4. Deterministic approximation methods in Bayesian statistics

Laplace method, Variational Bayes compared to Expectation maximization

Sun Woo Lim

Mar 19, 2022

Basics: Entropy, Cross Entropy, KL Divergence, and ELBO

Reference : https://www.youtube.com/watch?v=ErfnhcEV1O8

All comes from a paper "A mathematical theory of communication", Claude E. Shannon, 1948.

We want useful information to communicate with each other as not all information is useful.

Bits

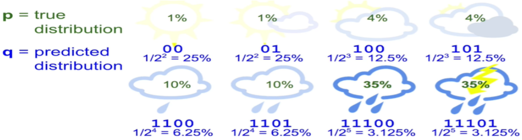
- In digital era, messages are composed of **bits** = 0 or 1.
- In Shannon's theory, a message of 1 bit **reduces the recipient's uncertainty by a factor of 2**.
  e.g) Sunny w.p 0.5 & rainy w.p 0.5. When weather forecast tells "rainy", the message has $1 = -log_2(0.5)$ bit of info.
  e.g) 8 possible (& equally likely) states and weather forecast tells "sunny and cloudy", $3 = -log_2(0.125)$ bits of info.

Entropy $H(p) := E_{x\sim p}[\frac{1}{log(p(x))}] = \int_x p(x)\frac{1}{logp(x)}dx = -\int_x p(x)logp(x)dx.$

- e.g)Sunny w.p $\frac{3}{4}$ and rainy w.p. $\frac{1}{4}$. "sunny": 0.41 bit of info, "rainy": 2 bits of info.
  On average, $H(p) = 0.75 \cdot 0.41 + 0.25 \cdot 2 = 0.81$
- Interpretation) **Expected message length = amount of info per data for given pmf/pdf** p **= how unpredictable** p **is.**
- Note) The log with base e is used more frequently than the base 2 although base 2 fits the definition.
- Facts) Uniform distribution has maximum entropy → usage in nonparametric statistics.

Cross Entropy of distribution Q relative to distribution P over the same domain χ

- $H(p,q) := E_{x\sim p}[\frac{1}{logq(x)}] = -\int_x p(x)logq(x)dx$
- Interpretation) **Expected message length per data assuming wrong distribution Q, while true distribution is P.**
- If $q = p$, $H(p,q) = H(q) = H(p)$.



Kullback–Leibler divergence (= Relative Entropy) from Q to P over the same domain χ

- $KL(p||q) := H(p,q) - H(p) = -\int_x p(x)logq(x)dx - [-\int_x p(x)logp(x)dx] = \int_x p(x)log\frac{p(x)}{q(x)}dx.$
- Interpretation) **Expected surplus of message length (=surprise)** of modeling the true distribution P as Q.
- **Properties**
  1. $KL(p||q) = 0$ iff $p = q$ almost everywhere.
  2. $KL(p||q) \geq 0$: "Non-negativity". Proven by Jensen's inequality.
  3. $KL(p||q) \neq KL(q||p)$: "Asymmetry" : disqualifies KL as a "metric".
  4. Triangle inequality not satisfied : disqualifies KL as a "metric".

Diagnostic Question When is cross entropy minimized?

A) When $q = p$. Understand both intuitively, and relating to the KL divergence!

Forward or reverse KL : When the true distribution of x is p and false (or, approximate) distribution is q,
Note, the point of following analysis is that p is fixed and q is not. Intuitively makes sense.

- **Forward KL = M-projection = moment projection** $KL(p||q) = \int_x p(x) log \frac{p(x)}{q(x)} dx$
  - Goes to ∞ when $q(x) \to 0$ and $p(x) > 0$. So, if $p(x) > 0$, $q(x)$ must > 0 to avoid $KL(p||q) \neq 0$.
  - Zero avoiding for q. Thus, q will overestimate support of p. Why? Think of def'n of support.
  - Intuitively makes sense to find q minimizing $KL(p||q)$ but bad at finding mode if p were multimodal.

- **Reverse KL = I-projection = information projection** $KL(q||p) = \int_x q(x) log \frac{q(x)}{p(x)} dx$
  - Goes to ∞ when $p(x) \to 0$ and $q(x) > 0$. So, if $p(x) \to 0$, $q(x)$ must $= 0$ to avoid $KL(q||p) \neq 0$.
  - Zero forcing for q. Thus, q will underestimate support of p.
  - Might neglect some other modes if p were multimodal.

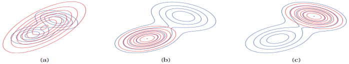


Figure 21.1 Illustrating forwards vs reverse KL on a bimodal distribution. The blue curves are the contours of the true distribution p. The red curves are the contours of the unimodal approximation q. (a) Minimizing forwards KL: q tends to "cover" p. (b-c) Minimizing reverse KL: q locks on to one of the two modes. Based on Figure 10.3 of (Bishop 2006b). Figure generated by KLfwdReverseMixGauss.

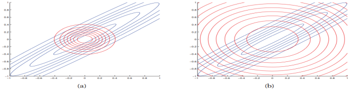


Figure 21.2 Illustrating forwards vs reverse KL on a symmetric Gaussian. The blue curves are the contours of the true distribution p. The red curves are the contours of a factorized approximation q. (a) Minimizing $\mathbb{KL}(q||p)$. (b) Minimizing $\mathbb{KL}(p||q)$. Based on Figure 10.2 of (Bishop 2006b). Figure generated by KLpqGauss.

Figure: From Murphy (2012)

Def) Evidence Lower Bound (ELBO) $L(q(\theta)) := E_{\theta\sim q}[log(\frac{p(\mathcal{D},\theta)}{q(\theta)})]$ for $p(\mathcal{D},\theta)$: joint dist'n of $\mathcal{D}$ & θ , $q(\theta)$: any pdf of θ.

Fact 1) log evidence (constant w.r.t. θ) = ELBO (varying w.r.t. θ) + KL from p to q (varying w.r.t. θ)

$KL(q(\theta)||p(\theta|\mathcal{D})) = E_{\theta\sim q}[log\frac{q(\theta)}{p(\theta|\mathcal{D})}]$

$= E_{\theta\sim q}logq(\theta) - E_{\theta\sim q}[logp(\theta|\mathcal{D})]$

$= E_{\theta\sim q}logq(\theta) - (E_{\theta\sim q}[logp(\theta,\mathcal{D}) - logp(\mathcal{D})])$

$= -E_{\theta\sim q}[logp(\theta,\mathcal{D})] - E_{\theta\sim q}[logq(\theta)] + logp(\mathcal{D})$

$= logp(\mathcal{D}) - E_{\theta\sim q}[log(\frac{p(\mathcal{D},\theta)}{q(\theta)})] = logp(\mathcal{D}) - L(q(\theta))$: KL = log evidence - ELBO $\leftrightarrow$ log evidence = ELBO + KL

Fact 2) ELBO inequality: shows that Evidence Lower Bound truly is lower bound of (log) evidence

1. proof method 1: Use Fact 1 + positiveness of KL divergence.
2. proof method 2: $log(p(\mathcal{D})) = log(\int_\theta p(\mathcal{D},\theta)d\theta)$

   $= log(\int_\mathcal{D} p(\mathcal{D},\theta)\frac{q(\theta)}{q(\theta)}d\mathcal{D}) = log(E_q[\frac{p(\mathcal{D},\theta)}{q(\theta)}])$ by applying definition of expectation with measure q.

   $\geq E_q[log(\frac{p(\mathcal{D},\theta)}{q(\theta)})] = L(q)$ by Jensen's inequality.

Note) ELBO inequality becomes inequality when $KL(q(\theta)||p(\theta|\mathcal{D})) = 0$

Note) The reverse KL, not the forward KL! Be careful.

Sampling Method vs Deterministic Approximation

Sampling Methods

- Obtain independent / dependent samples from a target distribution
- Samples from $p(\theta|\mathcal{D})$ in Bayesian Statistics. In Metropolis-Hastings, uses unnormalized density $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$.
- Unbiased but slow performance for large dimensions ($=$ not scalable)

Deterministic Approximation Methods

- Obtain approximate functional form of the target distribution. $p(\theta|\mathcal{D}) \approx q(\theta)$ ($\in C$: restricted function class)
- Modal Approximation (Laplace method), distributional approximation (Variational Bayes / Expectation Propagation)
- This slide deals with Laplace method and Variational Bayes.
  Other concepts not dealt include expectation propagation and Approximate Bayesian Computation.
- Scalable ($=$ not terrible in high-dimension) and biased solution.

Laplace Method

Idea: Laplace method approximates the posterior by 2nd order Taylor approximation at $\theta = \theta_{MAP} := argmax_\theta p(\theta|\mathcal{D})$.

More specifically, approximates log posterior $logp(\theta|\mathcal{D})$ by 2nd order Taylor approximation at $\theta = \theta_{MAP}$.

$$\widehat{log\ p(\theta|\mathcal{D})} = log\ p(\theta_{MAP}|\mathcal{D}) + (\nabla log\ p(\theta_{MAP}|\mathcal{D})^T(\theta - \theta_{MAP}) + \frac{1}{2}(\theta - \theta_{MAP})^T[\frac{d^2 log\ p(\theta|\mathcal{D})}{d\theta^2}_{\theta=\theta_{MAP}}](\theta - \theta_{MAP}).$$

$$\widehat{log\ p(\theta|\mathcal{D})} = log\ p(\theta_{MAP}|\mathcal{D}) + \frac{1}{2}(\theta - \theta_{MAP})^T[\frac{d^2 log\ p(\theta|\mathcal{D})}{d\theta^2}_{\theta=\theta_{MAP}}](\theta - \theta_{MAP}) \because \nabla log\ p(\theta|\mathcal{D})|_{\theta=\theta_{MAP}} = 0.$$

$$\rightarrow \widehat{p(\theta|\mathcal{D})} = p(\theta_{MAP}|\mathcal{D}) \times exp[-\frac{1}{2}(\theta - \theta_{MAP})^T[-\frac{d^2 log\ p(\theta|\mathcal{D})}{d\theta^2}_{\theta=\theta_{MAP}}](\theta - \theta_{MAP})]$$

$$\propto exp[-\frac{1}{2}(\theta - \theta_{MAP})^T[-\frac{d^2 log\ p(\theta|\mathcal{D})}{d\theta^2}_{\theta=\theta_{MAP}}](\theta - \theta_{MAP})]$$

$\widehat{p(\theta|\mathcal{D})} = dMVN(\theta,\ mean = \theta_{MAP},\ \Sigma = [-\frac{d^2 logp(\theta|\mathcal{D})}{d\theta^2}|_{\theta=\theta_{MAP}}]^{-1})$ using MVN fact!

Diagnostic Questions

- Q) Why of all $\theta = \theta_{MAP}$? A) θ_{MAP} is the point where $\nabla logp(\theta|\mathcal{D}) = 0$, advances to the **normal approximation**!
- Q) How can I calculate θ_{MAP}? A) Newton-Rhapson method / stepwise ascent / EM Algorithm
- Q) Then, is the posterior the normal distribution? A) No, Taylor polynomial truncated up to order two.

Expectation Maximization (EM) Algorithm

Situation : Known, observed data $X = x$, latent variables Z and unknown and **fixed** parameter θ.

Explanation of Z:

1. Really missing observation (missing at random, missing not at random, etc). e.g) Truncation
2. Model formulation is better by assuming latent variable
   1. Example 1) Gaussian Mixture Model (GMM): latent variable = group identifier
      - Data: $X = (\vec{x}_1, ..., \vec{x}_n), \vec{x}_i \in \mathbb{R}^d$: n observations from a mixture of k MVN_d distributions
      - Latent variable: $Z = (\vec{z}_1, ..., \vec{z}_n)$, $\vec{z}_i \in \{1, ..., d\}$: latent variable concerning the component = group of each x_i
      - Parameters
        1. $\vec{\tau} = (\tau_1, ..., \tau_d)$ where $\tau_j := P(Z_i = j), \forall i \in \{1, ..., n\}$. $j \in \{1, ..., d\}$.
        2. $(\vec{\mu}_1, ..., \vec{\mu}_d)$ where $\vec{\mu}_j$: mean vector of j^{th} Gaussian component.
        3. $(\Sigma_1, ..., \Sigma_d)$ where Σ_j: covariance matrix of j^{th} Gaussian component.
   2. Example 2) K Means Clustering: similar setting!
      - Data : $X = (\vec{x}_1, ..., \vec{x}_n), \vec{x}_i \in \mathbb{R}^d$: n observations
      - Latent variable: (r_{ij}), $i \in \{1, ..., n\}, j \in \{1, ..., d\}$. $r_{ij} = I[i^{th}\ data\ \in\ Group\ j]$
      - Parameters : Centroids $(\vec{\mu}_1, ..., \vec{\mu}_d)$, $\mu_j \in \mathbb{R}^d$

      Solve $min_{\mu_1,...,\mu_d,r_{11},...,r_{nd}} \sum_{i=1}^{n} \sum_{j=1}^{d} r_{ij} ||\vec{x}_i - \mu_j||_2^2$: total square distance from each point to its centroid.

Goal of EM : Obtain **MLE** of $\theta = argmax_\theta L(\theta)$ where $L(\theta) = p(x|\theta) = \int_z p(x, z|\theta)dz = \int_z p(x|z, \theta)p(z|\theta)dz$

Hardship: With latent variables Z, usually impossible $\because$ 1) z unobserved, 2) $p(z|\theta)$ unknown without knowledge of θ.

EM Idea: Find the argmax of log marginal likelihood of θ by repeatedly in a way avoiding above issue .

1) finding a function that minorizes $l(\theta; x)$ (**E-step**) and 2) finding the maximum of that function (**M-step**)
(Q "log" likelihood? No problem? A) I am concerned with argmax, and log is monotone increasing ftn)

Iterative representation of EM Algorithm

- E-step: Calculate $Q(\theta|\theta_{(t)}) := E_{Z|X,\theta_{(t)}} l(\theta; X, Z) = \int_z [log p_{X,Z}(x, z|\theta) \cdot p_{Z|X}(z|x, \theta_{(t)})]dz$

  Proof of minorization : hint) use Jensen's inequality

- M-step: Calculate $\theta_{(t+1)} = argmax_\theta Q(\theta|\theta_{(t)})$

Ex) EM for GMM (suppose number of cluster = 2 for simplicity and dimension = 2 for visualization)

Setting

1. Data: $X = (\vec{x}_1, ..., \vec{x}_n), \vec{x}_i \in \mathbb{R}^2$: n observations from a mixture of 2 MVN_2 distributions.
2. Latent variable: $Z = (\vec{z}_1, ..., \vec{z}_n)$, $\vec{z}_i \in \{1, , 2\}$: latent variable concerning the component = group of each x_i
3. Parameters: $\Theta := (\vec{\tau} = (\tau_1, \tau_2), \mu_1, \mu_2, \Sigma_1, \Sigma_2)$ where $\tau_1 := P(Z_i = 1), \tau_2 = 1 - \tau_1$.

Comparison of incomplete likelihood and complete likelihood

- Incomplete Likelihood $L(\theta; x) = \prod_{i=1}^n \sum_{j=1}^2 \tau_j f(x_i; \mu_j, \Sigma_j) = \prod_{i=1}^n [\tau_1 f(x_i; \mu_1, \Sigma_1) + (1 - \tau_1) f(x_i; \mu_2, \Sigma_2)]$
- Complete Likelihood: $L(\theta; x, z) = \prod_{i=1}^n \prod_{j=1}^2 [\tau_j f(x_i; \mu_j, \Sigma_j)]^{I(z_i = j)}$
  $exp[\sum_{i=1}^n \sum_{j=1}^2 I(z_i = j)[-\frac{d}{2}log(2\pi) + log\tau_j - \frac{1}{2}log|\Sigma_j| - \frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)]]$

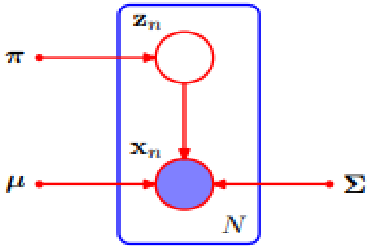


Figure: Graphical Notation of Frequentist EM GMM (Source: Bishop & Nasrabadi, 2006)

E Step

$Q(\theta|\theta^{(t)}) = E_{Z|X,\theta^{(t)}} l(\theta; X, Z) = E_{Z|X,\theta^{(t)}} log \prod_{i=1}^{n} L(\theta; x_i, Z_i) = E_{Z|X,\theta^{(t)}} \sum_{i=1}^{n} l(\theta; x_i, Z_i)$

$= \sum_{i=1}^{n} \sum_{j=1}^{2} \{l(\theta; x_i, Z_i) \times Pr(Z_i = j|X_i = x_i; \theta^{(t)})\}$.

Denote $P_{i,j}^{(t)} := Pr(Z_i = j|X_i = x_i; \theta^{(t)}) = \frac{\tau_j^{(t)} f(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{j=1}^{2} \tau_j^{(t)} f(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})} = \frac{\tau_j^{(t)} f(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}{\tau_1^{(t)} f(x_i; \mu_1^{(t)}, \Sigma_1^{(t)}) + \tau_2^{(t)} f(x_i; \mu_2^{(t)}, \Sigma_2^{(t)})}$

Then, $Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{2} P_{i,j}^{(t)} [-\frac{d}{2} log(2\pi) + log\tau_j - \frac{1}{2} log|\Sigma_j| - \frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)]$

M step : Θ consists of five sub parameters to optimize: $\vec{\tau}, \mu_1, \Sigma_1, \mu_2, \Sigma_2$. Regarding τ, notice $\tau_2 = 1 - \tau_1$.

Note) $Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{2} P_{i,j}^{(t)} [-\frac{d}{2} log(2\pi) + log\tau_j - \frac{1}{2} log|\Sigma_j| - \frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)]$

1. $\tau_1^{(t+1)} = argmax_{\tau_1} Q(\theta|\theta^{(t)}) = argmax_{\tau_1} \sum_{i=1}^{n} \sum_{j=1}^{2} P_{i,j}^{(t)} log\tau_j =$

   $argmax_{\tau_1}(log\tau_1 \cdot \sum_{i=1}^{n} P_{i,1}^{(t)} + log(1-\tau_1) \cdot \sum_{i=1}^{n} P_{i,2}^{(t)} = \frac{\sum_{i=1}^{n} P_{i,1}^{(t)}}{\sum_{i=1}^{n}(P_{i,1}^{(t)} + P_{i,2}^{(t)})} = \frac{1}{n} \sum_{i=1}^{n} P_{i,1}^{(t)}$

2. $(\mu_1^{(t+1)}, \Sigma_1^{(t+1)}) = argmax_{\mu_1, \Sigma_1} Q(\theta|\theta^{(t)}) = argmax_{\mu_1} \sum_{i=1}^{n} P_{i,1}^{(t)} [-\frac{1}{2} log|\Sigma_1| - \frac{1}{2}(x_i - \mu_1)^T \Sigma_1^{-1}(x_i - \mu_1)]$

   $= (\frac{\sum_{i=1}^{n} P_{i,1}^{(t)} x_i}{\sum_{i=1}^{n} P_{i,1}^{(t)}}, \frac{\sum_{i=1}^{n} P_{i,1}^{(t)}(x_i - \mu_1^{(t+1)})(x_i - \mu_1^{(t+1)})^T}{\sum_{i=1}^{n} P_{i,1}^{(t)}})$

3. By symmetry, $\mu_2^{(t+1)} = \frac{\sum_{i=1}^{n} P_{i,2}^{(t)} x_i}{\sum_{i=1}^{n} P_{i,2}^{(t)}}$ and $\Sigma_2^{(t+1)} = \frac{\sum_{i=1}^{n} P_{i,2}^{(t)}(x_i - \mu_2^{(t+1)})(x_i - \mu_2^{(t+1)})^T}{\sum_{i=1}^{n} P_{i,2}^{(t)}}$

Variational Bayes

Situation : Known, observed data $X = x$, latent variables Z **including parameter (random vector)** θ.

Explanation of Z:

- includes the explanation about latent variable in p7.
  Missing data + latent variable that affects the **data generation process**
- In Bayesian approach, θ is also a **random variable** which is latent $\rightarrow Z$ has to incorporate θ
- Example 1) Gaussian Mixture Model (GMM)
  - Data: $X = (\vec{x}_1, ..., \vec{x}_n), \vec{x}_i \in \mathbb{R}^d$: n observations from a mixture of k MVN_d distributions
  - Latent variable including parameter: $Z, \vec{\tau}, \mu_1, ..., \mu_d, \Sigma_1, ..., \Sigma_d$.

Notation: In the Variational Bayes, Z incorporates θ. However, I conform to the usual notation of Bayesian statistics and write the R.V parameters as θ.

VB Idea: Obtain approximated density $q(\theta)$ minimizing the **reverse KL divergence** $KL(q(\theta)||p(\theta|\mathcal{D}))$.
: $q(\theta) = argmin_q KL(q||p) = argmin_q - E_q(log(\frac{p(\theta|\mathcal{D})}{q(\theta)})) = argmin_q - \int_q log(\frac{p(\theta|\mathcal{D})}{q(\theta)})q(\theta)d\theta$.

Note) Reverse KL, not forward! Why? in multimodal p, reverse KL easier to compute and more sensible statistically.
(Reference: p733, Murphy (2012))

Two main hardships and solutions

1. Q)How find pdf q minimizing $KL(q(\vec{\theta})||p(\vec{\theta}|\mathcal{D}))$ when $p(\vec{\theta}|\mathcal{D})$ is not known?

   Intuitively, how do I find a way to the target where I do not know the target?

   A) **log evidence (constant) = ELBO (varying) + KL (varying)**.

   Instead of impossible task of **directly minimizing KL**, detour by **maximizing ELBO** that is possible.

2. Q) How to find $q(\vec{\theta}) = argmin_q KL(q(\vec{\theta})||p(\vec{\theta}|\mathcal{D}))$ where q is a function? Function optimization is hard.

   A) Assume restricted, simple (but, preserving dimension of $\vec{\theta}$) functional form.
   I.O.W, assume $q \in C$, C: restricted class of functions.

   1. **Mean Field Approximation** $q(\vec{\theta}) = \prod_{j=1}^{m} q_j(\theta_j)$ where $\vec{\theta} = (\theta_1, \ldots, \theta_m)$. θ_j might be vector valued.

      - Assuming that each parameter component is independent.

      - For $q_j(\theta_j)$, no restricted form. **Individual function optimization problem**.

   2. **Parametric Approximation** $q(\vec{\theta}) = q(\vec{\theta}|\vec{\phi})$ with hyperparameter $\vec{\phi}$. **Converts ftn optimization to parametric optimization quest**.

      - Initial guess of ϕ and iteratively update $\vec{\phi}$ using EM-like method that decreases KL.

Note) Can use both methods also: $q(\theta) = \prod_j g(\theta_j|\phi_j)$, which we will focus on today.

Hardship 1. How minimize $KL(g(\theta)||p(\theta|\mathcal{D}))$ when $p(\theta|\mathcal{D})$ is not known?

Setting: $\mathcal{D}$: observed data, θ: latent variable including parameter. $p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})}$ is intractable due to the normalizing constant $p(\mathcal{D})$ while $p(\theta, \mathcal{D})$ is tractable!

Then, refer to the facts about **ELBO** and maximize **ELBO**, which equivalent to minimizing **KL**, but is possible.

I.O.W, $q = argmin_q KL(q(\theta)||p(\theta||\mathcal{D}))$

$= argmax_q L(q(\theta)) = argmax_q E_q[log(\frac{p(\mathcal{D},\theta)}{q(\theta)})] = argmax_q \int_\theta log(\frac{p(\mathcal{D},\theta)}{q(\theta)})q(\theta)d\theta$.

Hardship 2. How to find $q(\theta) = argmin_q KL(q(\theta)||p(\theta|\mathcal{D}))$ where q is a (density) function ?

By solution of issue 1, $q(\theta) = argmin_{q\in C} KL(q(\theta)||p(\theta|\mathcal{D})) = argmax_{q\in C} L(q(\theta))$ where $L(q(\theta)) = E_q[log\frac{p(x,\theta)}{q(\theta)}]$

But, q is still a function and "restricted function class" C should be 'really' restricted.

Mean Field Approximation $q(\vec{\theta}) = \prod_{j=1}^{m} q_j(\theta_j)$ where $\vec{\theta} = (\theta_1, \dots, \theta_m)$.

How: Block Coordinate Ascent

Since $q(\vec{\theta})$ is divided by m different function components, fix all other $\{q_{i\neq j}\}$ and optimize q_j : $argmax_{q_j} L(q(\vec{\theta}))$.

$L(q(\vec{\theta})) = E_q[log\frac{p(\mathcal{D},\vec{\theta})}{q(\vec{\theta})}] = E_q logp(\mathcal{D},\vec{\theta}) - E_q logq(\vec{\theta})$

$= E_q logp(\mathcal{D},\vec{\theta}) - \sum_{j=1}^{m} E_{q_j} logq_j(\theta_j)$. Note that q_j is a function $q_j(\theta_j)$.

$= E_{q_j(\theta_j)}[E_{q_{i\neq j}} logp(x,\theta)] - E_{q_j(\theta_j)} logq_j(\theta_j) + Const$: with respect to j^{th} component function.

$= E_{q_j(\theta_j)} log\frac{r_j(\theta_j)}{q_j(\theta_j)} + Const = -KL(q_j(\theta_j)||r_j(\theta_j)) + Const,$

where $r_j(\theta_j) := \frac{1}{c(Z_j)} exp(E_{q_{i\neq j}} logp(\mathcal{D},\theta))$. $c(Z_j)$ is a normalizing constant to make $r_j(\theta_j)$ be density.

Since $L(q(\vec{\theta})) = -KL(q_j(\theta_j)||r_j(\theta_j)) + Const$ is maximized w.r.t θ_j when $q_j(\theta_j) = r_j(\theta_j)$ (property of KL divergence),

$q_j^*(\theta_j) = r_j(\theta_j) = \frac{1}{c(Z_j)} exp(E_{q_{i\neq j}} logp(\mathcal{D},\theta))$

Mean Field Approximation Algorithm

1. Initialize $q(\theta) = \prod_{j=1}^{m} q_j(\theta_j)$ by assumption of Mean field.
2. Iterate the following until convergence (of ELBO)
    1. Update each function component $q_1, \dots, q_m$: $q_j^*(\theta_j) = r_j(\theta_j) = \frac{1}{c(Z_j)} exp(E_{q_{i\neq j}} logp(\mathcal{D},\theta))$
    2. Calculate ELBO: $L(q(\theta))$

Question) When applicable?

Answer) $logq_j^*(\theta_j) = E_{q_{i\neq j}} logp(\mathcal{D}, \theta) + Const$

I.O.W, $E_{q_{i\neq j}} logp(\mathcal{D}, \theta)$ should be analytically calculated.

Satisfied if **Semi conjugacy** of likelihood and prior on each component θ_j conditioned on all other components!

: $p(\theta_j|\theta_{i\neq j}) \in \mathcal{F} \rightarrow p(\theta_j|\mathcal{D}, \theta_{i\neq j}) \in \mathcal{F}$

Example) Joint posterior of univariate Gaussian (Murphy, 2012) : example of Mean Field Parametric approximation!

- **True distribution**: Assume **full conjugacy prior**: $\tilde{\sigma}^2 := \frac{1}{\sigma^2} \sim \Gamma(a_0, b_0)$, $\mu|\tilde{\sigma}^2 \sim N(\mu_0, \frac{1}{\tilde{\sigma}^2\kappa_0})$. $\vec{\theta} := (\mu, \tilde{\sigma}^2)$

  $\rightarrow$ Pedagogical example to see how close $q(\mu, \tilde{\sigma}^2)$ is to $p(\mu, \tilde{\sigma}^2|\mathcal{D})$ because $p(\mu, \tilde{\sigma}^2|\mathcal{D})$ is tractable here!

- For the approximated posterior $q(\mu, \tilde{\sigma}^2) \approx p(\mu, \tilde{\sigma}^2|\mathcal{D})$, use Mean Field $q(\mu, \tilde{\sigma}^2) = q(\mu)q(\tilde{\sigma}^2)$.

  Note) For the **Mean Field**, modify the method to handle a **semi-conjugate** prior
  $p(\mu, \tilde{\sigma}^2) = dnorm(\mu, \mu_0, \tau_0) \times dgamma(\tilde{\sigma}^2, a_0, b_0)$,
  which will make the inference approximate (using different prior setting from the true one!)

  Note) Both $q(\mu)$ and $q(\tilde{\sigma}^2)$ are pdf's and no need to specify parametric forms of each (will fall out automatically).

Log Unnormalized posterior $log\ p(\theta, \mathcal{D}) = logp(\mu, \tilde{\sigma}^2, \mathcal{D}) = logp(\tilde{\sigma}^2)p(\mu|\tilde{\sigma}^2)p(\mathcal{D}|\mu, \tilde{\sigma}^2)$
$= \frac{n}{2}log\tilde{\sigma}^2 - \frac{\tilde{\sigma}^2}{2}\sum_{i=1}^n (x_i - \mu)^2 - \frac{\kappa_0\tilde{\sigma}^2}{2}(\mu - \mu_0)^2 + \frac{1}{2}log(\kappa_0\tilde{\sigma}^2) + (a_0 - 1)log\tilde{\sigma}^2 - b_0\tilde{\sigma}^2 + Const$

Update $q_\mu(\mu)$. Since this is Parametric approximation, update parameters of $q_\mu(\mu)$
$logq_\mu(\mu) = E_{q_{\tilde{\sigma}^2}}[logp(\mathcal{D}|\mu, \tilde{\sigma}^2) + logp(\mu|\tilde{\sigma}^2)] + Const$ Q) What happened to $logp(\tilde{\sigma}^2)$?
$= -\frac{E_{q_{\tilde{\sigma}^2}}(\tilde{\sigma}^2)}{2}[\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^n (x_i - \mu)^2] + Const$
$\leftrightarrow q_\mu(\mu) \propto exp[-\frac{E_{q_{\tilde{\sigma}^2}}(\tilde{\sigma}^2)}{2}[\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^n (x_i - \mu)^2]]$
By normal distribution fact 5 + completing sum of squares, $q_\mu(\mu)$ is $N(\mu_{new}, \frac{1}{\kappa_{new}})$ with
$\mu_{new} = \frac{\kappa_0\mu_0 + \sum x_i}{\kappa_0 + n}$ and $\kappa_{new} = (\kappa_0 + n)E_{q_{\tilde{\sigma}^2}}(\tilde{\sigma}^2)$.
Since I do not know $E_{q_{\tilde{\sigma}^2}}(\tilde{\sigma}^2)$ because of not knowing $q_{\tilde{\sigma}^2}(\tilde{\sigma}^2)$, illustrated in the next page.

Update $q_{\tilde{\sigma}^2}(\tilde{\sigma}^2)$. Since this is Parametric approximation, update parameters of $q_{\tilde{\sigma}^2}(\tilde{\sigma}^2)$
$logq_{\tilde{\sigma}^2}(\tilde{\sigma}^2) = E_{q_\mu}[logp(\mathcal{D}|\mu, \tilde{\sigma}^2) + logp(\mu|\tilde{\sigma}^2) + logp(\tilde{\sigma}^2)] + Const$
Think of why here $logp(\tilde{\sigma}^2)$ is considered.
$= (a_0 - 1)log\tilde{\sigma}^2 - b_0\tilde{\sigma}^2 + \frac{1}{2}log\tilde{\sigma}^2 + \frac{n}{2}log\tilde{\sigma}^2$
$= -\frac{\tilde{\sigma}^2}{2}E_{q_\mu}[\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^n (x_i - \mu)^2] + Const$ is a log of gamma density.
$q_{\tilde{\sigma}^2}(\tilde{\sigma}^2) = dgamma(\tilde{\sigma}^2, a_{new}, b_{new})$ with $a_{new} = a_0 + \frac{n+1}{2}$, and $b_{new} = b_0 + \frac{E_{q_\mu}[\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^n (x_i - \mu)^2]}{2}$

Expectation calculation

$E_{q_\mu}(\mu) = \mu_{new}$ and $E_{q_\mu}(\mu^2) = \mu_{new}^2 + \frac{1}{\kappa_{new}}$ using $q(\mu) = dnorm(\mu, \mu_{new}, \frac{1}{\kappa_{new}})$

$E_{q_{\tilde{\sigma}^2}}(\tilde{\sigma}^2) = \frac{a_{new}}{b_{new}}$ using $q(\tilde{\sigma}^2) = dgamma(\tilde{\sigma}^2, a_{new}, b_{new})$

Final update equation with no unknown variables

$q(\mu) = dnorm(\mu, \mu_{new}, \frac{1}{\kappa_{new}})$

where $\mu_{new} = \frac{\kappa_0 \mu_0 + \sum x_i}{\kappa_0 + n}$ and $\kappa_{new} = (\kappa_0 + n)\frac{a_{new}}{b_{new}}$

$q(\tilde{\sigma}^2) = dgamma(\tilde{\sigma}^2, a_{new}, b_{new})$

where $a_{new} = a_0 + \frac{n+1}{2}$ and $b_{new} = b_0 + \kappa_0(E[\mu^2] + \mu_0^2 - 2E(\mu)\mu_0) + \frac{\sum_{i=1}^n (x_i^2 + E[\mu^2] - 2E[\mu]x_i)}{2}$

Note that parameters μ_{new} and a_{new} are fixed for all iterations, only update κ_{new} and b_{new}.

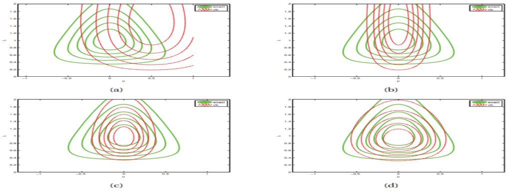


Figure 21.5 Factored variational approximation (red) to the Gaussian-Gamma distribution (green). (a) Initial guess. (b) After updating q_μ. (c) After updating q_λ. (d) At convergence (after 5 iterations). Based on 10.4 of (Bishop 2006b). Figure generated by unigaussVbDemo.

Figure: From Murphy, 2012

Variational Bayes EM

VB until now: Infer the **parameters** that are random variables and a concept of **latent variable** not existing.

VBEM : Bayesian method with both **latent variables** and **parameters**.

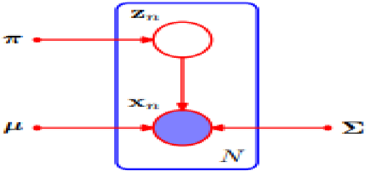
EM Algorithm

1. **Goal** : Obtain $\theta^* = argmax_\theta L(q(\theta))$ where $L(q(\theta)) = p(x|\theta) = \int_z p(x,z|\theta)dz = \int_z p(x|z,\theta)p(z|\theta)dz$
2. **E-step** : Calculate $Q(\theta|\theta_{(t)}) := E_{Z|X,\theta_{(t)}} l(\theta; X, Z) = \int_z [log p_{X,Z}(x,z|\theta) \cdot p_{Z|X}(z|x,\theta_{(t)})]dz.$
3. **M-step** : Calculate $\theta_{(t+1)} = argmax_\theta Q(\theta|\theta_{(t)})$

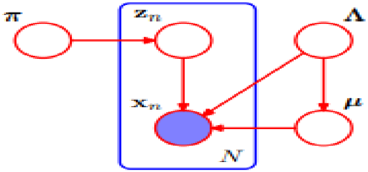
VBEM Algorithm

1. **Goal** : Obtain a function $q(\vec{\theta}, \vec{z})$ that is $\approx p(\vec{\theta}, \vec{z}|\mathcal{D})$.
   Specifically, **maximize ELBO** $\mathcal{L}(q(\vec{\theta}, \vec{z})) = \int_z \int_\theta q(\vec{\theta}, \vec{z}) log \frac{p(\mathcal{D}, \vec{\theta}, \vec{z})}{q(\vec{z}, \vec{\theta})} \leftrightarrow$ **minimizing reverse KL** $KL(q(\vec{\theta}, \vec{z})||p(\vec{\theta}, \vec{z}|\mathcal{D}).$
2. **Basic setting to make model easier** : $q(\vec{\theta}, \vec{z}) = q(\vec{\theta})q(\vec{z}) = q(\vec{\theta}) \prod_i q(\vec{z}_i)$. First equality comes from **Mean field assumption** and second equality is that latent variables are iid conditional on $\vec{\theta}$.
3. **Variational E-step** : Update $q(\vec{z}_i)$, or think of this as $q(\vec{z}_i|\mathcal{D}, \bar{\vec{\theta}})$: similar to E-step except that here we use posterior mean, not MAP.
4. **Variational V-step** : Update $q(\vec{\theta})$. Compared to M-step, which updates θ_{t+1}, update hyperparameters of $q(\vec{\theta})$

Example : VBEM for GMM compared to EM for GMM



(a) Graphical notation : EM for GMM (Bishop & Nasrabadi, 2006)



(b) Graphical notation : VBEM for GMM (Bishop & Nasrabadi, 2006)

Parameters: $\Theta := (\tau, \mu_1, ..., \mu_d, \Lambda_1, ..., \Lambda_d)$

1. $\tau = (p(Z_i = 1), ..., p(Z_i = d))$, d: number of groups (=2 in previous example). $p(\tau) = Dir(a_0, ..., a_0)$ for constant a_0.
2. $(\vec{\mu}_1, ..., \vec{\mu}_d)$ where $\vec{\mu}_j$: mean vector of j^{th} Gaussian component.
   $(\Lambda_1, ..., \Lambda_d)$ where Λ_j: covariance matrix of j^{th} Gaussian component.
   : $p(\mu, \Lambda) = p(\mu|\Lambda)p(\Lambda)$: Gaussian-Wishart prior for μ, Λ

Latent Variables: $Z = (\vec{z}_1, ..., \vec{z}_n)$, $\vec{z}_i \in \{1, ..., d\}$. $p(z_i|\tau) = Cat(\tau)$: Categorical Distribution: multinomial dist'n with 1 trial.

Observable Data : $x_1, ..., x_n$.

Joint probability

$p(X, Z, \tau, \mu_1, ..., \mu_d, \Lambda_1, ..., \Lambda_d) = p(X|Z, \mu_1, ..., \mu_d, \Lambda_1, ..., \Lambda_d)p(Z|\tau)p(\tau)p(\mu_1, ..., \mu_d|\Lambda_1, ..., \Lambda_d)p(\Lambda_1, ..., \Lambda_d)$.

Decompose the approximate function : $q(Z, \tau, \mu, \Lambda) = q(Z)q(\tau, \mu, \Lambda) = q(\tau, \mu, \Lambda)\prod_i q(z_i)$.

Derivation of $q(\mathbf{z})$ (variational E step)

The form for $q(\mathbf{z})$ can be obtained by looking at the complete data log joint, ignoring terms that do not involve $\mathbf{z}$, and taking expectations of what's left over wrt all the hidden variables except for $\mathbf{z}$. We have

$$\begin{aligned} \log q(\mathbf{z}) &= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})\right] + \text{const} &\quad (21.126)\\ &= \sum_i \sum_i z_{ik} \log \rho_{ik} + \text{const} &\quad (21.127)\end{aligned}$$

where we define

$$\begin{aligned} \log \rho_{ik} \triangleq\ & \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log \pi_k\right] + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log |\boldsymbol{\Lambda}_k|\right] - \frac{D}{2}\log(2\pi) \\ & -\frac{1}{2}\mathbb{E}_{q(\boldsymbol{\theta})}\left[(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_i - \boldsymbol{\mu}_k)\right] \quad (21.128)\end{aligned}$$

Using the fact that $q(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi})$, we have

$$\log \tilde{\pi}_k \triangleq \mathbb{E}\left[\log \pi_k\right] = \psi(\alpha_k) - \psi\Big(\sum_{k'} \alpha_{k'}\Big) \quad (21.129)$$

(a) Variational E-step (Murphy, 2012)

Derivation of $q(\boldsymbol{\theta})$ (variational M step)

Using the mean field recipe, we have

$$\begin{aligned} \log q(\boldsymbol{\theta}) &= \log p(\boldsymbol{\pi}) + \sum_k \log p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \sum_i \mathbb{E}_{q(\mathbf{z})}\left[\log p(\mathbf{z}_i|\boldsymbol{\pi})\right] \\ &+ \sum_k \sum_i \mathbb{E}_{q(\mathbf{z})}\left[z_{ik}\right] \log \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + \text{const} \quad (21.135)\end{aligned}$$

We see this factorizes into the form

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\pi}) \prod_k q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \quad (21.136)$$

For the $\boldsymbol{\pi}$ term, we have

$$\log q(\boldsymbol{\pi}) = (\alpha_0 - 1)\sum_k \log \pi_k + \sum_k \sum_i r_{ik} \log \pi_k + \text{const} \quad (21.137)$$

Exponentiating, we recognize this as a Dirichlet distribution:

$$\begin{aligned} q(\boldsymbol{\pi}) &= \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) &\quad (21.138)\\ \alpha_k &= \alpha_0 + N_k &\quad (21.139)\\ N_k &= \sum_i r_{ik} &\quad (21.140)\end{aligned}$$

(b) Variational M-step (Murphy, 2012)

Summary of differences between EM and VB

EM

1. Goal: Maximum Likelihood with latent variables. In Bayesian context, find MAP.
2. Not a Bayesian method because it finds an optimal point estimate of θ
3. Sets apart latent variable Z and parameter θ.

VB

1. Goal: A distribution closed to $p(\theta|\mathcal{D})$
2. Bayesian method because it treats θ as a random variable
3. Treats both types of unobserved variables Z and θ as the same.
4. In the variational EM, the distinction between E and M disappears (https://en.wikipedia.org/wiki/Expectation-maximization-algorithm)

Code examples

https://tinyheero.github.io/2016/01/03/gmm-em.html : EM for 1 dimensional GMM

https://rpubs.com/cakapourani/variational-bayes-lr : VB Linear Regression

https://rpubs.com/cakapourani/variational_bayes_gmm : VB for multivariate GMM

References

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. : Ch 13.3 for Laplace method
https://www.youtube.com/watch?v=ErfnhcEV1O8 : Entropy, Cross Entropy, and KL Divergence
Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press. : forward and reverse KL, VB for gaussian
https://en.wikipedia.org/wiki/Evidence_lower_bound : ELBO
https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf : ELBO
https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm : EM Algorithm
http://www.columbia.edu/ mh2078/MachineLearningORFE/EM_Algorithm.pdf : EM Algorithm
https://tinyheero.github.io/2016/01/03/gmm-em.html : EM for 1 dimensional GMM
https://www.youtube.com/watch?v=xH1mBw3tb_ct=1281s : Variational inference, mostly for Mean Field
https://hun-learning94.github.io/posts/2020-08-25-variational-inference/ : a blog post about variational inference
https://en.wikipedia.org/wiki/Variational_Bayesian_methods
https://rpubs.com/cakapourani/variational-bayes-lr : VB Linear Regression
https://rpubs.com/cakapourani/variational_bayes_gmm : VB for multivariate GMM
Bishop, C. M., Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer. : EM vs VBEM for GMM.