10/17: Estimation

EM Algorithm

Suppose that

$$f(x \mid \theta) = \theta\, f_1(x) + (1-\theta)\, f_2(x)$$

How can we estimate θ.

EM Algorithm

Simple numerical techniques are fine here; likelihood function is concave.

However, illustrates EM algorithm

[on board]

Non-Existence of MLE

$$f(x \mid \theta) \;=\; \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

It must be that $\theta > X_{(}n) = \max_i X_i$.

So, MLE does not exist!

(You should feel cheated!)

Non-Uniqueness of MLE

$$f(x \mid \theta) = \begin{cases} 1 & \theta \le x \le \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{L}(\boldsymbol{X} \mid \theta) = \begin{cases} 1 & \max(X_i) - 1 \le \theta \le \min(X_i) \\ 0 & \text{otherwise} \end{cases}$$

Any θ in this interval is an MLE.

Non-Uniqueness of MLE

Take $X_i \sim \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(\mu, \sigma^2)$

$$\mathcal{L}(\boldsymbol{X} \mid \theta) = \left(\frac{1}{2}\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \left(e^{-\frac{X_i^2}{2}} + \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(X_i - \mu)^2}\right)$$

What happens?

Non-Existence of MLE

Let $\mu = X_1$ and $\sigma \to 0$.

$$\begin{aligned} \mathcal{L}(\boldsymbol{X} \mid \theta) &\ge \left(\frac{1}{2}\frac{1}{\sqrt{2\pi}}\right)^n \left(e^{-\frac{X_1^2}{2}} + \frac{1}{\sigma}\right) \prod_{i=2}^{n} e^{-\frac{X_i^2}{2}} \\ &\overset{\sigma \to 0}{\longrightarrow} \infty \end{aligned}$$

Non-Existence of MLE

Typically, non-existance involves the *boundary* of the domain of θ in some way.

Properties of the MLE

Invariance. If g is a 1-1 function, and $\hat{\theta}_n$ is the MLE of θ then $g(\hat{\theta}_n)$ is the MLE of $\tau = g(\theta)$.

We already used this for the normal.

Remember: $\hat{\theta}_n$ is a random variable.

Consistency

Under mild regularity conditions,

$$\hat{\theta}_n \longrightarrow \theta$$

in probability

Computation

Typically the MLE must be computed numerically

Computation

$X_i \sim \text{Gamma}(\alpha)$, $f(x \mid \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$.

$$\mathcal{L}(\boldsymbol{X} \mid \alpha) = \frac{1}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} X_i \right)^{\alpha-1} e^{-\sum_i X_i}$$

$$\ell(\boldsymbol{X} \mid \alpha) = (\alpha - 1) \sum_i \log X_i - n \log \Gamma(\alpha) - \sum_i X_i$$

Computation

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} = 0 \quad & \Rightarrow \\ 0 \quad & = \quad \sum_i \log X_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ \Psi(\alpha) \quad & = \quad \frac{1}{n}\sum_i \log X_i \end{aligned}$$

where $\Psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the *digamma function*.

We'll talk about solving such equations later.

Statistics

Any function of a sample:

$$T = g(X_1, X_2, \ldots, X_n)$$

Any estimator of θ is a statistic (not conversely).

Sufficient Statistics

Informally, a statistic $T = g(X_1, X_2, \ldots, X_n)$ is *sufficient* for θ if it contains all of the necessary information needed to estimate θ.

Best definition is in terms of information theory, using the notion of mutual information

Example

For a Bernoulli, we don't need to record the exact sequence of coin flips, only the number of heads.

$$g(X_1, \ldots, X_n) = \sum_{i=1} X_i$$

Sufficient Statistics

$T = g(X_1, \ldots, X_n)$ is a *sufficient statistic* in case

$$\mathrm{Pr}(X_1, X_2, \ldots, X_n \,|\, T, \theta)$$

does not in fact depend on θ. (Note that θ is not a random variable here.)

Example – Bernoulli

$$\begin{aligned} &\mathrm{Pr}(X_1, \ldots, X_n \,|\, \sum_i X_i = k, \theta) \\ &= \frac{\prod_{i=1}^n \theta_i^X (1-\theta)^{1-X_i}}{\sum_{X_1,\ldots,X_n : \sum X_i = k} \prod_{i=1}^n \theta_i^X (1-\theta)^{1-X_i}} \\ &= \frac{1}{\binom{n}{k}} \end{aligned}$$

not depending on θ.

Factorization Criterion

$T = g(X_1, X_2, \ldots, X_n)$ is sufficient for θ if and only if we can write

$$\prod_{i=1}^n f(X_i \,|\, \theta) \quad = \quad u(X_1, \ldots, X_n)\, v(T(X_1, \ldots, X_n), \theta)$$

So, the only part that depends on θ only depends on $X_1, X_2, \ldots, X_n$ through the "summary" T.

Poisson

What is a sufficient statistic for the Poisson parameter λ?

Poisson

$$
\begin{aligned}
f(k \mid \lambda) &= e^{-\lambda}\frac{\lambda^k}{k!} \\
\prod_{i=1}^{n} f(K_i \mid \lambda) &= e^{-n\lambda}\frac{\lambda^{\sum_i K_i}}{\prod_i K_i!} \\
&= \underbrace{\left(\frac{1}{\prod_i K_i!}\right)}_{u(\boldsymbol{K})} \underbrace{e^{-n\lambda}\lambda^{\sum_i K_i}}_{v(\lambda,T)}
\end{aligned}
$$

So we can take $T(K_1, \ldots, K_n) = \sum_i K_i$, as for the Bernoulli.

Gamma

What is a sufficient statistic for the Gamma parameter α?

Uniform

Take

$$
f(x \mid \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}
$$

What is a sufficient statistic?

MLE and Sufficient Statistics

Suppose we have a sufficient statistic T. Then

$$
\mathcal{L}(\boldsymbol{X} \mid \theta) = u(\boldsymbol{X})\, v(T(\boldsymbol{X}), \theta)
$$

To determine $\hat{\theta}_{ML}$, we only need $T(\boldsymbol{X})$.

Bayesian perspective

Consider θ as a random variable. We are *uncertain* about θ.

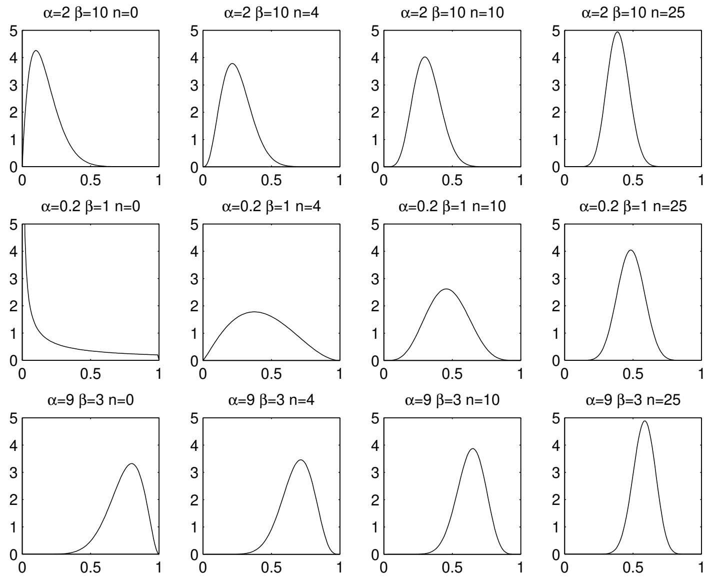
Bayesian Perspective

Under a *prior* $\Pr(\theta)$, the *posterior distribution* is given by

$$\Pr(\theta \mid \boldsymbol{X}) \quad \propto \quad \mathcal{L}(\boldsymbol{X} \mid \theta) \Pr(\theta)$$

The choice of prior $\Pr(\theta)$ is central, and the topic of great debate.

Example: Binomial with Beta Prior



Bayes Estimators

Recall our discussion of Loss and Utility.

$$\begin{aligned} \text{Risk}(a) &= \int_\Theta L(\theta, a) \Pr(\theta)\, d\theta \\ &= E[L(\theta, a)] \end{aligned}$$

Bayes Estimators

After observing some data $\boldsymbol{X}$ the risk is computed with respect to the new "updated" distribution (posterior)

$$\begin{aligned} \mathsf{Risk}(a \mid \boldsymbol{X}) &= \int_{\Theta} L(\theta, a) \, \mathsf{Pr}(\theta \mid \boldsymbol{X}) \, d\theta \\ &= E[L(\theta, a) \mid \boldsymbol{X}] \end{aligned}$$

Bayes Estimators

A *Bayes estimator* is an estimator of θ chosen to minimize the expected loss

$$\theta^*(\boldsymbol{X}) = \arg\min_a E[L(\theta, a) \mid \boldsymbol{X}]$$

Bayes Estimators – Special Case

Under the square loss $L(\theta, a) = (\theta - a)^2$, the Bayes estimator is the posterior mean:

$$\theta^*(\boldsymbol{X}) = E[\theta \mid X]$$