathrm{E}[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)} - \mathrm{E}[A(\eta_j)]\Big) \\ &\propto h(\mathbf{z}^{(j)}) \exp\Big(\, \mathrm{E}[\eta_j(\mathbf{z}_{-j}, \mathbf{y})] \cdot \mathbf{z}^{(j)}\Big) \end{aligned}$$

  is also in the same exponential family.
- C.f. Gibbs conditional densities.

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

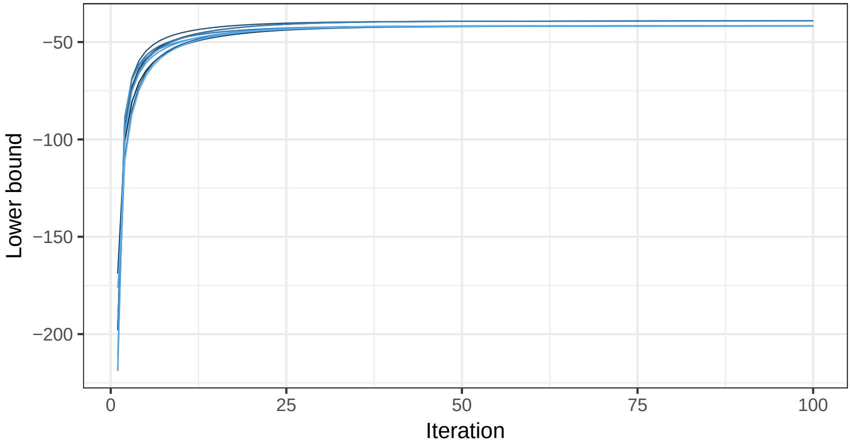
$$p(\mathbf{z}^{(j)}|\mathbf{z}_{-j},\mathbf{y}) = h(\mathbf{z}^{(j)})\exp\big(\eta_j(\mathbf{z}_{-j},\mathbf{y})\cdot\mathbf{z}^{(j)} - A(\eta_j)\big).$$

- Then, from (??),

$$\begin{aligned}
\tilde{q}_j(\mathbf{z}^{(j)}) &\propto \exp\big(\,\mathrm{E}_{-j}[\log p(\mathbf{z}^{(j)}|\mathbf{z}_{-j},\mathbf{y})]\big) \\
&= \exp\Big(\log h(\mathbf{z}^{(j)}) + \mathrm{E}[\eta_j(\mathbf{z}_{-j},\mathbf{y})]\cdot\mathbf{z}^{(j)} - \mathrm{E}[A(\eta_j)]\Big) \\
&\propto h(\mathbf{z}^{(j)})\exp\Big(\,\mathrm{E}[\eta_j(\mathbf{z}_{-j},\mathbf{y})]\cdot\mathbf{z}^{(j)}\Big)
\end{aligned}$$

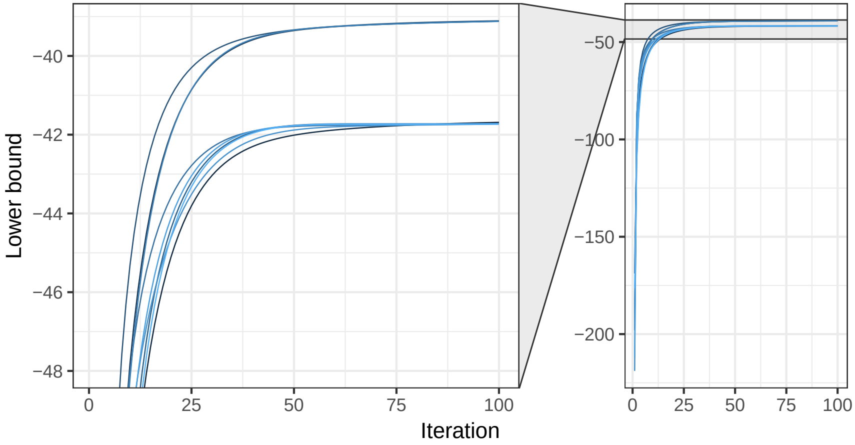
  is also in the same exponential family.
- C.f. Gibbs conditional densities.
- **ISSUE**: What if not in exponential family? Importance sampling or Metropolis sampling.

Non-convexity of ELBO



- CAVI only guarantees converges to a local optimum.
- Multiple local optima may exist.

Non-convexity of ELBO



- CAVI only guarantees converges to a local optimum.
- Multiple local optima may exist.

Zero-forcing vs Zero-avoiding

- Back to the KL divergence:

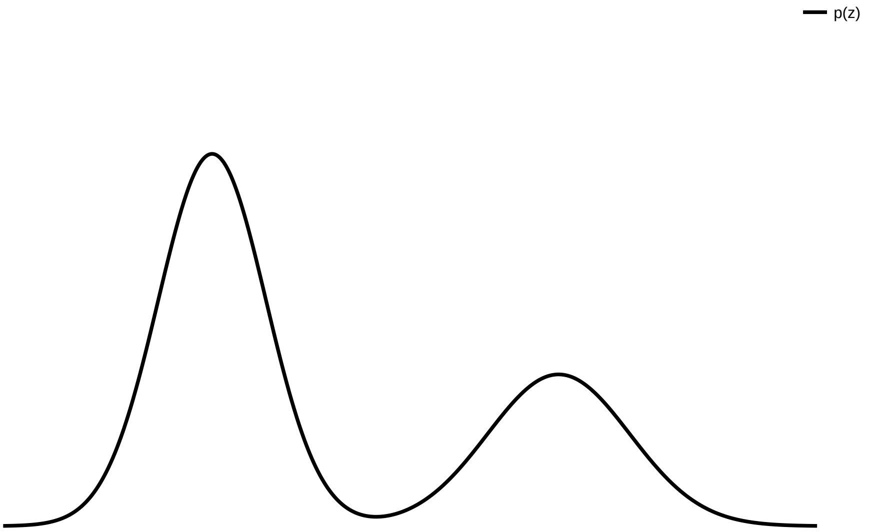
$$\mathrm{KL}(q\|p) = \int \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})} q(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

- $\mathrm{KL}(q\|p)$ is large when $p(\mathbf{z}|\mathbf{y})$ is close to zero, unless $q(\mathbf{z})$ is also close to zero (*zero-forcing*).
- **ISSUE**: What about other measures of closeness? For instance,

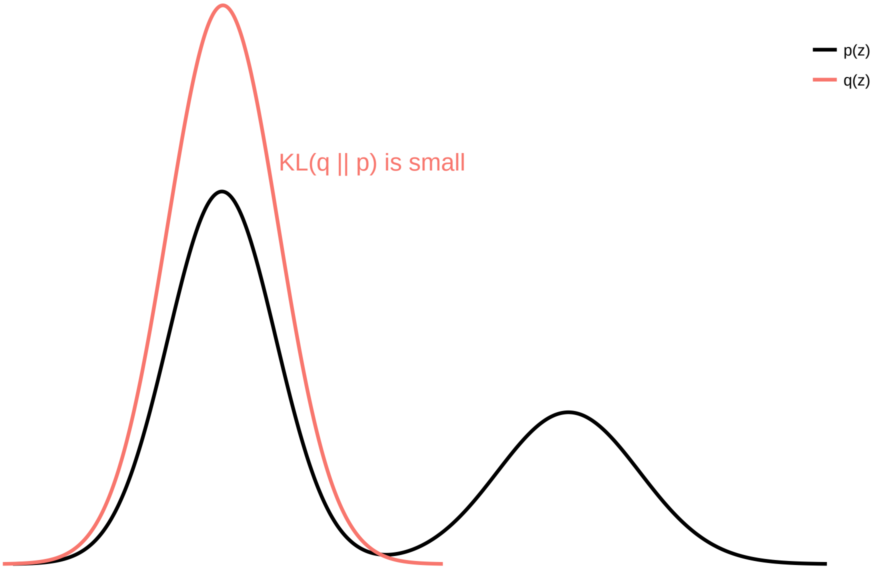
$$\mathrm{KL}(p\|q) = \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{y})} p(\mathbf{z}|\mathbf{y})\,\mathrm{d}\mathbf{z}.$$

- This gives the Expectation Propagation (EP) algorithm.
- It is *zero-avoiding*, because $\mathrm{KL}(p\|q)$ is small when both $p(\mathbf{z}|\mathbf{y})$ and $q(\mathbf{z})$ are non-zero.

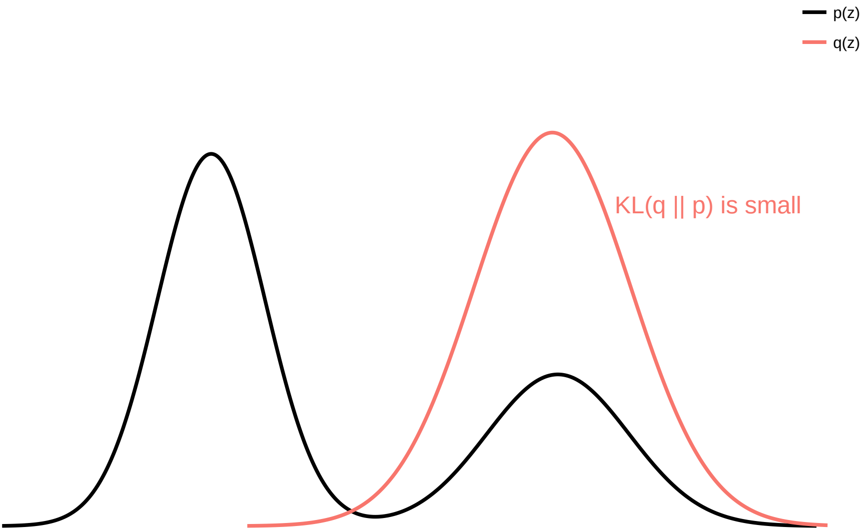
Zero-forcing vs Zero-avoiding (cont.)



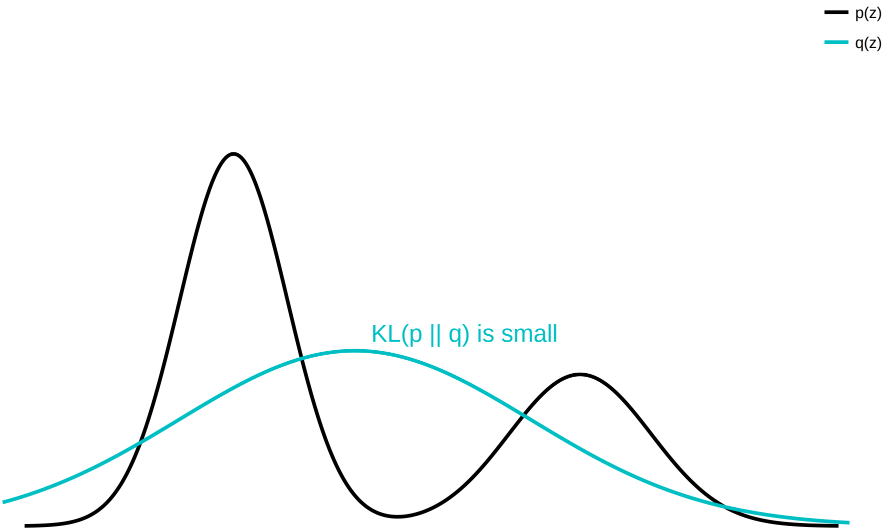
Zero-forcing vs Zero-avoiding (cont.)



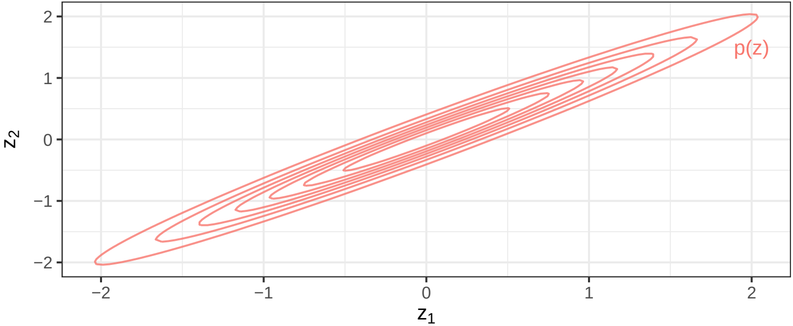
Zero-forcing vs Zero-avoiding (cont.)



Zero-forcing vs Zero-avoiding (cont.)

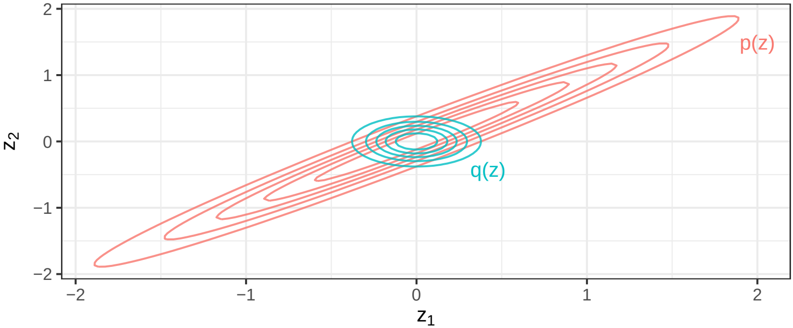


Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.

Distortion of higher order moments

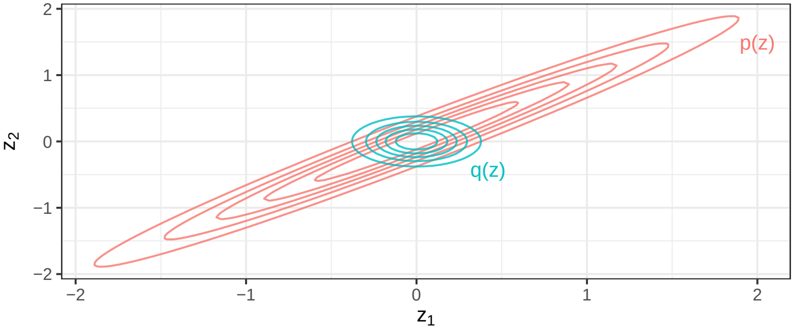


- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q(z_1)q(z_2)$ yields

$$\tilde{q}(z_1) = \mathrm{N}(z_1 | \mu_1, \boldsymbol{\Psi}_{11}^{-1}) \quad \text{and} \quad \tilde{q}(z_2) = \mathrm{N}(z_2 | \mu_2, \boldsymbol{\Psi}_{22}^{-1})$$

and by definition, $\mathrm{Cov}(z_1, z_2) = 0$ under $\tilde{q}$.

Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q(z_1)q(z_2)$ yields

$$\tilde{q}(z_1) = \mathrm{N}(z_1|\mu_1, \boldsymbol{\Psi}_{11}^{-1}) \ \text{ and } \ \tilde{q}(z_2) = \mathrm{N}(z_2|\mu_2, \boldsymbol{\Psi}_{22}^{-1})$$

  and by definition, $\mathrm{Cov}(z_1, z_2) = 0$ under $\tilde{q}$.
- This leads to underestimation of variances (widely reported in the literature—Zhao and Marriott 2013).

Quality of approximation

- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).

Quality of approximation

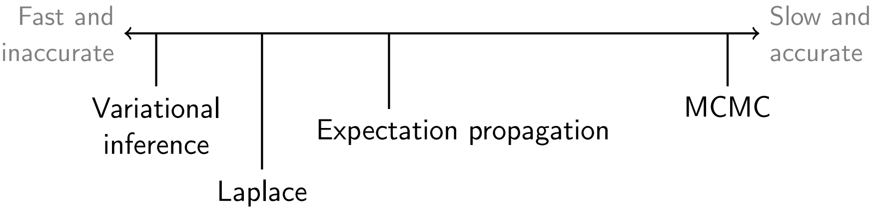
- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).
- But not much can be said about the quality of approximation.

Quality of approximation

- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).
- But not much can be said about the quality of approximation.
- Statistical properties not well understood—what is its statistical profile relative to the exact posterior?

Quality of approximation

- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).
- But not much can be said about the quality of approximation.
- Statistical properties not well understood—what is its statistical profile relative to the exact posterior?
- Speed trumps accuracy?



Advanced topics

- Local variational bounds
  - Not using the mean-field assumption.
  - Instead, find a bound for the marginalising integral $\mathcal{I}$.
  - Used for Bayesian logistic regression as follows:

$$\mathcal{I} = \int \text{expit}(x^\top \beta) p(\beta) \, \mathrm{d}\beta \geq \int f(x^\top \beta, \xi) p(\beta) \, \mathrm{d}\beta.$$

Advanced topics

- Local variational bounds
  - Not using the mean-field assumption.
  - Instead, find a bound for the marginalising integral $\mathcal{I}$.
  - Used for Bayesian logistic regression as follows:

$$\mathcal{I} = \int \text{expit}(x^\top \beta) p(\beta) \, \text{d}\beta \geq \int f(x^\top \beta, \xi) p(\beta) \, \text{d}\beta.$$

- Stochastic variational inference
  - VI on its own doesn't offer much computational advantages.
  - Use ideas from stochastic optimisation—gradient based improvement of ELBO from subsamples of the data.
  - Scales to massive data.

Advanced topics

- Local variational bounds
  - Not using the mean-field assumption.
  - Instead, find a bound for the marginalising integral $\mathcal{I}$.
  - Used for Bayesian logistic regression as follows:

$$\mathcal{I} = \int \operatorname{expit}(x^\top \beta) p(\beta) \, \mathrm{d}\beta \geq \int f(x^\top \beta, \xi) p(\beta) \, \mathrm{d}\beta.$$

- Stochastic variational inference
  - VI on its own doesn't offer much computational advantages.
  - Use ideas from stochastic optimisation—gradient based improvement of ELBO from subsamples of the data.
  - Scales to massive data.
- Black box variational inference
  - Beyond exponential families and model-specific derivations.

End

Thank you!

Slides and source code are made available at: http://socialstats.haziqj.ml

References I

Gunawardana, A. and W. Byrne (2005). "Convergence theorems for generalized alternating minimization procedures". *Journal of machine learning research* 6.Dec, pp. 2049–2073.

Zhao, H. and P. Marriott (2013). "Diagnostics for Variational Bayes approximations". arXiv: 1309.5117.

4 Additional material