Class Notes (experimental)

Jonathan Rosenblatt

April 7, 2015

1 Estimation

In this section, we present several estimation principeles. Their properties are not discussed, as the section is merely a reminder and a preparation for Section 2.

1.1 Moment matching

The fundamental idea: match empirical moments to theoretical. I.e., estimate

$$E[g(X)]$$

by

$$\mathbb{E}[g(X)]$$

where $\mathbb{E}[g(X)] := \frac{1}{n}\sum_i g(X_i)$, is the empirical mean.

Example 1 (Exponential Rate)**.** Estimate λ in $X_i \sim exp(\lambda)$, $i = 1, \dots, n$, i.i.d. $E[X] = 1/\lambda$. $\Rightarrow \hat{\lambda} = 1/\mathbb{E}[X]$

Example 2 (Linear Regression)**.** Estimate β in $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, a p dimensional random vector. $E[Y] = X\beta$ and $\mathbb{E}[Y] = y$. Clearly, moment mathing won't work because no β satistifes $X\beta = Y$. A technical workaround: Since β is p dimensional, I need to find some $g(Y) : \mathbb{R}^n \mapsto \mathbb{R}^p$. Well, $g(Y) := XY$ is such a mapping. I will use it, even though my technical justification is currently unsatisfactory. We thus have: $E[X'Y] = X'X\beta$ which I match to $\mathbb{E}[X'Y] = X'y$:

$$X'X\beta = X'y \Rightarrow \hat{\beta} = (X'X)^{-1}X'y.$$

1.2 Quantile matching

The fundamental idea: match empirical quantiles to theoretical. Denoting by $F_X(t)$ the CDF of X, then $F_X^{-1}(\alpha)$ is the α quantile of X. Also denoting by $\mathbb{F}_X(t)$ the Empirical CDF of $X_1, \dots, X_n$, then $\mathbb{F}_X^{-1}(\alpha)$ is the α quantile of $X_1, \dots, X_n$. The quantile matching method thus implies estimating

$$F_X^{-1}(\alpha)$$

by

$$\mathbb{F}_X^{-1}(\alpha).$$

Example 3 (Exponential rate)**.** Estimate λ in $X_i \sim exp(\lambda)$, $i = 1, \dots, n$, i.i.d.

$$\begin{aligned} F_X(t) =& 1 - \exp(-\lambda t) = \alpha \Rightarrow \\ F_X^{-1}(\alpha) =& \frac{-\log(1-\alpha)}{\lambda} \Rightarrow \\ F_X^{-1}(0.5) =& \frac{-\log(0.5)}{\lambda} \Rightarrow \\ \hat{\lambda} =& \frac{-\log(0.5)}{\mathbb{F}_X^{-1}(0.5)}. \end{aligned}$$

1.3 Maximum Likelihood

The fundamental idea is that if the data generating proces (i.e., the *sampling distribution*) can be assumed, then the observations are probably some high probability instance of this process, and not a low probability event: Let $X_1, \dots, X_n \sim P_\theta$, with density (or probability) $p_\theta(X_1, \dots, X_n)$. Denote the likelihood, as a function of θ: $\mathcal{L}(\theta) : p_\theta(X_1, \dots, X_n)$. Then $\hat{\theta}_{ML} := argmax_\theta\{\mathcal{L}(\theta)\}$.

Example 4 (Exponential rate)**.** Estimate λ in $X_i \sim exp(\lambda)$, $i = 1, \dots, n$, i.i.d. Using the exponential PDF and the i.i.d. assumption

$$\mathcal{L}(\lambda) = \lambda^n \exp(-\lambda \sum_i X_i).$$

Using a monotone mapping such as the log, does not change the *argmax*. Denoting $L(\theta) := \log(\mathcal{L}(\theta))$, we have

$$L(\lambda) = n \log(\lambda) - \lambda \sum_i X_i.$$

By differentiating and equating 0, we get $\hat{\lambda}_{ML} = 1/\mathbb{E}[X]$.

Example 5 (Discrete time Markov Chain). Estimate the transition probabilities, p_1 and p_2 in a two state, $\{0,1\}$, discrete time, Markov chain where: $P(X_{t+1}=1|X_t=0)=p_1$ and $P(X_{t+1}=1|X_t=1)=p_2$. The likelihood:

$$\mathcal{L}(p_1,p_2)=P(X_1,\dots,X_n;p_1,p_2)=\prod_{t=0}^{T}P(X_{t+1}=x_{t+1}|X_t=x_t).$$

We denote n_{ij} the number of observed transitions from i to j and get that $\hat{p}_1=\frac{n_{01}}{n_{01}+n_{00}}$, and that $\hat{p}_2=\frac{n_{11}}{n_{11}+n_{10}}$.

Remark 1. Well, this is a rather artificial example, as because of the Markov property, and the stationarity of the process, we only need to look at transition events, themselves Brenoulli distributed. This example does show, however, the power of the ML method to deal with non i.i.d. samples. As does the next example.

Example 6 (Brownian motion with drift). Estimate the drift parameter a, in a discrete time Gaussian process where: $X_{t+1}=X_t+\varepsilon;\varepsilon\sim\mathcal{N}(0,\sigma^2)\Rightarrow X_{t+1}|X_t\sim\mathcal{N}(aX_t,\sigma^2)$.

We start with the conditional density at time $t+1$:

$$p_{X_{t+1}|X_t=x_t}(x_{t+1})=(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(x_{t+1}-ax_t)^2\right).$$

Moving to the likelihood:

$$\mathcal{L}(a)=(2\pi\sigma^2)^{-T/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(x_{t+1}-ax_t)^2\right).$$

Differentiating with respect to a and equating 0 we get $\hat{a}_{ML}=\frac{\sum x_{t+1}x_t}{\sum x_t^2}$.

We again see the power of the ML device. Could we have arrive to this estimator by intuiton alone? Hmmmm... maybe. See that $Cov[X_{t+1},X_t]=a\ Var[X_t]\Rightarrow a=\frac{Cov[X_{t+1},X_t]}{Var[X_t]}$. So a can also be derived using the moment matching method which is probably more intuitive.

Example 7 (Linear Regression). Estimate β in $Y\sim\mathcal{N}(X\beta,\sigma^2I)$, a p dimensional random vector. Recalling the multivariate Gaussian PDF:

$$p_{\mu,\Sigma}(y)=(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

So in the regression setup:

$$\mathcal{L}(\beta)=p_{\beta,\sigma^2}(y)=(2\pi)^{-n/2}|\sigma^2I|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\|y-X\beta\|^2\right)$$

1.4 M-Estimation and Empirical Risk Minimization

M-Estimation, know as Empirical Risk Minimizaton (ERM) in the machine learning literature, is a very wide framework which stems from statistical desicion theory. The underlying idea is that each realization of X incurs some loss, and we seek to find a "policy", in this case a parameter, θ^* that minimizes the average loss. In the econometric literature, we dot not incur a loss, but rather a utility, we thus seek a policy that maximizes the average utility.

Define a loss function $l(X;\theta)$, and a risk function, being the expected loss, $R(\theta) := E[l(X;\theta)]$. Then

$$\theta^* := argmin_\theta\{R(\theta)\}. \tag{1}$$

As we do not know the distribution of X, we cannot solve Eq.(1), so we minimize the *empirical* risk. Define the empirical risk as $\mathbb{R}(\theta) := \mathbb{E}[l(X;\theta)]$, then

$$\hat{\theta} := argmin_\theta\{\mathbb{R}(\theta)\}. \tag{2}$$

Example 8 (Squared Loss)**.** Let $l(X;\theta) = (X-\theta)^2$. Then $R(\theta) = E[(X-\theta)^2] = (E[X]-\theta)^2 + Var[X]$. Clearly $Var[X]$ does not depend on θ so that $R(\theta)$ is minimized by $\theta^* = E[X]$. **We thus say that the expectation of a random variable is the minimizer of the squared loss.**

How do we estimate the population expectation? Well a natural estimator is the empirical mean, which is also the minimizer of the empirical risk $\mathbb{R}(X)$. The proof is immediate by differentiating.

2 From Estimation to Learning

2.1 Empirical Risk Minimization (ERM) and Inductive Bias

2.2 Linear Regression (OLS)

2.3 Ridge Regression

2.4 Logistic Regression

2.5 LASSO

2.6 Linear SVM

2.7 Generalized Additive Models (GAMs)

2.8 Neural Nets (NNETs)

2.9 Classification and Regression Trees (CARTs)