Parameter Estimation

Distances Between Distributions

1. **Total Variation Distance**: The total variation distance between distributions P and Q, with density functions p and q, is defined to be
$$TV(P,Q) = \sup_{A \subset E} |\mathbb{P}_p(A) - \mathbb{P}_q(A)|.$$
This is equivalent to half the L^1 distance between p and q:
$$TV(P,Q) = \frac{1}{2}\int_E |p(x) - q(x)|dx.$$
   (a) This is a genuine metric.
   (b) Unfortunately, it is hard to estimate.
2. **Kullback-Leibler Divergence**: The Kullback-Leibler (known as relative entropy in Information Theory) divergence between P and Q is defined to be
$$KL(P\|Q) = \int_E p(x)\log\Big(\frac{p(x)}{q(x)}\Big)dx,$$
where we assign the value ∞ if the support of p is *not* contained in the support of q (if it is, then anywhere $q = 0$, we will also have $p = 0$ and thus the points at which the integrand is not defined will all be removable discontinuities).
   (a) While positive semi-definite, KL-divergence is not a true metric, since it is anti-symmetric. It also fails to satify a trangle inequality.
   (b) It is, however, an expectation. Hence, it can be replaced with a sample mean and estimated.
      - Professor Rigollet calls the act of replating an expectation with a sample mean (i.e., the application of LLN) "the statistical hammer." The implication here is that it's our simplest (and often only) tool.

Examples

1. Let $X_n = \text{Poi}(1/n)$ and let δ_0 be a point mass centered at 0. Then $TV(X_n, \delta_0) \to 0$.
2. Let $P = \text{Bin}(n,p)$, $Q = \text{Bin}(n,q)$, where $p,q \in (0,1)$, and write their densities with one function
$$f(p,k) = \binom{n}{k}p^k(1-p)^{n-k},$$
and similarly for $f(n,q)$. Then it is actually a pretty straightforward calculation to show that
$$KL(P\|Q) = np\cdot\log\Big(\frac{p}{q}\Big) + (n-np)\cdot\log\Big(\frac{1-p}{1-q}\Big).$$

3. Let $P = N(a, 1)$ and let $Q = N(b, 1)$. Then (also pretty straightforward to calculate):
$$KL(P\|Q) = \frac{1}{2}(a-b)^2.$$

Maximum Likelihood Estimation

Definitions

1. Let $X_1, X_2, \ldots, X_n$ be an *iid* sample from a distribution with density $f(x;\theta)$. The *Likelihood* of the sample is
$$L(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(X_i;\theta).$$
2. The *log-likelihood* function, denoted $\ell(\theta)$ is
$$\ell(\theta) = \log(L(X_1, X_2, \ldots, X_n;\theta)).$$
Note we write ℓ as a random function of θ.
3. The *Fisher Information* is defined to be
$$I(\theta) = E\big[\nabla\ell(\theta)(\nabla\ell(\theta))^T\big] - E\big[\nabla\ell(\theta)\big]E\big[\nabla\ell(\theta)\big]^T = -E\big[\mathbf{H}\ell(\theta)\big],$$
where in this case the likelihood is of a one-element sample, and the bold $\mathbf{H}$ denotes the Hessian operator. In one dimension, this reduces to
$$I(\theta) = -E(\ell''(\theta)).$$
Equivalently, we also have
$$I(\theta) = Var(\ell'(\theta)).$$
This latter definition is usually harder to work with, but has a more direct connection to maximum likelihood estimators.

Throughout, we will be discussing ways to estimate the value of a "true" parameter θ^* of a distribution $\mathbb{P}_{\theta^*}$, given a model $(E, \{\mathbb{P}_\theta : \theta \in \Theta\})$. A noble goal might be to build an estimator $\widehat{TV}(P_\theta, P_{\theta^*})$ and compute the argmin using this estimator. However, TV distance is hard to estimate in general, so we use KL-divergence instead. Since this function is an expectation, it can be replaced by a sample mean (using LLN), and is therefore easy to estimate.

For the rest of this section, suppose we are estimating a distribution $\mathbb{P} = \mathbb{P}_{\theta^*}$ with a parametric family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$. We will proceed to do this by estimating the minimizer (argmin) of $KL(\mathbb{P}_\theta, \mathbb{P})$, which is θ^*, by the positive semidefiniteness (or nonnegative definiteness?) of KL.

The strategy for doing so will involve first estimating KL divergence and finding the minimizer of that estimator $\widehat{KL}$. That the argmin of $\widehat{KL}$ converges to the argmin of KL follows from "nice analytic properties" of these functions. I'm guessing that KL is at least C^1 and the convergence is relatively strong.

Estimating KL Divergence

Recall that $KL(\mathbb{P}_\theta, \mathbb{P})$ is an expectation: if f_θ and f are the densities of $\mathbb{P}_\theta$ and $\mathbb{P}$, respectively, then

$$KL(\mathbb{P}, \mathbb{P}_\theta) = E_{\mathbb{P}}\left(\log\left(\frac{f(x)}{f_\theta(x)} \right) \right) = E_{\mathbb{P}}(\log(f(x))) - E_{\mathbb{P}}(\log(f_\theta(x))).$$

As a function $\theta \mapsto KL(f_\theta, f)$, this has the form

$$KL(\mathbb{P}, \mathbb{P}_\theta) = \text{"constant"} - E_{\mathbb{P}}(\log(f_\theta(x))).$$

Thus, by LLN, we have

$$\widehat{KL}(\mathbb{P}, \mathbb{P}_\theta) = \text{"constant"} - \frac{1}{n}\sum_{i=1}^{n} \log(f_\theta(X_i)).$$

Finding the Minimum of $\widehat{KL}$

Starting with the above equation, we have

$$\begin{aligned}
\min_{\theta\in\Theta} \widehat{KL}(\mathbb{P}, \mathbb{P}_\theta) &\Leftrightarrow \min_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^{n} \log(f_\theta(X_i)) \\
&\Leftrightarrow \max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log(f_\theta(X_i)) \\
&\Leftrightarrow \max_{\theta\in\Theta} \log\left(\prod_{i=1}^{n} p_\theta(X_i) \right) \\
&\Leftrightarrow \max_{\theta\in\Theta} \prod_{i=1}^{n} p_\theta(X_i).
\end{aligned}$$

Therefore, the minimizer of $\widehat{KL}$ is the maximum likelihood estimator $\hat{\theta}$ of θ^*. Furthermore (avoiding a bunch of details necessary for this implication), we have

$$\hat{\theta} \xrightarrow{(p)} \theta^*.$$

Examples of Maximum Likelihood Estimators

1. **TODO**

The Asymptotic Variance of MLE

The MLE is not only consistent, but also satisfies a central limit theorem:

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{(d)} N\big(0, V(\theta^*)\big),$$

where $V(\theta^*)$ represents the asymptotic variance of $\hat{\theta}$. But what is this asymptotic variance?!? Turns out that under some mild conditions, the asymptotic variance of $\hat{\theta}$ is known.

Theorem Assume the following.

1. θ^* is identifiable.
2. θ^* is an interior point of Θ.
3. The Fisher information matrix $I(\theta)$ is invertible in a neighborhood of θ^*.
4. All the functions involved are "nice".
5. The support of $\mathbb{P}_\theta$ does not depend on θ.

Then

$$V(\hat{\theta}) = I(\theta^*)^{-1}.$$

Proof Write $\ell_i(\theta) = \log f_\theta(X_i)$. We start with a couple of observations:

1. Since $\hat{\theta}$ is the unique maximizer of $\log(L_n(X_1, X_2, \ldots, X_n; \theta))$,

$$\frac{d}{d\theta}\bigg|_{\theta=\hat{\theta}} \sum_{i=1}^n \ell_i = \sum_{i=1}^n \ell_i(\hat{\theta}) = 0.$$

2. Since θ^* is the unique minimizer of $KL(\mathbb{P}_\theta, \mathbb{P})$ and this differs from $E(\ell(\theta))$ by a constant, we have

$$E(\ell'(\theta^*)) = \frac{d}{d\theta}\bigg|_{\theta=\theta^*} E(\ell(\theta)) = 0.$$

Now, we start with a Taylor expansion at θ^*:

$$0 = \sum_{i=1}^n \ell_i'(\hat{\theta}) = \sum_{i=1}^n \big[\ell_i'(\theta^*) + (\hat{\theta} - \theta^*)\ell_i''(\theta^*) + \cdots\big].$$

Therefore scaling and applying observation 1, we have

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[\big(\ell_i'(\theta^*) - E[\ell_i'(\theta^*)]\big) + (\hat{\theta} - \theta^*)\ell_i''(\theta^*) + \cdots\big] \\
&\approx \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(\ell_i'(\theta^*) - E[\ell_i'(\theta^*)]\big) + \frac{1}{\sqrt{n}} \sum_{i=1}^n (\hat{\theta} - \theta^*)\ell_i''(\theta^*) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(\ell_i'(\theta^*) - E[\ell_i'(\theta^*)]\big) + \sqrt{n}(\hat{\theta} - \theta^*) \cdot \frac{1}{n} \sum_{i=1}^n \ell_i''(\theta^*).
\end{aligned}$$

By CLT, the term on the left converges to $N(0, I(\theta^*))$. By LLN, the term $n^{-1}\sum_i \ell_i''(\theta^*)$ converges to $E(\ell''(\theta^*)) = -I(\theta^*)$. Therefore, rearranging, we have

$$I(\theta^*) \cdot \sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{(d)} N(0, I(\theta^*)),$$

therefore,

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{(d)} N(0, I(\theta^*)^{-1}).$$

Remark: This proof only works in one dimension. In higher dimensions, there is a lack of commutativity that results in a more complicated expression in the end. (**TODO:** Write up something about it)

Remark: Recall the original definition of Fisher information as the Hessian of log-likelihood. This adds geometric intuition to the result: If the log-likelihood is more tightly curved at θ^*, then MLE will vary less around the maximum and vice versa. The word "information" is also more than superficial with this in mind; i.e., more "information" iff less variance, which translates to tighter confidence intervals around MLE.

Examples of Fisher Information

1. **TODO: Match the MLE examples above.**

Method of Moments

M-Estimation