

## Lecture 3: Density Estimation

**Note:** *LaTeX template courtesy of UC Berkeley EECS dept.*

**Disclaimer:** *These notes are adapted from ETH's Advanced Machine Learning Course and the book All Of Statistics, Larry Wasserman, Springer.*

### 3.1 Parametric Inference

We now turn our attention to parametric models, that is, models of the form:

$$\mathfrak{F} = \{f(x; \theta) : \theta \in \Theta\}$$

where the  $\theta \in \mathbb{R}^k$  is the parameter space and  $\theta = (\theta_1, \dots, \theta_k)$  is the parameter. The problem of inference then reduces to the problem of estimating the parameter  $\theta$ .

Often, we are only interested in some function  $T(\theta)$ . For example, if  $X \sim \mathcal{N}(\mu, \sigma^2)$  then the parameter is  $\theta = (\mu, \sigma)$ . If our goal is to estimate  $\mu$  then  $\mu = T(\theta)$  is called the parameter of interest and  $\sigma$  is called a nuisance parameter.

#### 3.1.1 Maximum Likelihood

The most common method for estimating parameters in a parametric model is the maximum likelihood method. Let  $X_1, \dots, X_n$  be IID with pdf  $f(x; \theta)$ .

**Definition 3.1** The *likelihood function* is defined by:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

The *log-likelihood function* is defined by  $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$ .

The likelihood function is just the joint density of the data, except that we treat it as a function of the parameter  $\theta$ . Thus,  $\mathcal{L}_n(\theta) : \Theta \rightarrow [0, \infty)$ . The likelihood function is not a density function: in general, it is not true that  $\mathcal{L}_n(\theta)$  integrates to 1 (with respect to  $\theta$ ).

**Definition 3.2** The *maximum likelihood estimator MLE*, denoted by  $\hat{\theta}_n$ , is the value of  $\theta$  that maximizes  $\mathcal{L}_n(\theta)$ .

The maximum of  $\ell_n(\theta)$  occurs at the same place as the maximum of  $\mathcal{L}_n(\theta)$ , so maximizing the log-likelihood leads to the same result as maximizing the likelihood. Often, it is easier to work with the log-likelihood.

**Claim 3.3** If we multiply  $\mathcal{L}_n(\theta)$  by any positive constant  $c$  (not depending on  $\theta$ ) then this will not change the MLE. Hence, we shall often drop constants in the likelihood function.

**Example 3.1** Let  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ . The parameter is  $\theta = (\mu, \sigma)$  and the likelihood function (ignoring some constants) is:

$$\begin{aligned}\mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp \left\{ -\frac{nS^2}{2\sigma^2} \right\} \exp \left\{ -\frac{n(\bar{X} - \mu)^2}{2\sigma^2} \right\}\end{aligned}$$

where  $\bar{X} = n^{-1} \sum_i X_i$  is the sample mean and  $S^2 = n^{-1} \sum_i (X_i - \bar{X})^2$ . The last equality above follows from the fact that  $\sum_i (X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$  which can be verified by writing  $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2$  and then expanding the square. The log-likelihood is:

$$l(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}$$

Solving the equations:

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial l(\mu, \sigma)}{\partial \sigma} = 0,$$

we conclude that  $\hat{\mu} = \bar{X}$  and  $\hat{\sigma} = S$ . It can be verified that these are indeed global maxima of the likelihood.

### 3.1.2 Properties of Maximum Likelihood Estimators

Under certain conditions on the model, the maximum likelihood estimator  $\hat{\theta}_n$  possesses many properties that make it an appealing choice of estimator. The main properties of the MLE are:

1. The MLE is **consistent**:  $\hat{\theta}_n \xrightarrow{P} \theta^*$  where  $\theta^*$  denotes the true value of the parameter  $\theta$ ;
2. The MLE is **equivariant**: if  $\hat{\theta}_n$  is the MLE of  $\theta$  then  $g(\hat{\theta}_n)$  is the MLE of  $g(\theta)$ ;
3. The MLE is **asymptotically Normal**:  $(\hat{\theta} - \theta^*)/\hat{\text{se}} \sim \mathcal{N}(0, 1)$ ; also, the estimated standard error  $\hat{\text{se}}$  can often be computed analytically;
4. The MLE is **asymptotically optimal** or **efficient**: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large samples. That is,  $\hat{\theta}_n$  minimizes  $\mathbb{E}[(\hat{\theta}_n - \theta^*)^2]$  as  $n \rightarrow \infty$ ;
5. The MLE is approximately the Bayes estimator.

The properties we discuss only hold if the model satisfies certain regularity conditions. These are essentially smoothness conditions on  $f(x; \theta)$ , unless otherwise stated we shall tacitly assume that these conditions hold.

### 3.1.3 Understanding Asymptotic efficiency

The expected square error is a measure for quantifying how good an estimator  $\hat{\theta}$  is:

$$\mathbb{E}[(\hat{\theta} - \theta_0)^2]$$

The Rao-Cramer bound shows that there does not exist an estimator that reaches  $\mathbb{E}[(\hat{\theta} - \theta_0)^2] = 0$

**Theorem 3.4** For any estimator  $\hat{\theta}$  of  $\theta$  it holds that:

$$\mathbb{E}_{x|\theta}[(\hat{\theta} - \theta)^2] \geq \frac{\left(\frac{\partial}{\partial \theta} b_{\hat{\theta}} + 1\right)^2}{\mathbb{E}_{x|\theta}[\Lambda^2]} + b_{\hat{\theta}}^2$$

Where:

$$\Lambda = \frac{\partial}{\partial \theta} \log p(x|\theta) = \frac{1}{p(x|\theta)} \frac{\partial}{\partial \theta} p(x|\theta) \quad \text{and} \quad b_{\hat{\theta}} = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta$$

**Proof:**

$$\mathbb{E}_{x|\theta}[\Lambda] = \int_x p(x|\theta) \Lambda \, dx = \int_x \frac{\partial}{\partial \theta} p(x|\theta) \, dx = \frac{\partial}{\partial \theta} \int_x p(x|\theta) \, dx \stackrel{=1}{=} 0$$

$$\mathbb{E}_{x|\theta}[\Lambda \hat{\theta}] = \int_x p(x|\theta) \Lambda \hat{\theta} \, dx = \int_x \frac{\partial}{\partial \theta} p(x|\theta) \hat{\theta} \, dx = \frac{\partial}{\partial \theta} \int_x p(x|\theta) \hat{\theta} \, dx = \frac{\partial}{\partial \theta} \mathbb{E}_{x|\theta}[\hat{\theta}] = \frac{\partial}{\partial \theta} (\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta) + 1 = \frac{\partial}{\partial \theta} b_{\hat{\theta}} + 1$$

Consider the covariance between  $\Lambda$  and  $\hat{\theta}$ :

$$\left( \mathbb{E}_{x|\theta} \left[ \left( \Lambda - \overbrace{\mathbb{E}_{x|\theta}[\Lambda]}^{=0} \right) (\hat{\theta} - \mathbb{E}_{x|\theta}[\hat{\theta}]) \right] \right)^2 = \left( \mathbb{E}_{x|\theta}[\Lambda \hat{\theta}] - \mathbb{E}_{x|\theta}[\Lambda] \mathbb{E}_{x|\theta}[\hat{\theta}] \right)^2 = \left( \mathbb{E}_{x|\theta}[\Lambda \hat{\theta}] - \overbrace{\mathbb{E}_{x|\theta}[\Lambda]}^{=0} \mathbb{E}_{x|\theta}[\hat{\theta}] \right)^2 = \left( \mathbb{E}_{x|\theta}[\Lambda \hat{\theta}] \right)^2$$

Now, let's consider Cauchy-Schwarz inequality i.e.  $(\mathbb{E}[xy])^2 \leq \mathbb{E}[x^2] \mathbb{E}[y^2]$  applied to the cross-correlation:

$$\begin{aligned} \left( \mathbb{E}_{x|\theta} \left[ \left( \Lambda - \overbrace{\mathbb{E}_{x|\theta}[\Lambda]}^{=0} \right) (\hat{\theta} - \mathbb{E}_{x|\theta}[\hat{\theta}]) \right] \right)^2 &\leq \mathbb{E}_{x|\theta}[\Lambda^2] \mathbb{E}_{x|\theta}[(\hat{\theta} - \mathbb{E}_{x|\theta}[\hat{\theta}])^2] = \mathbb{E}_{x|\theta}[\Lambda^2] \mathbb{E}_{x|\theta}[(\hat{\theta} - \theta - (\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta))^2] \\ &= \mathbb{E}_{x|\theta}[\Lambda^2] \mathbb{E}_{x|\theta}[(\hat{\theta} - \theta)^2 + (\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta)^2 - 2(\hat{\theta} - \theta)(\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta)] \\ &= \mathbb{E}_{x|\theta}[\Lambda^2] \left\{ \mathbb{E}_{x|\theta}[(\hat{\theta} - \theta)^2] + \overbrace{\mathbb{E}_{x|\theta}[(\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta)^2 - 2(\hat{\theta} - \theta)(\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta)]}^{-b_{\hat{\theta}}^2} \right\} = \mathbb{E}_{x|\theta}[\Lambda^2] \left\{ \mathbb{E}_{x|\theta}[(\hat{\theta} - \theta)^2] - b_{\hat{\theta}}^2 \right\} \end{aligned}$$

It's easy to verify that  $\mathbb{E}_{x|\theta}[(\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta)^2 - 2(\hat{\theta} - \theta)(\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta)] = -b_{\hat{\theta}}^2$ :

$$\begin{aligned} &\mathbb{E}_{x|\theta} \left[ \mathbb{E}_{x|\theta}^2[\hat{\theta}] + \theta^2 - \cancel{2\theta \mathbb{E}_{x|\theta}[\hat{\theta}]} - 2\hat{\theta} \mathbb{E}_{x|\theta}[\hat{\theta}] + 2\hat{\theta} \theta + \cancel{2\theta \mathbb{E}_{x|\theta}[\hat{\theta}]} - 2\theta^2 \right] \\ &= \mathbb{E}_{x|\theta}^2[\hat{\theta}] + \mathbb{E}_{x|\theta}[\theta^2] - 2\mathbb{E}_{x|\theta}^2[\hat{\theta}] + 2\theta \mathbb{E}_{x|\theta}[\hat{\theta}] - 2\mathbb{E}_{x|\theta}[\theta^2] \\ &= -\mathbb{E}_{x|\theta}^2[\hat{\theta}] + \mathbb{E}_{x|\theta}[\theta^2] + 2\theta \mathbb{E}_{x|\theta}[\hat{\theta}] = -\mathbb{E}_{x|\theta}^2[\hat{\theta}] - \theta^2 + 2\theta \mathbb{E}_{x|\theta}[\hat{\theta}] = -(\mathbb{E}_{x|\theta}[\hat{\theta}] - \theta)^2 = -b_{\hat{\theta}}^2 \end{aligned}$$

Finally, from the inequality proved earlier we know that:

$$(\mathbb{E}_{x|\theta}[\Lambda \hat{\theta}])^2 = \left( \frac{\partial}{\partial \theta} b_{\hat{\theta}} + 1 \right)^2 \leq \mathbb{E}_{x|\theta}[\Lambda^2] \mathbb{E}_{x|\theta}[(\hat{\theta} - \theta)^2 - b_{\hat{\theta}}^2]$$

It follows that:

$$\mathbb{E}_{x|\theta}[(\hat{\theta} - \theta)^2] \geq \frac{\left(\frac{\partial}{\partial \theta} b_{\hat{\theta}} + 1\right)^2}{\mathbb{E}_{x|\theta}[\Lambda^2]} + b_{\hat{\theta}}^2$$

■

### 3.1.4 Stein Estimator

For finite samples, the maximum-likelihood estimator is not necessarily efficient.

Consider a multivariate random variable with distribution  $\mathcal{N}(\theta_0, \sigma^2 I)$  with range  $\mathbb{R}^d$  and  $d \geq 3$ . If we sample a single point  $y$  from this distribution then the Stein Estimator is:

$$\hat{\theta}_{JS} := \left(1 - \frac{(d-2)\sigma^2}{\|y\|^2}\right)y$$

It is possible to prove that the Stein Estimator is better than the maximum-likelihood estimator for any  $\theta_0$ . That is:

$$\mathbb{E}\left[(\hat{\theta}_{JS} - \theta_0)^2\right] \leq \mathbb{E}\left[(\hat{\theta}_{ML} - \theta_0)^2\right] \text{ for any } \theta_0$$

Moreover, the inequality is strict for some values of  $\theta_0$ .

## 3.2 Bayesian Learning

Bayesian inference is usually carried out in the following way:

- $\theta$  is considered to be a **random variable** with distribution  $p(\theta|\mathcal{X})$ .
- $X \sim p(x)$  and  $p(x)$  is unknown.
- $p(x|\theta)$  is a statistical model that reflects our beliefs about  $x$  given  $\theta$ .

We are looking for  $p(X = x|\mathcal{X})$ , i.e., the probability of  $x$  given the sample set  $\mathcal{X}$  (class conditional density):

$$p(X = x|\mathcal{X}) = \int \underbrace{p(x, \theta|\mathcal{X})}_{p(x|\theta, \mathcal{X})p(\theta|\mathcal{X})} d\theta = \int p(x|\theta, \mathcal{X})p(\theta|\mathcal{X})d\theta = \int p(x|\theta)p(\theta|\mathcal{X})d\theta$$

Where  $p(x|\theta, \mathcal{X}) = p(x|\theta)$  since  $x_i \in \mathcal{X}$  and  $x$  are i.i.d.

Moreover, asymptotically it holds that  $p(\theta|\mathcal{X}) \sim \delta(\theta - \hat{\theta})$ ; intuitively, this follows from the fact that  $\hat{\theta} \xrightarrow{P} \theta_{true}$ . Thus, in the asymptotic case, we can approximate the integral with:

$$p(X = x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X})d\theta \approx \int p(x|\theta)\delta(\theta - \hat{\theta})d\theta = p(x|\hat{\theta})$$

This approximation was used in the early days of Bayesian inference when it was not possible to evaluate the integral.

### 3.2.1 Bayesian Learning of a Normal Distribution

Let us begin with a simple example in which we consider a single Gaussian random variable  $x$ . We shall suppose that the variance  $\sigma^2$  is known, and we consider the task of inferring the mean  $\mu$  given a set of  $N$  observations:

- The likelihood is  $p(x|\mu) = \mathcal{N}(\mu, \sigma^2)$
- The prior is  $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$
- The data is  $\mathcal{X} = \{x_1, \dots, x_n\}$

We want to compute the posterior distribution  $p(\mu|\mathcal{X})$ :

$$\begin{aligned} p(\mu|\mathcal{X}) &\propto p(\mathcal{X}|\mu)p(\mu) \implies p(\mu|\mathcal{X}) = \alpha \prod_{i \leq n} \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \right\} \cdot \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right) \\ &= \alpha' \cdot \prod_{i \leq n} \left\{ \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \right\} \cdot \exp\left(-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right) = \alpha' \cdot \exp\left\{-\frac{1}{2}\left(\sum_{i \leq n} \left(\frac{x_i - \mu}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right)\right\} \end{aligned}$$

Expanding the squares we get:

$$p(\mu|\mathcal{X}) = \alpha' \cdot \exp\left(\overbrace{\mu^2 \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}^a - 2\mu \overbrace{\left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2} \sum_{i \leq n} x_i^2\right)}^b + c\right)$$

Which we know is a Gaussian Distribution, i.e.  $p(\mu|\mathcal{X}) \sim \mathcal{N}(\mu_n, \sigma_n^2)$ , because the exponent is a quadratic form. Furthermore, by completing the square we know that:

$$\begin{aligned} \mu_n = \frac{b}{a} &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \\ \sigma_n^2 = \frac{1}{a} &= \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2} \end{aligned}$$

It is worth spending a moment studying the form of the posterior mean and variance. First of all, note that the mean of the posterior is a compromise between  $\mu_0$  and the maximum likelihood solution  $\hat{\mu}$ . If the number of observed data points  $n = 0$  then  $\mu_n$  reduces to the prior mean as expected. For  $n \rightarrow \infty$ ,  $\mu_n$  is given by the maximum likelihood solution.

Similarly, consider the result for the variance of the posterior distribution  $\sigma_n^2$ . With no observed data points, we have the prior variance, whereas if the number of data points  $n \rightarrow \infty$ , the variance goes to zero and the posterior distribution becomes infinitely peaked around the maximum likelihood solution.

We therefore see that the maximum likelihood result of a point estimate for  $\mu$  is recovered precisely from the Bayesian formalism in the limit of an infinite number of observations.