Variational Bayes Inference

Contents

1 Introduction

Variational Bayes Inference is a technique which is used a lot in Text Data Analysis and also in Neuroscience or Computational Biology. It is an interesting tool to treat modern problem with large data and large dimensions and it can be very fast.
I used the following conference to create this document https://www.youtube.com/watch?v=Moo4-KR5qNg&list=PLtrOFtjly3lT-fchyIrNkPLRqMkrbpvwV&index=10&t=0s

2 Notions

2.1 Kullback-Leibler divergence

On pourra se référer au cours Topics in Statistical Theory de Yoav Zemel, parties 5.1 et 5.2

The Kullback-Leibler (also called relative entropy) is a measure of how one probability distribution is different from a second, reference probability distribution.

$$KL(P||Q) = D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

In Bayesian inference, $D_{KL}(P||Q)$ is a measure of the information gained by revising one's beliefs from the prior distribution Q to the posterior distribution P. It represents the amount of information lost when Q is used to approximate P.

2.2 Bayesian inference

In Bayesian inference, the first goal is to identify a parameter θ unknown. We have the following formula (Bayes' formula) :

$$p(y_{1:N}|\theta) \cdot p(\theta) \propto_\theta p(\theta|y_{1:N})$$

- $p(\theta)$ is the prior. It represents our previous knowledge about the distribution of the parameter θ.
- $p(y_{1:N}|\theta)$ is the likelihood. It represents the probability to obtain the value $y_{1:N}$ given θ.
- $p(\theta|y_{1:N})$ is the posterior. It represents the distribution of the parameter θ given the information we have with the data $y_{1:N}$. The idea is : "we know more about θ after, we want our uncertainty to decrease.

The steps of Bayesian inference are :

1. Build a model : choose prior and choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances.

Why are steps 2 and 3 hard ?

- Typically no closed form for the posterior
- High dimensional integration (difficult when $d \geq 3$) as $p(y_{1:N}) = \int p(y_{1:N}, \theta) d\theta$

Thereby we use a class of computational methods called Approximate Bayesian Inference. Markov Chain Monte Carlo (MCMC) is a gold standard in Approximate Bayesian Inference. It is eventually accurate but can be slow. So we will describe an optimazation approach which can be faster : Variational Bayes Inference.

3 Variational Bayes Inference

3.1 Principle

The goal is to approximate the posterior with a distrubtion q^*. We want to find q^* which is the *closest* to $p(\theta|y)$ among a family of "*nice*" distributions (i.e easy to compute means, (co)variances).

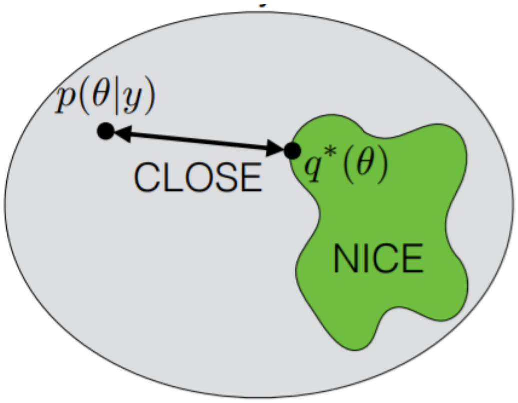


Figure 1 – VBI Principle

3.2 The optimization problem

Question : How to define a distance between distributions ?
Answer : We use the Kullback-Leibler divergence. We can rewrite the KL divergence as below :

$$D_{KL}(q(\cdot)||p(\cdot|y)) = \int q(\theta) \log \left(\frac{q(\theta)}{p(\theta|y)} \right) d\theta$$

$$D_{KL}(q(\cdot)||p(\cdot|y)) = \int q(\theta) \log \left(\frac{q(\theta)p(y)}{p(\theta, y)} \right) d\theta$$

$$D_{KL}(q(\cdot)||p(\cdot|y)) = log(p(y)) - \int q(\theta) \log \left(\frac{p(\theta, y)}{q(\theta)} \right) d\theta$$

Does not depend on q We know the distributions in the integral

The second term is called Evidence lower bound (ELBO). We have :

$$KL \geq 0 \Rightarrow \log(p(y)) \geq \text{ELBO}$$

We obtain the following optization problem :

$$\boxed{q^* = \underset{q \in Q}{\text{argmax}}\, \text{ELBO}(q)}$$

Question : Other divergences/metrics can be possible such as the Total Variation. Why do we use KL (in this direction) ?
Answer : To compute to trick above.

3.3 Mean-field variational Bayes

Question : How to choose a family of "nice" distributions ?
Answer : We use Mean-field variational Bayes (easy to compute means, (co)variances,...)

$$Q_{\text{MFVB}} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

Another example could be exponential family.

3.4 Examples

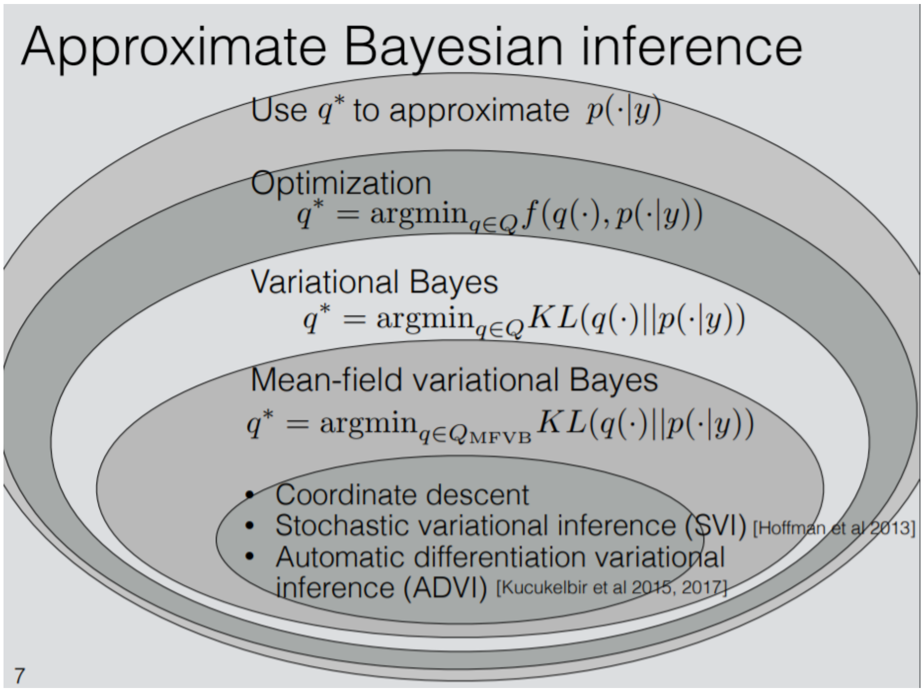
Cf a basic example on `https: // en. wikipedia. org/ wiki/ Variational_ Bayesian_ methods`

3.5 How to solve the optimization problem ?

Possibilities :

- Coordinate descente in $q_1, ..., q_J$
- Stochastic gradient descent (stochastic variational inference)
- Automatic differential variational inference (ADVI)

3.6 A summary



Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]

7

Figure 2 – VBI summary