QCB 408 / 508 – Notes on Week 06

Juechun Tang

2020-03-22

Summary

In week 6, we continued our discussion on Bayesian Estimation and classificaiton, different ways of doing priors, numerical methods and likihood; and the following concepts are discussed:

- Bayesian Estimation
  - Posterier Distribution
  - Posterior Expection
  - Posterior Intervals
  - Maximum a posteriori Probability
  - Loss Function
  - Bayes Risk
  - Bayes Estimation
- Bayes Classification
  - Posterior Probability
  - loss Function
  - Bayes Risk
  - Bayes Rule
- Priors
  - Conjugate Priors
  - Jeffreys Prior
  - Improper Prior
- Empirical Bayes
- Numerical methods
  - Latent varibale models, EM models
  - Markov chain Monte Carlo

Estimation

Bayesian estimation is an anolog of point estimation, such as posterior expectation, posterior median, MAP,etc.

Posterier distribution

Assume we have $(X_1, X_2, \ldots, X_n)|\theta \overset{\text{iid}}{\sim} F_\theta$ with prior distribution $\theta \sim F_\tau$ The posterior distribution of $\theta|\boldsymbol{X}$ is obtained through Bayes theorem:

$$f(\theta|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\theta)f(\theta)}{f(\boldsymbol{X})} = \frac{f(\boldsymbol{X}|\theta)f(\theta)}{\int f(\boldsymbol{X}|\theta^*)f(\theta^*)d\theta^*} \propto L(\theta;\boldsymbol{x})f(\theta)$$

Note $\boldsymbol{X}$ is bold and capital, representing r.v. Now θ is a random variable, so we write in term of $f(\boldsymbol{X}|\theta)$. Here we use a condensed notation (not writing τ). $f(\boldsymbol{X})$ is the mariginal pdf or pmf of the data. The denominator is what usually difficult to calculate. This posterior pdf $\propto L(\theta;\boldsymbol{x})f(\theta)$ because 1. $f(\boldsymbol{X}|\boldsymbol{\theta})$ is the likihood function and 2. the marginal on the denominator is not a function of θ.

Posterior Expectation

Once we havce the posterior pdf or pmf, we can do any probablity calcualtion we want. The the posterior expected value is:

$$\begin{aligned} \mathrm{E}[\theta|\boldsymbol{x}] &= \int \theta f(\theta|\boldsymbol{x})d\theta & (1)\\ &= \frac{\int \theta L(\theta;\boldsymbol{x})f(\theta)d\theta}{\int L(\theta;\boldsymbol{x})f(\theta)d\theta} & (2) \end{aligned}$$

Equation 2 is got from substiting the Posterier distribution $f(\theta|\boldsymbol{x})$ in and $f(x)$ cancels out eventually. This is equivalent to the motivating example from last week.

Posterior Interval

A Bayesian analog of the confidence interval: an interval within which the parameter of interest falls within. This is straight probability calculation. Note θ is posterior distribution conditioned on $\boldsymbol{x}$

$$1 - \alpha = \Pr(C_\ell \leq \theta \leq C_u|\boldsymbol{x})$$

Maximum *A Posteriori* Probability (MAP)

This is maximum likihood analog. We are maimizing the likihood times the prior (weighting) over θ.

$$\begin{aligned} \hat{\theta}_{\mathrm{MAP}} &= \mathrm{argmax}_\theta \Pr(\theta|\boldsymbol{x})f(\theta)\\ &= \mathrm{argmax}_\theta L(\theta;\boldsymbol{x})f(\theta) \end{aligned}$$

Note for many machine learning settings, if there are infinite amount of data $f(\theta)$ peaks to a true value.

Loss functions (Error in estimate)

Let $\mathcal{L}\left(\theta, \tilde{\theta}\right)$ be the loss function for a given estimator $\tilde{\theta}$.We can have for example squared error loss or absolute error loss:

$$\mathcal{L}\left(\theta, \tilde{\theta}\right) = \left(\theta - \tilde{\theta}\right)^2 \text{ or } \mathcal{L}\left(\theta, \tilde{\theta}\right) = \left|\theta - \tilde{\theta}\right|.$$

Fields like social sciences and economics commonly use loss functions.

$$\begin{aligned}\mathrm{E}\left[\left(\theta-\tilde{\theta}\right)^2\right] &= \left(\mathrm{E}\left[\tilde{\theta}\right]-\theta\right)^2+\mathrm{Var}\left(\tilde{\theta}\right)\\ &= \text{bias}^2+\text{variance}\end{aligned}$$

Let's work through how we get the result:

$$\begin{aligned}\mathrm{E}\left[\left(\theta-\tilde{\theta}\right)^2\right] &= \mathrm{E}\left[(\theta^2-2\theta\tilde{\theta}+\tilde{\theta}^2)\right]\\ &= \theta^2-2\theta\,\mathrm{E}[\tilde{\theta}]+\mathrm{E}[\tilde{\theta}^2]\\ &= \mathrm{E}[\tilde{\theta}]^2+\theta^2-2\theta\,\mathrm{E}[\tilde{\theta}]+\mathrm{E}[\tilde{\theta}^2]-\mathrm{E}[\tilde{\theta}]^2\\ &= \left(\mathrm{E}\left[\tilde{\theta}\right]-\theta\right)^2+\mathrm{Var}\left(\tilde{\theta}\right)\end{aligned}$$

Bayes Risk

The **Bayes risk**, $R\left(\theta,\tilde{\theta}\right)$, is the expected loss with respect to the posterior:

$$\mathrm{E}\left[\,\mathcal{L}\left(\theta,\tilde{\theta}\right)\middle|\,\boldsymbol{x}\right]=\int\mathcal{L}\left(\theta,\tilde{\theta}\right)f(\theta|\boldsymbol{x})d\theta$$

The intergral is taken over the posterior distribution $\theta|\boldsymbol{x}$. What we have here is that we first conditioned on observed data $\boldsymbol{x}$, and then do the probability calculation. Note both θ and $\tilde{\theta}$ are conditioned on data.

Bayes Estimators

Once we have Bayes risk from loss function, we can calculate Bayes Estimators to minimize the Bayes risk. Note everything is in finite sample size, n is not going to ∞. The challenge is what the prior distribution is, which we will discuss more in details.

The posterior expectation $\mathrm{E}\left[\,\theta\middle|\,\boldsymbol{x}\right]$ minimizes the Bayes risk of $\mathcal{L}\left(\theta,\tilde{\theta}\right)=\left(\theta-\tilde{\theta}\right)^2$ (squared loss).

The posterior median of $f(\theta|\boldsymbol{x})$, calculated by $F^{-1}_{\theta|\boldsymbol{x}}(1/2)$, minimizes the Bayes risk of $\mathcal{L}\left(\theta,\tilde{\theta}\right)=\left|\theta-\tilde{\theta}\right|$ (absolute loss).

Bayes Classification

Bayes classification is an analog of hypothesis testing. The **Neyman-Pearson Lemma** and **Generalized LRT** we covered before have a direct anaolog in baysesian inference. Typically we deal with composite hypothesis where ther are more than one value in each hypothesis.

We will set up a composite vs. composite in the following section.

Assumptions

Let $(X_1,X_2,\ldots,X_n)|\theta\overset{\text{iid}}{\sim}F_\theta$ where $\theta\in\Theta$ and $\theta\sim F_\tau$. Let $\Theta_0,\Theta_1\subseteq\Theta$ so that $\Theta_0\cap\Theta_1=\varnothing$ and $\Theta_0\cup\Theta_1=\Theta$.

Given observed data $\boldsymbol{x}$, we wish to classify whether $\theta\in\Theta_0$ or $\theta\in\Theta_1$.

The null hypothesis is $\theta\in\Theta_0$ and the alternative hypotehsis is $\theta\in\Theta_1$.

Prior Probability on H

Let H be a rv such that $H = 0$ when $\theta \in \Theta_0$ and $H = 1$ when $\theta \in \Theta_1$.

From the prior distribution on θ, we can calculate

$$\Pr(H = 0) = \int_{\theta \in \Theta_0} f(\theta) d\theta$$

and $\Pr(H = 1) = 1 - \Pr(H = 0)$.

A critisium for the Bayesian is that they often think parameters as a interval instead of a single value. From the Null hypothesis significance testing (NHST), it's not reasonable to have a simple null phypothesis. One exception that's toally scientific is genetics mapping where we consider if the recombination is $1/2$ or less than $1/2$.

Posterior Probability

Using Bayes theorem, we can also calculate

$$\begin{aligned} \Pr(H = 0|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|H = 0) \Pr(H = 0)}{f(\boldsymbol{x})} \\ &= \frac{\int_{\theta \in \Theta_0} f(\boldsymbol{x}|\theta) f(\theta) d\theta}{\int_{\theta \in \Theta} f(\boldsymbol{x}|\theta) f(\theta) d\theta} \end{aligned}$$

where $\Pr(H = 1|\boldsymbol{x}) = 1 - \Pr(H = 0|\boldsymbol{x})$.

Loss Function

We can only make 2 types of error for classification: 0 when it's 1 or 1 when it's 0.

Let $\tilde{H}$ be estimate for H, $\mathcal{L}\left(\tilde{H}, H\right)$ be such that

$$\begin{aligned} \mathcal{L}\left(\tilde{H} = 1, H = 0\right) &= c_I \\ \mathcal{L}\left(\tilde{H} = 0, H = 1\right) &= c_{II} \end{aligned}$$

for some $c_I, c_{II} > 0$, where c_I indicates Type I error and c_{II} indicates Type II error.

Bayes Risk

Now that we get loss function from posterior and prior, we will calculate Bayes Risk as we did for estimation, and then Bayes Estimate which minimize the Bayes Risk.

The Bayes risk, $R\left(\tilde{H}, H\right)$, is

$$\begin{aligned} \mathrm{E}\left[\mathcal{L}\left(\theta, \tilde{\theta}\right) \middle| \boldsymbol{x}\right] &= c_I \Pr(\tilde{H} = 1, H = 0) + c_{II} \Pr(\tilde{H} = 0, H = 1) \\ &= c_I \Pr(\tilde{H} = 1|H = 0) \Pr(H = 0) + c_{II} \Pr(\tilde{H} = 0|H = 1) \Pr(H = 1) \end{aligned}$$

Note that the equation is interpreted as the sum of Type I penalty times Type I error rate times the prior plus the Type II penalty times Type II error rate times the prior.

The challenge is to determine what c_I, c_{II} and prior are. If there are trillions of data points, we can get values for them.

Bayes Rule (Bayes Estimator)

The estimate $\tilde{H}$ that minimizes $R\left(\tilde{H}, H\right)$ is

$$\tilde{H} = 1 \text{ when } \Pr(H = 1|\boldsymbol{x}) \geq \frac{c_I}{c_I + c_{II}}$$

and $\tilde{H} = 0$ otherwise.

Priors

The challenge to get the Posterior distribution is how we get the prior distribution. Sometimes the prior is chosen for mathematical conveninence, other types are estimated using data.

Conjugate Priors

A **conjugate prior** is a prior distribution for a data generating distribution so that the posterior distribution is of the same type as the prior.

Conjugate priors are useful for obtaining stratightforward calculations of the posterior. If we can find a conjugate prior, we ca navoid calculating mariginal on the denominator.

There is a systematic method for calculating conjugate priors for exponential family distributions.

Example: Beta-Bernoulli

Suppose $\boldsymbol{X}|p \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ and suppose that the prior $p \sim \text{Beta}(\alpha, \beta)$. The posterior is:

$$\begin{aligned} f(p|\boldsymbol{x}) &\propto L(p; \boldsymbol{x}) f(p) \\ &= p^{\sum x_i}(1-p)^{\sum(1-x_i)} p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha-1+\sum x_i}(1-p)^{\beta-1+\sum(1-x_i)} \\ &\propto \text{Beta}(\alpha + \sum x_i, \beta + \sum(1-x_i)) \end{aligned}$$

Therefore,

$$\text{E}[p|\boldsymbol{x}] = \frac{\alpha + \sum x_i}{\alpha + \beta + n}.$$

If there's a true p, note that as $n \to \text{inf}$, p concentrates to true p (Strong Law of large numbers).

Example: Normal-Normal

Suppose $\boldsymbol{X}|\mu \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$, where σ^2 is known, and suppose that $\mu \sim \text{Normal}(a, b^2)$.

Then it can be shown that $\mu|\boldsymbol{x} \sim \text{Normal}(\text{E}[\mu|\boldsymbol{x}], \text{Var}(\mu|\boldsymbol{x}))$

$$\text{E}[\mu|\boldsymbol{x}] = \frac{b^2}{\frac{\sigma^2}{n} + b^2}\overline{x} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + b^2}a$$

$$\text{Var}(\mu|\boldsymbol{x}) = \frac{b^2 \frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + b^2}$$

Let's work through:

$$\begin{aligned} f(\mu|\boldsymbol{x}) &\propto L(\mu;\boldsymbol{x})f(\mu) \\ &= e^{\frac{-\sum(x_i-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-a)^2}{2b^2}} \\ &= e^{\frac{-\sum {x_i}^2 - n\mu^2 + 2\mu n\overline{x}}{2\sigma^2} - \frac{\mu^2+a^2-2\mu a}{2b^2}} \\ &\propto e^{\frac{-n\mu^2+2\mu n\overline{x}}{2\sigma^2} - \frac{\mu^2-2\mu a}{2b^2}} \\ &= e^{\frac{-\mu^2}{2}(\frac{n}{\sigma^2}+\frac{1}{b^2})+\mu(\frac{n\overline{x}}{\sigma^2}+\frac{a}{b^2})} \end{aligned}$$

Let's try to get a quadraticfunction in terms of μ:

Let $\lambda^2 = (\frac{n}{\sigma^2}+\frac{1}{b^2})^{-1}, \frac{\lambda'}{\lambda^2} = \frac{n\overline{x}}{\sigma^2}+\frac{a}{b^2}$, We have:

$$\begin{aligned} f(\mu|\boldsymbol{x}) &= e^{-\frac{\mu^2}{2\lambda^2}+\frac{\lambda'}{\lambda^2}\mu} \\ &\propto e^{-\frac{\mu^2}{2\lambda^2}+\frac{\lambda'}{\lambda^2}\mu+\frac{\lambda'^2}{2\lambda^2}} \\ &= e^{\frac{-(\mu-\lambda')^2}{2\lambda^2}} \\ &\propto \text{Normal}(\lambda', \lambda^2) \end{aligned}$$

where

$$\lambda' = (\frac{n\overline{x}}{\sigma^2}+\frac{a}{b^2})(\frac{b^2\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2}) = \frac{b^2}{\frac{\sigma^2}{n}+b^2}\overline{x} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2}a$$

$$\lambda^2 = \frac{\sigma^2 b^2}{nb^2+\sigma^2} = \frac{b^2\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2}$$

Note that the posterior mean is a weighted average of the sample mean and the prior a. It's weighted by the weighted averaged of the two variances. Again, as n goes to inf, if I'm a frequentist, then we will get the true value (variance goes to 0). Note the conjugate prior only tells which distribution to pick, but we still need to specify what the parameter values are.

Jeffreys Prior

Another way to pick prior for mathematical convinience is through **Jeffreys Prior**. After reparameterization, we get the same answer when doing inference.

If we do inference based on prior $\theta \sim F_\tau$ to obtain $f(\theta|\boldsymbol{x}) \propto L(\theta;\boldsymbol{x})f(\theta)$, it follows that this inference may *not* be invariant to transformations of θ, such as $\eta = g(\theta)$.

If we utilize a **Jeffreys prior**, which means it is such that

$$f(\theta) \propto \sqrt{I(\theta)}$$

then the prior will be invariant to transformations of θ. We would want to show that $f(\theta) \propto \sqrt{I(\theta)}$ implies $f(\eta) \propto \sqrt{I(\eta)}$.

Examples: Jeffreys Priors

Normal(μ, σ^2), σ^2 known: $f(\mu) \propto 1$. Note $f(\mu) \propto 1$ a problem because this is not a proper probability distribution

Normal(μ, σ^2), μ known: $f(\sigma) \propto \frac{1}{\sigma}$

Poisson(λ): $f(\lambda) \propto \frac{1}{\sqrt{\lambda}}$

Bernoulli(p): $f(p) \propto \frac{1}{\sqrt{p(1-p)}}$

Improper Prior

An **improper prior** is a prior such that $\int f(\theta)d\theta = \infty$. Therefore, it is not a proper probability distribution, but we can view it as a weighting function and the resulting posterior $f(\theta|\boldsymbol{x}) \propto L(\theta;\boldsymbol{x})f(\theta)$ yields a probability distribution.

Take for example the case where $\boldsymbol{X}|\mu \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$, where σ^2 is known, and suppose that $f(\mu) \propto 1$. Then $\int f(\theta)d\theta = \infty$, but

$$f(\theta|\boldsymbol{x}) \propto L(\theta;\boldsymbol{x})f(\theta) \sim \text{Normal}\left(\overline{x}, \sigma^2/n\right)$$

which is a proper probability distribution.

Empirical Bayes

Empirical Bayes uses the data to estimate the **prior parameter**. A lot of ML methods are **Empirical Bayes** in terms of still treating parameter as a random variable, but everything is evaluated from data. It is commonly used to estimate multiple parameters.

This is especially useful for high-dimensional data when many parameters are simultaneously drawn from a prior with multiple observations drawn per parameter realization.

There are both parametric and nonparametric empirical bayes. Nonparametric means doing inference in a way that we are making minimal assumptions of the distribution. And we focuses on parametric bayes here.

Approach

The usual approach is to integrate out the parameter to obtain

$$f(\boldsymbol{x};\tau) = \int f(\boldsymbol{x}|\theta)f(\theta;\tau)d\theta.$$

Now we have pdf or pmf my data in terms of prior parameters. An estimation method (such as MLE) is then applied to estimate τ. Then inference proceeds as usual under the assumption that $\theta \sim f(\theta;\hat{\tau})$.

Problem is that we tend to overfit the data. In ML, what people do in practice is to keep some test data, perfrom cross-validation to avoid overfitting.

MLE tends to fail when doing high-dimentional data. Some emprical bayes takes overfitting into account and gets much better results.

Example: Normal

Suppose that $X_i|\mu_i \sim \text{Normal}(\mu_i, 1)$ for $i = 1, 2, \ldots, n$ where these rv's are independent. Also suppose that $\mu_i \overset{\text{iid}}{\sim} \text{Normal}(a, b^2)$.

Note here we have a different μ for every observcation. Let this be a toy model for gene expression where each X_i is the sample mean from gene i, and we model this in aggregates. Note it's sufficient for μ_i per sufficient statistics. Now we integrate our μ_i

$$f(x_i; a, b) = \int f(x_i|\mu_i) f(\mu_i; a, b) d\mu_i \sim \text{Normal}(a, 1 + b^2).$$

My MLE are:

$$\implies \hat{a} = \overline{x},\ 1 + \hat{b}^2 = \frac{\sum_{k=1}^{n} (x_k - \overline{x})^2}{n}$$

If the $1 + \hat{b}^2 < 1$, this term gets truncated (Restricted).

From **conjugate prior**, we have:

$$\text{E}[\mu_i|x_i] = \frac{1}{1+b^2} a + \frac{b^2}{1+b^2} x_i \implies$$

$$\begin{aligned} \hat{\text{E}}[\mu_i|x_i] &= \frac{1}{1+\hat{b}^2} \hat{a} + \frac{\hat{b}^2}{1+\hat{b}^2} x_i \\ &= \frac{n}{\sum_{k=1}^{n} (x_k - \overline{x})^2} \overline{x} + \left(1 - \frac{n}{\sum_{k=1}^{n} (x_k - \overline{x})^2}\right) x_i \end{aligned}$$

This is called shrinkage estimator. We just take sample mean for gene 1 as μ_1 and so on. For empirical Bayes, we take weighted average with the overall pulled sample mean. Sum of squared error loss and empirical Bayes are uniformly better than the MLE even if the truncated version.

Numerical Methods

Challenges

Frequentist model:

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}$$

Bayesian model:

$$X_1, X_2, \ldots, X_n | \boldsymbol{\theta} \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}} \text{ and } \boldsymbol{\theta} \sim F_{\boldsymbol{\tau}}$$

Sometimes it's not possible to find formulas for $\hat{\boldsymbol{\theta}}_{\text{MLE}}$, $\hat{\boldsymbol{\theta}}_{\text{MAP}}$, $\text{E}[\boldsymbol{\theta}|\boldsymbol{x}]$, or $f(\boldsymbol{\theta}|\boldsymbol{x})$. We have to use numerical methods instead.

Approaches

Some numberical approach for likelihood based inference:

- Expectation-maximization (EM) algorithm
- Variational inference: an extension of the EM algorithm
- Markov chain Monte Carlo (MCMC)
  - Metropolis sampling
  - Metropolis-Hastings sampling
  - Gibbs sampling

In class, we mainly discussed EM and MCMC.

Latent Variable Models

One way think about EM algorithm is through latent variable models.

Latent variables (or hidden variables) are random variables that are present in the model, but unobserved.

We will denote latent variables by Z, and we will assume

$$(X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n) \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}.$$

A realized value of Z is z, $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)^T$, etc.

Bayesian models are a special case of latent variable models: the unobserved random parameters are latent variables.

Empirical Bayes Revisited

In the earlier EB example, we supposed that $X_i|\mu_i \sim \text{Normal}(\mu_i, 1)$ for $i = 1, 2, \ldots, n$ where these rv's are independent, and also that $\mu_i \overset{\text{iid}}{\sim} \text{Normal}(a, b^2)$.

The unobserved parameters $\mu_1, \mu_2, \ldots, \mu_n$ are latent variables. In this case, $\boldsymbol{\theta} = (a, b^2)$.

Normal Mixture Model

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}$ be a sample of n independent observations from a mixture of k normal distributions with pdf

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right\}.$$

where π_k represents the probability of which component from the mixture is drawn from and $\sum \pi_k = 1$. In the simple case of 2 Gaussian mixture, $\pi_1 + \pi_2 = 1$. Note this is a messy function which is not normal.

The goal is to estimate the unknown parameters $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. In this case, the MLEs of the unknown paramaters cannot be found analytically. Mixture models allow us to estimate the parameters.

Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n \overset{\text{iid}}{\sim} \text{Multinomial}_K(1, \boldsymbol{\pi})$ be the latent variables that determine the component k from which the observation is drawn from, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. Note that $Z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} Z_{ik} = 1$. Note each $\boldsymbol{Z}_i$ is a vector of length k.

Let $[X_i|Z_{ik} = 1] \sim \text{Normal}(\mu_k, \sigma_k^2)$, where $\{X_i|\boldsymbol{Z}_i\}_{i=1}^n$ are jointly independent. The joint pdf is

$$f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right\} \right]^{z_{ik}}.$$

from which we can form an analytic solution.

Note that

$$f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{z}_i; \boldsymbol{\theta}).$$

It can be verified that $f(\boldsymbol{x}; \boldsymbol{\theta})$ is the marginal distribution of this latent variable model:

$$f(x_i; \boldsymbol{\theta}) = \sum_{\boldsymbol{z}_i} f(x_i, \boldsymbol{z}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right\}.$$

Bernoulli Mixture Model

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, p_1, \ldots, p_K)$ with pmf

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k p_k^{x_i} (1 - p_k)^{1 - x_i}.$$

As in the Normal mixture model, the MLEs of the unknown paramaters cannot be found analytically.

As before, there is a latent variable model that produces the same marginal distribution and likelihood function. Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n \overset{\text{iid}}{\sim} \text{Multinomial}_K(1, \boldsymbol{\pi})$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. Note that $Z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} Z_{ik} = 1$. Let $[X_i | Z_{ik} = 1] \sim \text{Bernoulli}(p_k)$, where $\{X_i | \boldsymbol{Z}_i\}_{i=1}^{n}$ are jointly independent.

The joint pmf is

$$f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k p_k^{x_i} (1 - p_k)^{1 - x_i} \right]^{z_{ik}}.$$

EM Algorithm

We start from an intractable pdf or pmf. After we introduce a latent variable, the likihood becomes tractable and then we can derive MLE or MAP. If we have clinical studies and have missing data, this framework works well.

Conceptually, what we basically do is that suppose we have some data points that may come from a mixutre of known distributions, we randomly assign parameters values and do soft-clutering, that is, calculate the probablity of which component each data point come from. Then the parameters are optimized and iterate until convergence.

Rationale

For any likelihood function, $L(\boldsymbol{\theta}; \boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta})$, there is an abundance of optimization methods that can be used to find the MLE or MAP. However:

- Optimization methods can be messy to implement
- There may be probabilistic structure that we can use to simplify the optimization process and also provide theoretical guarantees on its convergence
- Optimization isn't necessarily the only goal, but one may also be interested in point estimates of the latent variable values

Requirement

The expectation-maximization (EM) algorithm allows us to calculate MLEs and MAPs when certain geometric properties are satisfied in the probabilistic model.

In order for the EM algorithm to be a practical approach, then we should have a latent variable model $f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$ that is used to do inference on $f(\boldsymbol{x}; \boldsymbol{\theta})$ or $f(\boldsymbol{\theta} | \boldsymbol{x})$.

Note: Sometimes $(\boldsymbol{x}, \boldsymbol{z})$ is called the **complete data** and $\boldsymbol{x}$ is called the **observed data** when we are using the EM as a method for dealing with missing data.

To do EM, we first need to know the distribution, probably from EDA to justify the validity of the probablistic model selected.

The Algorithm

1. Choose initial value $\boldsymbol{\theta}^{(0)}$
2. Calculate $f(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})$
3. Calculate
$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}} \left[\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}); \boldsymbol{\theta}^{(t)} \right]$$
where $\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta})$ is the complete data log likihood. It's a function of Z which we replace with $\mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}$ given current estimate of $\boldsymbol{\theta}^{(t)}$. As we integrate out Z, we can then do MLE or MAP.
4. Set
$$\boldsymbol{\theta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$$
5. Iterate until convergence and set $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(\infty)}$

$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$

Continuous $\boldsymbol{Z}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \int \log f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}) d\boldsymbol{z}$$

Discrete $\boldsymbol{Z}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{\boldsymbol{z}} \log f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

EM for MAP

If we wish to calculate the MAP we replace $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ with

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}} \left[\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}); \boldsymbol{\theta}^{(t)} \right] + \log f(\boldsymbol{\theta})$$

where $f(\boldsymbol{\theta})$ is the prior distribution on $\boldsymbol{\theta}$ as we shown before in the bayesian framework.

EM Examples

Normal Mixture Model

Returning to the Normal mixture model, we first calculate the log likihood of the complete data:

$$f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right\} \right]^{z_{ik}}.$$

$$\log f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \pi_k + z_{ik} \log \phi(x_i; \mu_k, \sigma_k^2)$$

where

$$\phi(x_i; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right\}.$$

In caculating

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}\left[\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}); \boldsymbol{\theta}^{(t)}\right]$$

we only need to know $\mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}[Z_{ik}|\boldsymbol{x}; \boldsymbol{\theta}]$, which turns out to be

$$\mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}[Z_{ik}|\boldsymbol{x}; \boldsymbol{\theta}] = \frac{\pi_k \phi(x_i; \mu_k, \sigma_k^2)}{\sum_{j=1}^K \pi_j \phi(x_i; \mu_j, \sigma_j^2)}.$$

We got the result from Bayes rule the the numerator is the prior times the normal pdf and the denominator is the marignal $f(x)$.

Note that we take

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}\left[\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}); \boldsymbol{\theta}^{(t)}\right]$$

so the parameter in $\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta})$ is a free $\boldsymbol{\theta}$, but the paramaters used to take the conditional expectation of $\boldsymbol{Z}$ are fixed at $\boldsymbol{\theta}^{(t)}$. Let's define

$$\hat{z}_{ik}^{(t)} = \mathrm{E}\left[z_{ik}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right] = \frac{\pi_k^{(t)} \phi(x_i; \mu_k^{(t)}, \sigma_k^{2,(t)})}{\sum_{j=1}^K \pi_j^{(t)} \phi(x_i; \mu_j^{(t)}, \sigma_j^{2,(t)})}.$$

E-Step

We calculate

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) &= \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}\left[\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\theta}); \boldsymbol{\theta}^{(t)}\right] \\ &= \sum_{i=1}^n \sum_{k=1}^K \hat{z}_{ik}^{(t)} \log \pi_k + \hat{z}_{ik}^{(t)} \log \phi(x_i; \mu_k, \sigma_k^2) \end{aligned}$$

At this point the parameters making up $\hat{z}_{ik}^{(t)}$ are fixed at $\boldsymbol{\theta}^{(t)}$. Note μ_k, σ_k^2, π_k are all free parameters.

M-Step

We now caculate $\boldsymbol{\theta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.

To get $\pi_k^{(t+1)}$, we use a Lagrange multiplier. Let $\sum_{i=1}^n \hat{z}_{ik}^{(t)} = a_k$, we have

$$\max \sum_{k=1}^K a_k \log \pi_k$$

s.t. $\pi_k \geq 0, \sum \pi_k = 1$

$$\max_{\pi_k} \min_{\lambda} \sum_{k=1}^{K} a_k \log \pi_k + \lambda(\sum(\pi_k - 1)) \leq \max_{\pi_k} \sum_{k=1}^{K} a_k \log \pi_k + \lambda^*(\sum(\pi_k - 1)) = f(\pi)$$

where λ^* is a random λ

$$\frac{\partial f(\pi)}{\partial \pi_k} = \frac{a_k}{\pi_k} + \lambda^* = 0 \quad \forall k$$

$$\sum \pi_k = \sum -\frac{a_k}{\lambda^*} = 1$$

we get $\lambda^* = -\sum a_k$, $\pi_k = \frac{a_k}{\sum a_k} \to \pi_k^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}}{n}$.

Similarly, we can get;

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_i}{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}}$$

$$\sigma_k^{2,(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}}$$

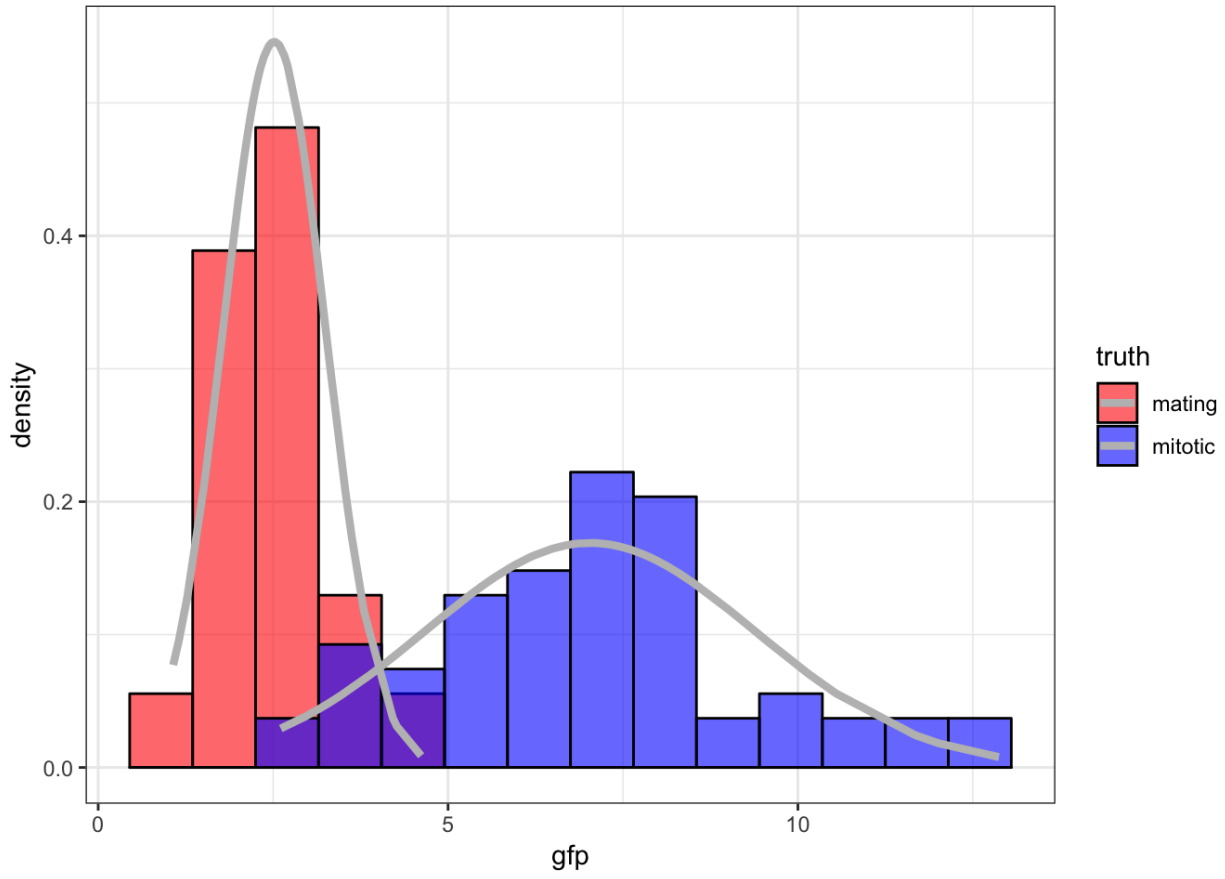
Caveat

If we assign one and only one data point to mixture component k, meaning $\mu_k^{(t)} = x_i$ and $\hat{z}_{ik}^{(t)} = 1$ for some k and i, then as $\sigma_k^{2,(t)} \to 0$, the likelihood goes to ∞.

Therefore, when implementing the EM algorithm for this particular Normal mixture model, we have to be careful to bound all $\sigma_k^{2,(t)}$ away from zero and avoid this scenario.

Yeast Gene Expression

Measured ratios of the nuclear to cytoplasmic fluorescence for a protein-GFP construct that is hypothesized as being nuclear in mitotic cells and largely cytoplasmic in mating cells. We will run an EM and then compare with the the true distribution.



Initialize Values

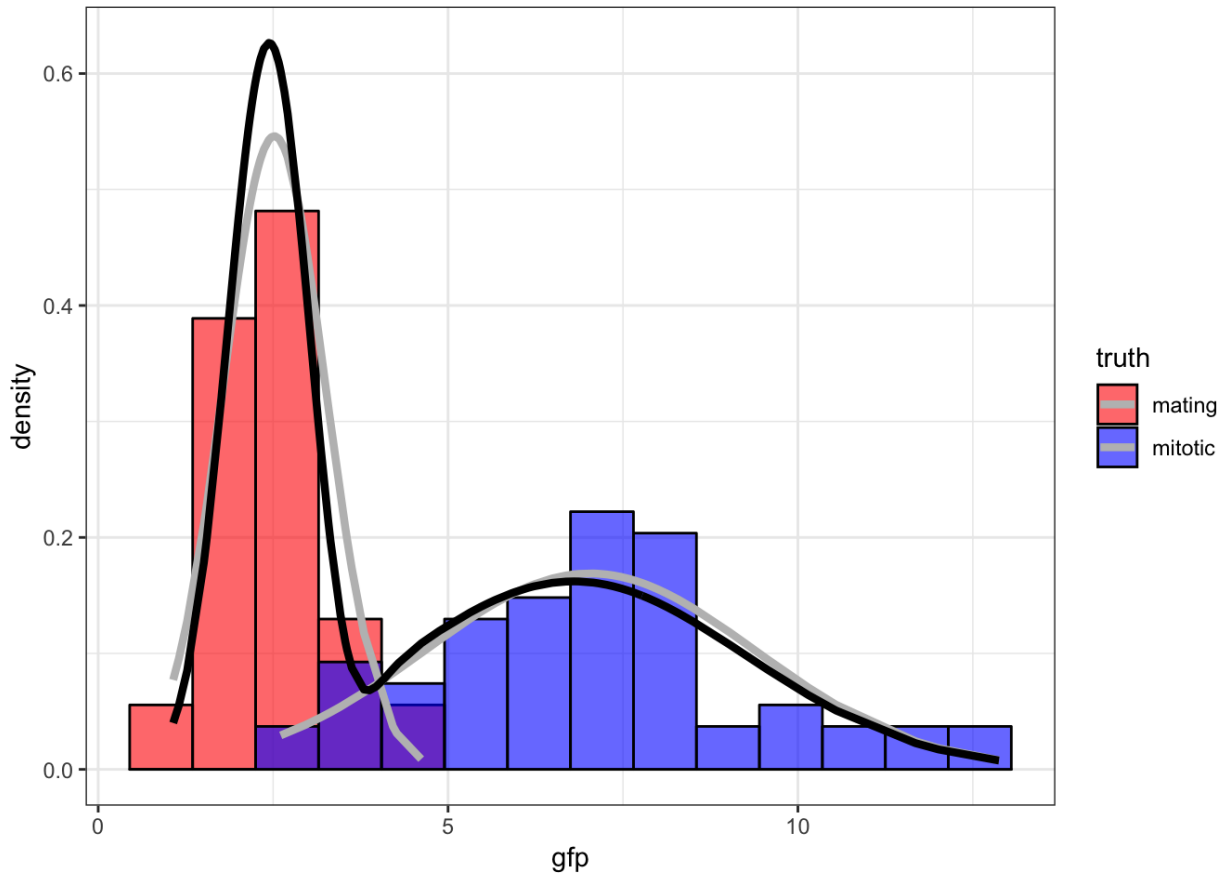
```
> set.seed(508)
> B <- 100 #number of iteration, although need to check if converge at 100
> p <- rep(0,B) #pi vector
> x <- rep(0,B) #added just to compile, don't have the data locally
> mu1 <- rep(0,B)
> mu2 <- rep(0,B)
> s1 <- rep(0,B)
> s2 <- rep(0,B)
> p[1] <- runif(1, min=0.1, max=0.9)
> mu.start <- sample(x, size=2, replace=FALSE)
> #x is a vector of data, draw 2 data potin ramdomly and set as initial mu
> mu1[1] <- min(mu.start)
> mu2[1] <- max(mu.start)
> s1[1] <- var(sort(x)[1:60]) #bottome half from component 1
> s2[1] <- var(sort(x)[61:120])
> z <- rep(0,120) #indication variable whether it's from population 2 or 1
```

A smarter way of setting of mean can be done as the same as variance

Run EM Algorithm

```
> for(i in 2:B) {
+   z <- (p[i-1]*dnorm(x, mean=mu2[i-1], sd=sqrt(s2[i-1])))/
+     (p[i-1]*dnorm(x, mean=mu2[i-1], sd=sqrt(s2[i-1])) +
+        (1-p[i-1])*dnorm(x, mean=mu1[i-1], sd=sqrt(s1[i-1])))
+   # z is indicator of whether it is from component 2.
+   # It's 1 if from 2 0 if from 1
+   #denominator is the marginal distribution of x
+   mu1[i] <- sum((1-z)*x)/sum(1-z)
+   mu2[i] <- sum(z*x)/sum(z)
+   s1[i] <- sum((1-z)*(x-mu1[i])^2)/sum(1-z)
+   s2[i] <- sum(z*(x-mu2[i])^2)/sum(z)
+   p[i] <- sum(z)/length(z)
+ }
> tail(cbind(mu1, s1, mu2, s2, p), n=3)
       mu1 s1 mu2 s2  p
 [98,]  NA NA  NA NA NA
 [99,]  NA NA  NA NA NA
[100,]  NA NA  NA NA NA
> #see if convergence
```

Fitted Mixture Distribution



Grey line is when we know the truth (complete data likihood estimates); Black line is when we don't know the truth (observed likihood estimates). EM performed really well in this case.

Other Applications of EM

- Dealing with missing data:
- Multiple imputation of missing data: what missing data looks like under repeated sampling
- Truncated observations
- Bayesian hyperparameter estimation:when have a Bayesian model, treat unobverseed variables as latent variable (Empirical Bayes)
- Hidden Markov models: model data with some underlying state e.g. a Hidden Markov state: gene in/out of open readig frame (ORF).

EM Increases Likelihood

Since $\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$, it follows that

$$Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}).$$

which is a non-decreasing iteration

Also, by the properties of KL divergence (broader way to do numerical method) stated above (more details in Ch 43 in the reading), we have

$$\text{KL}(f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \| f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})) \geq \text{KL}(f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}) \| f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})).$$

Putting these together we have

$$\log f(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \geq \log f(\boldsymbol{x}; \boldsymbol{\theta}^{(t)}).$$

Markov Chain Monte Carlo

A very popular technique in Bayesian inference, but it's very slow and doesn't scale well to high-dimensional datasets; The tuning steps are very tricky and need a lot of experience.

Motivation

When performing Bayesian inferece, it is often (but not always) much easier to calculate

$$f(\boldsymbol{\theta}|\boldsymbol{x}) \propto L(\boldsymbol{\theta}; \boldsymbol{x}) f(\boldsymbol{\theta})$$

compared with

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{\theta}; \boldsymbol{x}) f(\boldsymbol{\theta})}{f(\boldsymbol{x})}$$

Say we have a plot of the

$$f(\boldsymbol{\theta}|\boldsymbol{x}) \propto L(\boldsymbol{\theta}; \boldsymbol{x}) f(\boldsymbol{\theta}).$$

There will be areas that are very high and very low. High areas are with high posterior probability. In order to take the product and turn into a probability distribution, we need the denominator, which is ofte nhard to calculate, that's where MCMC comes in.

MCMC tries to spend more time in high probability areas. If we have two θ values and I take the ratio of the two pdfs, now the mariginal disappears. Markov chain Monte Carlo is a method for simulating data approximately from $f(\boldsymbol{\theta}|\boldsymbol{x})$ with knowledge of only $L(\boldsymbol{\theta};\boldsymbol{x})f(\boldsymbol{\theta})$.

But once we get into high-dimensions, say the gaussian mixture model with multiple parameters, it would be hard to calculate by MCMC.

Note

MCMC can be used to approximately simulate data from any distribution that is only proportionally characterized, but it is probably most well know for doing so in the context of Bayesian infererence.

We will explain MCMC in the context of Bayesian inference.

Big Picture

We draw a Markov chain (a sequence of R.V. that the current variable only depends on the previous one) of $\boldsymbol{\theta}$ values so that, in some asymptotic sense, these are equivalent to iid draws from $f(\boldsymbol{\theta}|\boldsymbol{x})$.

The draws are done competitively so that the next draw of a realization of $\boldsymbol{\theta}$ depends on the current value.

The Markov chain is set up so that it only depends on $L(\boldsymbol{\theta};\boldsymbol{x})f(\boldsymbol{\theta})$.

A lot of practical decisions need to be made by the user, so utilize MCMC carefully.

Metropolis-Hastings Algorithm

1. Initialize $\boldsymbol{\theta}^{(0)}$
2. Generate $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ for some pdf or pmf $q(\cdot|\cdot)$ which we have to choose the distribution
3. With probablity
$$A(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(b)}) = \min\left(1, \frac{L(\boldsymbol{\theta}^*;\boldsymbol{x})f(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}^{(b)};\boldsymbol{x})f(\boldsymbol{\theta}^{(b)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(b)})}\right)$$
Set $\boldsymbol{\theta}^{(b+1)} = \boldsymbol{\theta}^*$. Otherise, set $\boldsymbol{\theta}^{(b+1)} = \boldsymbol{\theta}^{(b)}$

This is taking the ratio of the posterior pdf of new $\boldsymbol{\theta}^*$ vs. old $\boldsymbol{\theta}^{(b)}$, and weigh that by $\frac{q(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^b)}{q(\boldsymbol{\theta}^b|\boldsymbol{\theta}^{(*)})}$. The q function here is to introduce some stochastic behavior so it's not deterministic and stuck at local maxima.

4. Continue for $b = 1, 2, \ldots, B$ iterations and *carefully* select a subset of $\boldsymbol{\theta}^{(b)}$ which are utilized to approximate iid observations from $f(\boldsymbol{\theta}|\boldsymbol{x})$

Metropolis Algorithm

It's a simplification of Metropolis-Hastings Algorithm. The Metropolis algorithm restricts $q(\cdot,\cdot)$ to be symmetric so that $q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(b)})$ and

$$A(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(b)}) = \min\left(1, \frac{L(\boldsymbol{\theta}^*;\boldsymbol{x})f(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}^{(b)};\boldsymbol{x})f(\boldsymbol{\theta}^{(b)})}\right).$$

Utilizing MCMC Output

Two common uses of the output from MCMC are as follows: 1. The posterior expectation $\mathrm{E}[f(\boldsymbol{\theta})|\boldsymbol{x}]$ is approximated by

$$\hat{\mathrm{E}}[f(\boldsymbol{\theta})|\boldsymbol{x}] = \frac{1}{B}\sum_{b=1}^{B} f\left(\boldsymbol{\theta}^{(b)}\right).$$

2. Some subsequence $\boldsymbol{\theta}^{(b_1)}, \boldsymbol{\theta}^{(b_2)}, \ldots, \boldsymbol{\theta}^{(b_m)}$ from $\left\{\boldsymbol{\theta}^{(b)}\right\}_{b=1}^{B}$ is utilized as an empirical approximation to iid draws from $f(\boldsymbol{\theta}|\boldsymbol{x})$.

Remarks

- The random draw $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ perturbs the current value $\boldsymbol{\theta}^{(b)}$ to the next value $\boldsymbol{\theta}^{(b+1)}$. It is often a Normal distribution for continuous $\boldsymbol{\theta}$ with some μ and σ^2.
- Choosing the variance of $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ is important as it requires enough variance for the theory to be applicable within a reasonable number of computations, but it cannot be so large that new values of $\boldsymbol{\theta}^{(b+1)}$ are rarely generated.
- $A(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(b)})$ is called the acceptance probability.
- The algorithm must be run for a certain number of iterations ("burn in") before observed $\boldsymbol{\theta}^{(b)}$ can be utilized.

  -There's a burn in time to reach stable state.
- The generated $\boldsymbol{\theta}^{(b)}$ are typically "thinned" (only sampled every so often) to reduce Markov dependence.

  -If we want to approximate iid draws from the posterior, since $\boldsymbol{\theta}^{(b_2)}|\boldsymbol{\theta}^{(b_1)}, \boldsymbol{\theta}^{(b_3)}|\boldsymbol{\theta}^{(b_2)}, \ldots$, sometimes we need to do thinning, that is taking, say every 100th, θ.

Session Information

```
> sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS  10.15.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gapminder_0.3.0 forcats_0.4.0   stringr_1.4.0   dplyr_0.8.4
 [5] purrr_0.3.3     readr_1.3.1     tidyr_1.0.2     tibble_2.1.3
 [9] ggplot2_3.2.1   tidyverse_1.3.0 knitr_1.28
```

```
loaded via a namespace (and not attached):
 [1] tidyselect_1.0.0 xfun_0.12        haven_2.2.0      lattice_0.20-38
 [5] colorspace_1.4-1 vctrs_0.2.2      generics_0.0.2   htmltools_0.4.0
 [9] yaml_2.2.1       rlang_0.4.4      pillar_1.4.3     withr_2.1.2
[13] glue_1.3.1       DBI_1.1.0        dbplyr_1.4.2     modelr_0.1.5
[17] readxl_1.3.1     lifecycle_0.1.0  munsell_0.5.0    gtable_0.3.0
[21] cellranger_1.1.0 rvest_0.3.5      evaluate_0.14    fansi_0.4.0
[25] broom_0.5.4      Rcpp_1.0.3       scales_1.1.0     backports_1.1.5
[29] jsonlite_1.6.1   fs_1.3.1         hms_0.5.3        digest_0.6.23
[33] stringi_1.4.3    grid_3.6.0       cli_2.0.0        tools_3.6.0
[37] magrittr_1.5     lazyeval_0.2.2   crayon_1.3.4     pkgconfig_2.0.3
[41] xml2_1.2.2       reprex_0.3.0     lubridate_1.7.4  assertthat_0.2.1
[45] rmarkdown_2.1    httr_1.4.1       rstudioapi_0.11  R6_2.4.1
[49] nlme_3.1-143     compiler_3.6.0
```