Advanced probabilistic methods

Lecture 6: Variational inference

Pekka Marttinen

Aalto University

March, 2019

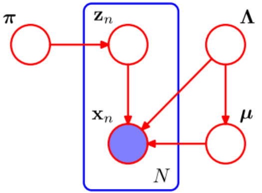
Lecture 6 overview

- Variational inference overview
- KL-divergence
- Mean-field variational inference
- Simple example using variational inference
- Suggested reading: Bishop: *Pattern Recognition and Machine Learning*
  - p. 461-474
  - *simple_vb_example.pdf* for the derivation of the VB updates for a simple GMM.
  - The general VB formulation for GMMs p. 474-486 (optional)

Approximate inference

- A central task in probabilistic modeling is to evaluate the posterior distribution
$$p(Z|X)$$
of latent variables Z given the observed variables X.
- In a fully Bayesian model, model parameters θ may be given priors and included as part of Z (unlike in the EM).

  - Often, computation of $p(Z|X)$ may not be possible in a closed form, and approximations are needed
    - variational inference (today)
    - stochastic variational inference (later)
    - sampling (→Bayesian data analysis)



Variational inference

- **Idea:** Approximate the posterior distribution of unknowns $p(Z|X)$ with a tractable distribution $q(Z)$.
- For example, $q(Z)$ may be assumed to have a simple form, e.g., Gaussian, or to factorize in a certain way.
- For the GMM, it would be sufficient to assume

$$q(\mathbf{z}, \pi, \Lambda, \mu) = q(\mathbf{z})q(\pi, \Lambda, \mu)$$

Basis of variational inference

- When $q(\mathbf{z})$ is an approximation for $p(\mathbf{z}|\mathbf{x})$, it is always true that

$$\log p(\mathbf{x}) = \mathcal{L}(q) + KL(q||p),$$

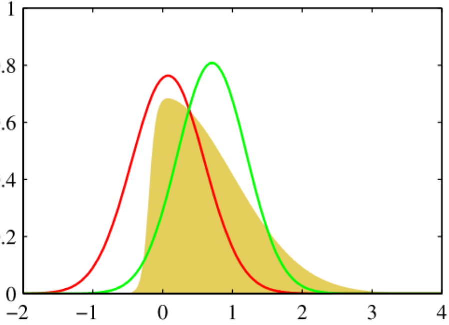
where

$$\mathcal{L}(q) = \int q(\mathbf{z}) \log \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} d\mathbf{z} \quad \text{(lower bound for } \log p(\mathbf{x}))$$

$$KL(q||p) = -\int q(\mathbf{z}) \log \left\{ \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} \right\} d\mathbf{z} \quad \text{(KL-divergence btw } q \text{ and } p).$$

- **Goal**: to maximize $\mathcal{L}(q)$ or, equivalently, to minimize the $KL(q||p)$.
- Note: $\mathcal{L}(q)$ is also called the 'ELBO' (evidence lower bound)

Variational Gaussian approximation



- Figure shows approximation of the original distribution (yellow) with a Gaussian at the mode (red, Laplace) or with a Gaussian that minimizes the KL-divergence (green).

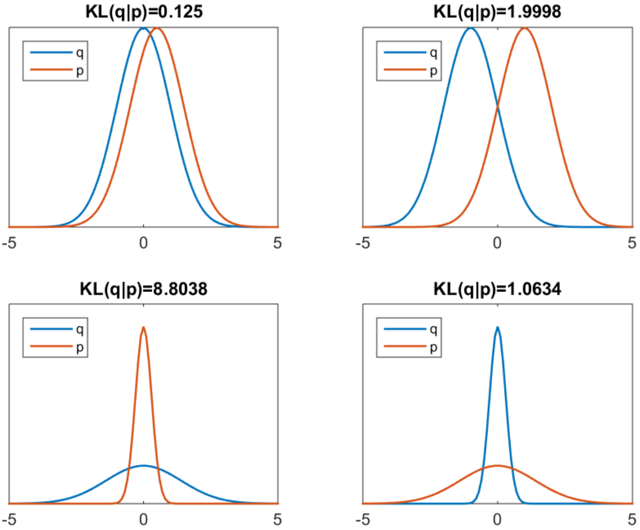
Kullback-Leibler divergence

- **KL-divergence**. For two distributions $q(x)$ and $p(x)$

$$KL(q|p) \equiv \int_x q(x) \log \frac{q(x)}{p(x)} dx$$

  - $KL(q|p) \geq 0$ (follows from Jensen's inequality)
  - $KL(q|p) = 0$ if and only if $q = p$
  - KL-divergence between q and p can be thought of as a 'distance' of p from q. However, $KL(q|p) \neq KL(p|q)$. Hence it's rather called 'divergence'.

Kullback-Leibler divergence - Example



Mean-field variational Bayes

- **Mean-field variational Bayes**: assume that the approximating distribution q factorizes according to M disjoint groups of $\mathbf{z}$

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i)$$

- Distributions $q(\mathbf{z}_i)$ are called **factors**
- NB: above $\mathbf{z}$ is a generic notation for all unobserved variables in the model, and comprises both parameters (e.g. π, Λ, μ in a GMM) and latent variables (e.g. cluster labels $\mathbf{z}$ in a GMM!)
- For example, assuming:

$$q(\mathbf{z}, \pi, \Lambda, \mu) = q(\mathbf{z})q(\pi, \Lambda, \mu)$$

leads to a tractable solution for the posterior $p(\mathbf{z}, \pi, \Lambda, \mu|\mathbf{x})$ of a GMM.

Mean-field variational Bayes updates

- Assume some current values for all factors $q_i(\mathbf{z}_i)$
- It can be shown (p. 465-466) that by keeping other factors $q_i(z_i)$ fixed for $i \neq j$, the lower bound $\mathcal{L}(q)$ of $\log p(\mathbf{x})$ can be mazimized (or $KL(q||p)$ minimized) by updating factor $q_j(\mathbf{z}_j)$ using

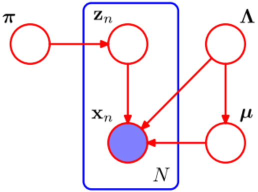
$$\log q_j^*(\mathbf{z}_j) = E_{q(\mathbf{z}_{\setminus j})}[\log p(\mathbf{x}, \mathbf{z})] + \text{const.}$$

- Here $q(\mathbf{z}_{\setminus j})$ is a short-hand for $\prod_{i \neq j} q_i(\mathbf{z}_i)$
- **Important formula**, as it forms the basis of deriving VB algorithms using factorized distributions
- **Algorithm**: update each factor in turn until convergence

Mean-field VB in practice (1/2)

- Assume a factorization, e.g., $q(\mathbf{z}, \pi, \Lambda, \mu) = q(\mathbf{z})q(\pi)q(\Lambda, \mu)$
- Write the log of the joint distribution

$$\log p(\mathbf{x}, \mathbf{z}, \mu, \Lambda, \pi) = \log p(\mathbf{x}|\mathbf{z}, \Lambda, \mu) + \log p(\mu|\Lambda) \\ + \log p(z|\pi) + \log p(\Lambda) + \log p(\pi)$$



Mean-field VB in practice (2/2)

- When updating a certain factor, for example $q(\mathbf{z})$, we identify terms in the log of the joint distribution that depend on $\mathbf{z}$, and compute their expectation over other unobserved variables

$$\begin{aligned}\log q^*(\mathbf{z}) &= E_{q(\pi)q(\Lambda,\mu)}\left[\log p(\mathbf{x},\mathbf{z},\mu,\Lambda,\pi)\right] + \text{const} \\ &= E_{q(\Lambda,\mu)}\left[\log p(\mathbf{x}|\mathbf{z},\Lambda,\mu)\right] + E_{q(\pi)}\left[\log p(\mathbf{z}|\pi)\right] + \text{const}\end{aligned}$$

- Finally, we exponentiate and normalize to give the updated $q^*(\mathbf{z})$

$$q^*(\mathbf{z}) = \frac{\exp\left(E_{\pi,\Lambda,\mu}\left[\log p(\mathbf{x},\mathbf{z},\mu,\Lambda,\pi)\right]\right)}{\int \exp\left(E_{\pi,\Lambda,\mu}\left[\log p(\mathbf{x},\mathbf{z},\mu,\Lambda,\pi)\right]\right) d\mathbf{z}}$$

  If conjugate priors are used, this belongs to the same family as the prior.

- Notation: instead of $E_{q(\pi,\Lambda,\mu)}$ we may simply use $E_{\pi,\Lambda,\mu}$ or just E.

Idea of derivation of the mean-field VB update*

- Assume just two hidden variables z_1 and z_2 and $q(z_1, z_2) = q_1(z_1)q_2(z_2)$. Then

$$\mathcal{L}(q) = \int q(\mathbf{z}) \log \frac{p(x, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} = \int q_1(z_1)q_2(z_2) \log \frac{p(x, z_1, z_2)}{q_1(z_1)q_2(z_2)} dz_1 dz_2$$
$$= \cdots = \int q_1(z_1) \log \frac{\widetilde{p}(x, z_1)}{q_1(z_1)} dz_1 + \text{const} = -KL(q_1, \widetilde{p}) + \text{const},$$

where $\widetilde{p}(x, z_1)$ is a distribution defined by

$$\log \widetilde{p}(x, z_1) = E_{q_2(z_2)}[\log p(x, z_1, z_2)] + \text{const}.$$

- We see that $\mathcal{L}(q)$ is maximized w.r.t. to q_1 when $KL(q_1, \widetilde{p})$ is minimized, i.e. when

$$q_1(z_1) = \widetilde{p}(x, z_1).$$

Simple example

- Model: assume that we have observations $\mathbf{x} = (x_1, \ldots, x_N)$ s.t.

$$p(x_n|\theta, \tau) = (1-\tau)N(x_n|0,1) + \tau N(x_n|\theta, 1)$$

  Prior:

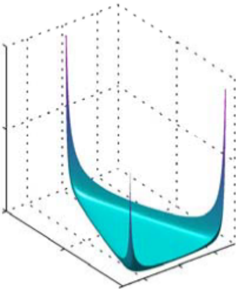
$$\tau \sim Beta(\alpha_0, \alpha_0) \qquad \theta \sim N(0, \beta_0^{-1})$$

  Formulation using latent variables $\mathbf{z} = (z_1, \ldots, z_n)$:

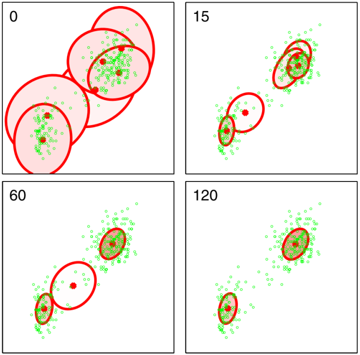
$$p(\mathbf{z}|\tau) = \prod_{n=1}^{N} \tau^{z_{n2}} (1-\tau)^{z_{n1}}$$
$$p(\mathbf{x}|\mathbf{z}, \theta) = \prod_{n=1}^{N} N(x_n|0,1)^{z_{n1}} N(x_n|\theta, 1)^{z_{n2}}$$

- *simple_vb_example.pdf*, and the next exercise.

Mean-field VB for the general GMM*

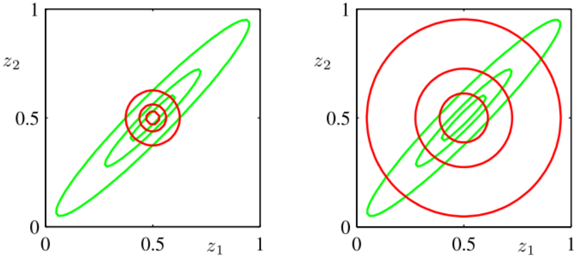


Bishop, Fig 2.5



- $Dirichlet(\pi|\alpha_0)$ prior on mixture coefficients with $\alpha_0 < 1$ favors **sparse solutions** →some components remain empty, with corresponding parameters μ_k, Λ_k following prior distributions
- Avoids overfitting and singularities present in the EM algorithm.

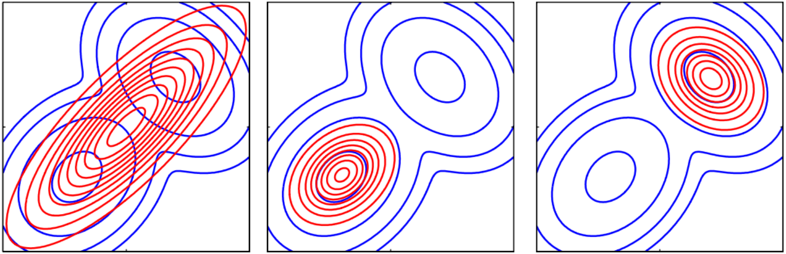
Properties of factorized approximations (1/2)



- Green: $p(\mathbf{z}|\mathbf{x})$, red: $q(\mathbf{z})$
- **Left**: q that minimizes $KL(q||p)$
- **Right**: q that minimizes $KL(p||q)$

→variational approximation (left) **underestimates uncertainty**.

Properties of factorized approximations (2/2)



- Blue: $p(\mathbf{z}|\mathbf{x})$, red: $q(\mathbf{z})$
- **Left**: q that minimizes $KL(p||q)$
- **Center**: q represents a local minimum of $KL(q||p)$
- **Right**: q represents another local minimum of $KL(q||p)$

$\rightarrow$variational approximation usually captures only a single mode.

Variational lower bound (ELBO)

- The derivation of the VB algorithm was based on minimizing $KL(q||p)$ in
$$\log p(\mathbf{x}) = \mathcal{L}(q) + KL(q||p)$$
- When conjugate priors and exponential family distributions are used, we can compute the variational lower bound $\mathcal{L}(q)$ directly
$$\mathcal{L}(q) = \int q(\mathbf{z}) \log \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} d\mathbf{z}$$
- Computing $\mathcal{L}(q)$ gives:
    1. alternative way to define the factor updates by maximizing $\mathcal{L}(q)$.
    2. simple check of the algorithm - $\mathcal{L}(q)$ should never decrease.
    3. criterion to monitor convergence.
    4. an estimate of $\log p(x)$ to be used in model selection

Computing the lower bound in practice

- For the GMM

$$\begin{aligned}\mathcal{L} =& E\left[\log p(\mathbf{x}, \mathbf{z}, \pi, \mu, \Lambda)\right] - E\left[\log q(\mathbf{z}, \pi, \mu, \Lambda)\right] \\ =& E\left[\log p(\mathbf{x}|\mathbf{z}, \pi, \mu, \Lambda)\right] + E\left[\log p(\mathbf{z}|\pi)\right] \\ &+ E\left[\log p(\pi)\right] + E\left[\log p(\mu, \Lambda)\right] \\ &- E\left[\log q(\mathbf{z})\right] - E\left[\log q(\pi)\right] - E\left[\log q(\mu, \Lambda)\right],\end{aligned}$$

where we have used

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}, \pi, \mu, \Lambda) &= p(\mathbf{x}|\mathbf{z}, \pi, \mu, \Lambda)p(\mathbf{z}|\pi)p(\pi)p(\mu, \Lambda) \quad \text{and} \\ q(\mathbf{z}, \pi, \mu, \Lambda) &= q(\mathbf{z})q(\pi)q(\mu, \Lambda)\end{aligned}$$

- All of these can be computed in a closed form.

Important points

- Variational Bayes aims to find a tractable approximation $q(\mathbf{z})$ for the posterior distribution $p(\mathbf{z}|\mathbf{x})$.
- $q(\mathbf{z})$ is found by maximizing the ELBO $\mathcal{L}(q)$ or, equivalently, by minimizing $KL(q||p)$.
- Mean-field VB: if $q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i)$, factor $q_j(\mathbf{z}_j)$ can be updated using

$$\log q_j^*(\mathbf{z}_j) = E_{q(\mathbf{z}_{\setminus j})}\left[\log p(\mathbf{x}, \mathbf{z})\right] + \text{const.}$$

- Variational approximation for a fully Bayesian model with prior distributions avoids some of the problems related to the ML estimation of the GMM (overfitting, singularities).