STK4051/9051 Computational statistics

Geir Storvik

Variational inference

Bayesian inference

- Assume observations $\boldsymbol{x}$, variables of interest $\boldsymbol{z}$
  - $\boldsymbol{z}$ may include parameters and/or latent variables
- Core: Posterior distribution

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})}$$

- Inference:

$$E[h(\boldsymbol{z})|\boldsymbol{x}] = \int_{\boldsymbol{z}} h(\boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}$$

- Main problem: $p(\boldsymbol{z}|\boldsymbol{x})$ can be very complex
- "Standard" methods such as MCMC do not scale well with big data

Variational inference

- Main source: Blei et al. (2017)
- Idea:
  - Approximate $p(\boldsymbol{z}|\boldsymbol{x})$ by a simpler $q^*(\boldsymbol{z})$
  - Perform inference by

$$E[h(\boldsymbol{z})|\boldsymbol{x}] \approx \int_{\boldsymbol{z}} h(\boldsymbol{z}) q^*(\boldsymbol{z}) d\boldsymbol{z} \qquad (*)$$

- Question: How to find $q^*(\boldsymbol{z})$?
- Approach: Define $q^*(\boldsymbol{z})$ as

$$q^*(\boldsymbol{z}) = \underset{q(\boldsymbol{z}) \in \mathcal{Q}}{\arg\min}\, \mathcal{D}(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x}))$$

  where
  - $\mathcal{Q}$ is some class of distributions; variational family
  - $\mathcal{D}(\cdot, \cdot)$ is some distance measure between distributions
  - Note: q^* will depend on $\boldsymbol{x}$ although not explicitly shown in the notation!
- Integration problem now mainly transformed to an optimization problem
  - Integration of (*) assumed to be simple

Specification of distance measure

- Most common: Kullback-Leibler divergence:

$$\begin{aligned}\mathcal{D}(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x}) =& KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) \\ =& \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}}\right) d\boldsymbol{z} \\ =& E^q[\log q(\boldsymbol{z})] - E^q[\log p(\boldsymbol{z}|\boldsymbol{x})]\end{aligned}$$

- Properties:
  - $KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) \geq 0$
  - $KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) = 0 \Rightarrow q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x})$
  - $KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) \neq KL(p(\boldsymbol{z}|\boldsymbol{x})||q(\boldsymbol{z}))$

The evidence lower bound

- Rewriting of the KL divergence:

$$\begin{aligned} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) =& E^q[\log q(\boldsymbol{z})] - E^q[\log p(\boldsymbol{z}|\boldsymbol{x})] \\ =& E^q[\log q(\boldsymbol{z})] - E^q[\log p(\boldsymbol{z}, \boldsymbol{x})] + E^q[\log p(\boldsymbol{x})] \\ =& E^q[\log q(\boldsymbol{z})] - E^q[\log p(\boldsymbol{z}, \boldsymbol{x})] + \log p(\boldsymbol{x}) \end{aligned}$$

- Minimizing KL is equivalent maximizing

$$\text{ELBO}(q) = E^q[\log p(\boldsymbol{z}, \boldsymbol{x})] - E^q[\log q(\boldsymbol{z})]$$

- ELBO = Evidence lower bound:

$$\text{ELBO}(q) = \log p(\boldsymbol{x}) - KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) \leq \log p(\boldsymbol{x})$$

- Practical formulation:

$$\begin{aligned} \text{ELBO}(q) =& E^q[\log p(\boldsymbol{x}|\boldsymbol{z})] + E^q[\log p(\boldsymbol{z})] - E^q[\log q(\boldsymbol{z})] \\ =& E^q[\log p(\boldsymbol{x}|\boldsymbol{z})] - KL(q(\boldsymbol{z})||p(\boldsymbol{z})) \end{aligned}$$

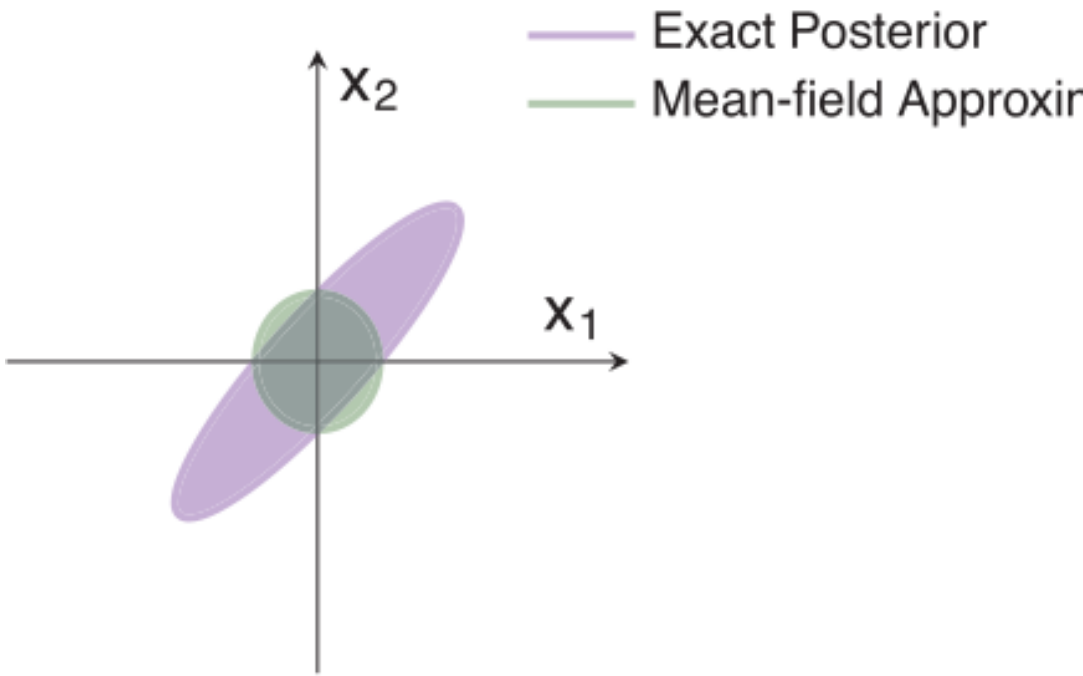
  - First term: Fit to data
  - Second part: Penalty term

Mean-field variational family

- Common choice: Mean-field variational family

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j(z_j)$$

  that is independence structure
- Also need to specify $q_j(z_j)$.
  - For continuous variables, Gaussian is possible
  - For categorical variables, any discrete distribution.



- Typically, point esimates quite accurate, uncertainty measures too low

Example - Gaussian mixtures

- Model

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2), && k = 1, \ldots, K \\ c_i &\sim \text{Categorical}(1/K, \ldots, 1/K), && i = 1, \ldots, n \\ x_i | c_i, \boldsymbol{\mu} &\sim \mathcal{N}(\mu_{c_i}, 1) \end{aligned}$$

  Of interest

$$p(\boldsymbol{\mu}, \boldsymbol{c} | \boldsymbol{x}) \propto p(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

- Variational approximation:

$$q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \boldsymbol{\phi}_i)$$

  where $\boldsymbol{\phi}_i = (\phi_{i,1}, \ldots, \phi_{i,K})$, $\sum_{j=1}^{K} \phi_{i,k} = 1$ and

$$\begin{aligned} q(\mu_k; m_k, s_k^2) &= \mathcal{N}(\mu_k; m_k, s_k^2) \\ q(c_i; \boldsymbol{\phi}_i) &= \phi_{i,c_i} \end{aligned}$$

Optimization - CAVI

- CAVI = Coordinate ascent variational inference
- Given all other variables, the optimal $q_j^*(z_j)$ is

$$\begin{aligned} q_j^*(z_j) &\propto \exp\{E_{\boldsymbol{z}_j}^q[\log p(z_j|\boldsymbol{z}_{-j}, \boldsymbol{x})]\} \\ &\propto \exp\{E_{\boldsymbol{z}_j}^q[\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x})]\} \end{aligned}$$

- Algorithm

Input: A model $p(\boldsymbol{x}, \boldsymbol{z})$, a data set $\boldsymbol{x}$
Output: A variational density $q(\boldsymbol{x}) = \prod_{j=1}^m q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while *the ELBO has not converged* **do**
  for $j \in \{1, \ldots, m\}$ **do**
    Set $q_j(z_j) \propto \exp\{E_{\boldsymbol{z}_j}^q[\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x})]\}$
  end
  Compute ELBO$(q) = E^q[\log p(\boldsymbol{z}, \boldsymbol{x})] - E^q[\log q(\boldsymbol{z})]$
end

- Note: Similar structure as the Gibbs sampler!

Example - updating for c_i:

$$
\begin{aligned}
q^*(c_i; \boldsymbol{\phi}_i) &\propto \exp\{E_{\boldsymbol{c}_{-i},\boldsymbol{\mu}}[\log p(c_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu}, \boldsymbol{x})]\} \\
&\propto \exp\{E_{\boldsymbol{c}_{-i},\boldsymbol{\mu}}[\log(p(c_i)) + \log(p(\boldsymbol{c}_{-i})) + \log(\boldsymbol{\mu}) + \sum_{j=1}^{n} \log p(x_j|c_j, \boldsymbol{\mu})]\} \\
&\propto \exp\{E_{\boldsymbol{c}_{-i},\boldsymbol{\mu}}[\log(p(c_i)) + \log p(x_i|c_i, \boldsymbol{\mu})]\} \\
&\propto \exp\{E_{\boldsymbol{c}_{-i},\boldsymbol{\mu}}[\tfrac{1}{K} + \log p(x_i|\mu_{c_i})]\} \\
&\propto \exp\{E_{\boldsymbol{\mu}}[[-\tfrac{1}{2}(x_i - \mu_{c_i})^2]\} \\
&\propto \exp\{E_{\boldsymbol{\mu}}[-\tfrac{1}{2}(x_i^2 - 2x_i\mu_{c_i} + \mu_{c_i}^2)]\} \\
&\propto \exp\{E_{\boldsymbol{\mu}}[x_i m_{c_i} - \tfrac{1}{2}s_{c_i}^2 - \tfrac{1}{2}m_{c_i}^2]\}
\end{aligned}
$$

Example, opdating for μ_k

$$
\begin{aligned}
q^*(\mu_k; m_k, s_k^2) &\propto \exp\{E_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}[\log p(c, \boldsymbol{\mu}, \boldsymbol{x})]\} \\
&\propto \exp\{E_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}[\log p(\mu_k) + \sum_{i=1}^{n} \log p(x_i|c_i, \boldsymbol{\mu})]\} \\
&\propto \exp\{E_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}[-\tfrac{1}{2\sigma^2}\mu_k^2 + \sum_{i=1}^{n} I(c_i = k) \log p(x_i|\boldsymbol{\mu}_k)]\} \\
&\propto \exp\{E_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}[-\tfrac{1}{2\sigma^2}\mu_k^2 + \sum_{i=1}^{n} \phi_{ik}[-\tfrac{1}{2}(x_i - \mu_k)^2]\} \\
&\propto \exp\{E_{\boldsymbol{c},\boldsymbol{\mu}_{-k}}[-\tfrac{1}{2\sigma^2}\mu_k^2 + -\tfrac{1}{2}\sum_{i=1}^{n} \phi_{ik}[x_i^2 - 2x_i\mu_k + \mu_k^2]\} \\
&\propto \exp\{-\tfrac{1}{2}[(\tfrac{1}{\sigma^2} + \sum_{i=1}^{n} \phi_{ik})\mu_k^2 - 2\sum_{i=1}^{n} \phi_{ik} x_i \mu_k]\} \\
&\propto \exp\{-\tfrac{1}{2}(\tfrac{1}{\sigma^2} + \sum_{i=1}^{n} \phi_{ik})[\mu_k - \frac{\sum_{i=1}^{n} \phi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^{n} \phi_{ik}}]^2\}
\end{aligned}
$$

Example - calculating ELBO

$$
\begin{aligned}
\text{ELBO}(q) =& E^q[\log p(\boldsymbol{z}, \boldsymbol{x})] - E^q[\log q(\boldsymbol{z})] \\
=& E^q[\sum_{i=1}^{n} \log p(c_i) + \sum_{k=1}^{K} \log p(\mu_k) + \sum_{i=1}^{n} \log p(x_i|c_i, \boldsymbol{\mu})] - \\
& E^q[\sum_{i=1}^{n} \log q(c_i) + \sum_{k=1}^{K} \log q(\mu_k)] \\
=& E^q[\sum_{i=1}^{n} \log(\tfrac{1}{K}) - \tfrac{1}{2\sigma^2} \sum_{k=1}^{K} \mu_k^2 - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} I(c_i = k)(x_i - \mu_k)^2] - \\
& E^q[\sum_{i=1}^{n} \sum_{k=1}^{K} \phi_{ik} \log(\phi_{ik}) + \sum_{k=1}^{K} [-\tfrac{1}{2} \log(s_k^2) - \tfrac{1}{2s_k^2}(\mu_k - m_k)^2)]] \\
=& E^q[-\tfrac{1}{2\sigma^2} \sum_{k=1}^{K} [s_k^2 + m_j^2] - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \phi_{ik} [x_i^2 - 2x_i\mu_k + \mu_k^2] - \\
& E^q[\sum_{i=1}^{n} \sum_{k=1}^{K} \phi_{ik} \log(\phi_{ik}) + \sum_{k=1}^{K} [-\tfrac{1}{2} \log(s_k^2) - \tfrac{1}{2}] \\
=& -\tfrac{1}{2\sigma^2} \sum_{k=1}^{K} [s_k^2 + m_j^2] + \sum_{i=1}^{n} \sum_{k=1}^{K} \phi_{ik} [x_i m_k - \tfrac{1}{2} s_k^2 - \tfrac{1}{2} m_k^2] - \\
& \sum_{i=1}^{n} \sum_{k=1}^{K} \phi_{ik} \log(\phi_{ik}) + \tfrac{1}{2} \sum_{k=1}^{K} \log(s_k^2)
\end{aligned}
$$

References

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.