

# Nonparametric Estimation

Nonparametric estimation refers to estimation of signals, regression functions, or probability distributions that do not admit a fixed parameterization.

## 1 Nonparametric Signal Estimation

Consider the standard problem of estimating a signal  $S$  given noisy data

$$Y_i = S_i + N_i, \quad 1 \leq i \leq n \quad (1)$$

where  $N_i$  are iid  $\mathcal{N}(0, \sigma^2)$ . As seen in previous lectures, this problem can be approached using

- Bayesian estimation, in case a prior on the signal is available; or
- maximum-likelihood (ML) estimation, in case the signal admits a parametric form.

### 1.1 Failure of ML

Assume no prior is available. A simple but naive idea is to view each individual sample  $S_i$  as an unknown parameter, and apply the ML principle to estimate them. The ML estimator for the problem (1) is

$$\begin{aligned} \hat{\underline{s}}_{ML} &= \arg \max_{\underline{s} \in \mathbb{R}^n} \ln p(\underline{y} | \underline{s}) \\ &= \arg \max_{\underline{s} \in \mathbb{R}^n} \sum_{i=1}^N \ln p_N(y_i - s_i). \end{aligned}$$

Since the cost function is additive over  $\{s_i\}$ , the maximization problem over  $\underline{s}$  reduces to  $n$  independent maximization problems:

$$\begin{aligned} \hat{s}_{ML,i} &= \arg \max_{s_i} \ln p_N(y_i - s_i) \\ &= y_i, \quad 1 \leq i \leq n \end{aligned}$$

where the last line holds because the noise pdf  $p_N$  has its maximum at 0. Hence the ML estimator of  $\underline{s}$  is the noisy data  $\underline{y}$  themselves. ML fails because the number of data ( $n$ ) is not large enough relative to the number of parameters ( $n$  also): the usual justification of ML using asymptotic arguments does not apply here.

## 1.2 Filtering

An example of nonparametric estimation for the model (1) is simple filtering of the noisy data  $Y_i$ , using a lowpass filter  $h_i$ :

$$\hat{s}_i = (h \star Y)_i = \sum_j h_j Y_{i-j}, \quad 1 \leq i \leq n.$$

The estimator is characterized by the smoothing filter chosen, and most importantly by the bandwidth of the filter. We shall not further elaborate on this method as it lacks flexibility and is not easily generalized to models that are more complex than (1).

## 1.3 Semiparametric Estimation

Nonparametric signal estimation cannot be successful unless the set of signals is somehow constrained. In Bayesian estimation, the prior penalizes unlikely signals and may be thought of as a soft constraint.

The semiparametric approach presented here may be thought of as introducing a hard constraint: approximate the signal with a model characterized by  $m$  parameters  $\theta_j$ ,  $1 \leq j \leq m$ , and use ML to estimate these parameters. Now we have two conflicting requirements:

- To properly approximate the signal, the model should capture the presumably complex characteristics of the signal, and thus  $m$  should be large enough.
- To reliably estimate the  $m$  unknown parameters, we need  $n$  to be large enough relative to  $m$ .

We conclude that  $m$  should be neither too large nor too small and should somehow depend on  $n$ . Before addressing the issue of selecting the model order  $m$ , we present a few popular semiparametric models.

**Piecewise-constant approximation.** Assume  $n/m$  is an integer and let  $\theta$  represent equispaced samples of the signal:  $\theta_j = s_{jn/m}$  for  $0 \leq j < m$ . Then approximate the signal as

$$\tilde{s}_i = \theta_j, \quad \text{for all } \frac{jn}{m} \leq i < \frac{(j+1)n}{m} \quad \text{and } 0 \leq j < m.$$

The problem with this approximation is the jumps at the locations  $i = jn/m$ . Such jumps are unnatural and uncharacteristic of many signals.

**Piecewise-linear approximation.** Again assume  $n/m$  is an integer and let  $\theta_j = s_{jn/m}$  for  $0 \leq j < m$ . Then approximate the signal by performing a linear interpolation between these samples:

$$\tilde{s}_i = \left( j + 1 - \frac{im}{n} \right) \theta_j + \left( \frac{im}{n} - j \right) \theta_{j+1}, \quad \text{for all } \frac{jn}{m} \leq i < \frac{(j+1)n}{m} \quad \text{and } 0 \leq j < m.$$

**Truncated Fourier Series.** Here the approximation consists of the first  $m$  terms in the Fourier series representation of the signal:

$$\tilde{s}_i = \theta_0 + \sum_{j=1}^{m/2} \left[ \theta_j \cos \frac{2\pi i j}{n} + \theta_{j+m/2} \sin \frac{2\pi i j}{n} \right], \quad 0 \leq i < n.$$

Note that in all three examples above, the approximation is of the form

$$\tilde{s}_i = \sum_{j=1}^m \theta_j \psi_j(i), \quad 0 \leq i < n \quad (2)$$

where  $\{\psi_j(i)\}$  are basis functions. The choice of the basis reflects our assumptions about the underlying characteristics of the signal (e.g., continuous, slowly varying, oscillatory, etc.) Other choices of basis functions include splines and wavelets.

## 1.4 Model Order Selection

Consider the semiparametric model (2). Having defined a family of models indexed by  $m$ , the problem is now to determine  $m$ . As mentioned above, there is a fundamental tradeoff between accuracy of the model (large  $m$  is better) and statistical accuracy of the estimates (small  $m$  is better). The following example illustrates this tradeoff.

**Example.** Consider the estimation problem (1) again, assume the basis  $\{\psi_j(i)\}$  is orthonormal, and denote by

$$S_j = \sum_{i=1}^n s_i \psi_j(i), \quad 1 \leq j \leq n$$

the coefficients of the signal in this basis. (The signal can be reconstructed from its coefficients as  $s_i = \sum_{j=1}^n S_j \psi_j(i)$ ,  $1 \leq i \leq n$ . In matrix form, we may write  $\underline{S} = \Psi \underline{s}$  and  $\underline{s} = \Psi^T \underline{S}$ , where  $\Psi^T = \Psi^{-1}$  by orthonormality of the basis.) Since iid Gaussian statistics are preserved by orthonormal transforms, the problem (1) may be equivalently written as

$$Y_j = S_j + N_j, \quad 1 \leq j \leq n$$

where  $\{N_j\}$  are iid  $\mathcal{N}(0, \sigma^2)$ . The ML estimator of  $S_j$  is  $\hat{S}_{ML,j} = Y_j$ . For a slowly varying signal, only the first few components  $S_j$  are significant, and the other ones are very small. Therefore component  $j$  is worth retaining in the model if  $S_j$  is “large” relative to  $\sigma$ ; otherwise  $\hat{S}_{ML,j}$  is just dominated by noise. But we do not know  $S_j$  and need to make decisions based on noisy data.

There are several ways of choosing  $m$ , including *fixed designs* such as  $m = n^\alpha$  for some  $0 < \alpha < 1$ . Modern inference methods use a *data-driven estimator*. Consider the so-called *complexity regularization criterion*

$$\mathcal{E}(m, \theta) = -\ln p(\underline{y} | \underline{s}(\theta)) + C_n m \quad (3)$$

where  $\underline{s}(\theta)$  is a  $m$ -dimensional parameterization of the signal (e.g., (2)), and  $C_n$  is some constant that depends on  $n$  but not on the data  $\underline{y}$ . Common choices include

- $C_n = 1$ : Akaike criterion [1];
- $C_n = \frac{1}{2} \ln n$ : Minimum Description Length (MDL) criterion (Rissanen [2]), *aka* Bayes Information Criterion (BIC) (Schwarz [3]);
- $C_n = \ln \ln n$ : Hannan and Quinn criterion [4].

The *complexity-regularized estimator* minimizes  $\mathcal{E}(m, \theta)$  over  $m$  and  $\theta$ . For any fixed  $m$ ,  $\max_{\theta} \mathcal{E}(m, \theta)$  is achieved by the ML estimator  $\hat{\theta}_{ML}^{(m)}$ . The estimated  $\hat{m}$  is therefore the maximizer of

$$\mathcal{E}(m, \hat{\theta}_{ML}^{(m)}) = -\ln p(\underline{y} | \underline{s}(\hat{\theta}_{ML}^{(m)})) + C_n m.$$

In our example above, we have

$$-\ln p(\underline{y} | \underline{s}(\theta)) = C + \frac{1}{2\sigma^2} \sum_{j=1}^n |Y_j - \theta_j|^2$$

and thus

$$\hat{\theta}_{ML,j}^{(m)} = \begin{cases} Y_j & : 1 \leq j \leq m \\ 0 & : m < j \leq n. \end{cases}$$

Hence

$$\begin{aligned} \mathcal{E}(m, \hat{\theta}_{ML}^{(m)}) &= -\ln p(\underline{y} | \underline{s}(\hat{\theta}_{ML}^{(m)})) + C_n m \\ &= C + \frac{1}{2\sigma^2} \sum_{j=m+1}^n |Y_j|^2 + C_n m. \end{aligned}$$

Note that  $\mathbb{E}|Y_j|^2 = S_j^2 + \sigma^2$ , and therefore

$$\mathbb{E}[\mathcal{E}(m, \hat{\theta}_{ML}^{(m)})] = C + \frac{1}{2\sigma^2} \sum_{j=m+1}^n |S_j|^2 + \frac{n-m}{2} + C_n m.$$

For instance, if  $S_j = \rho^{j/2} A$  where  $0 < \rho < 1$ , we obtain

$$\begin{aligned} \mathbb{E}[\mathcal{E}(m, \hat{\theta}_{ML}^{(m)})] &= C + \frac{A^2}{2\sigma^2} \frac{1 - \rho^{n-1}}{1 - \rho} \rho^m + \frac{n-m}{2} + C_n m \\ &= C' + \frac{A^2}{2\sigma^2} \frac{1 - \rho^{n-1}}{1 - \rho} \rho^m + \left(C_n - \frac{1}{2}\right) m. \end{aligned}$$

For large  $n$ , the above expectation is minimized by  $m^* \approx \frac{\ln C_n}{-\ln \rho} \ll n$ . The actual criterion  $\mathcal{E}(m, \hat{\theta}_{ML}^{(m)})$  may be viewed as a “noisy version” of its expectation, and  $\hat{m}$  is a random variable which might heuristically be expected to be of the order of  $m^*$ , as discussed below.

## 1.5 Consistency

Asymptotic analysis of nonparametric signal estimators can be performed as  $n \rightarrow \infty$  provided that the limiting behavior of the signal is meaningfully described. One possible scenario arising in signal processing is that of a signal defined over a finite time window  $[0, T]$  and sampled at times  $iT/n$  for  $1 \leq i \leq n$ .

The performance of complexity-regularized estimators of the form (3) can be studied from several angles.

The first is to analyze performance when the signal  $\underline{s}$  happens to belong to a parametric class of order  $m^*$ . Then, under some regularity conditions, it may be shown that the model order estimator  $\hat{m}$  converges in probability to the correct value  $m^*$  as  $n \rightarrow \infty$ . Moreover, the parameter estimator  $\hat{\theta}_{ML}^{(\hat{m})}$  is consistent.

The second is to analyze performance when  $\underline{s}$  belongs to a “well-behaved” class of signals such that the approximation accuracy rapidly improves with increasing  $m$ . There is no “true model order”, and the question of interest is whether the estimated signal  $\hat{\underline{s}}$  converges to  $\underline{s}$  in some appropriate sense (e.g., MSE). And if so, how does MSE behave as a function of  $n$ ?

An example of a “well-behaved” class is the so-called ellipsoidal class [5]

$$\Sigma(r, M) = \left\{ \underline{s} : \frac{1}{n} \sum_{j=1}^n S_j^2 j^{2r} \leq M \right\}$$

for some  $r > 0$  and  $M < \infty$ . This class consists of signals whose coefficients in the basis  $\Psi$  have rapidly decreasing magnitude, where  $\alpha$  controls the decay rate. For instance  $n^{-1/2}|S_j| \leq j^{-(r+1+\epsilon)}$  ensures that  $\underline{s} \in \Sigma(r, M)$  for some finite  $M$ . It may then be shown that for the MDL estimator,  $\hat{m}$  is of the order of  $n^{1/(2r+1)}$ , and the squared-error risk (MSE) is of the order of  $n^{-2r/(2r+1)}$ . Hence, for signals with rapidly decaying coefficients ( $r \gg 1$ ), the convergence rate for the MSE approaches  $n^{-1}$ , which is the usual convergence rate in the parametric case. If the coefficients decays less rapidly, the signal is “more complex”, and the MSE converges *more slowly* to zero.

## 1.6 Other Model Order Estimation Problems

**AR Models.** Consider the classical autoregressive (AR) model of order  $m$ :

$$Y_i = \sum_{j=1}^m \theta_j Y_{i-j} + N_i, \quad 1 \leq i \leq n$$

where  $\{N_i\}$  are iid  $\mathcal{N}(0, \sigma^2)$ . If  $m$  is known, the ML estimator for the AR coefficients  $\{\theta_j\}$  is the solution to the linear least-squares problem

$$\min_{\theta} \sum_{i=1}^n \left( Y_i - \sum_{j=1}^m \theta_j Y_{i-j} \right)^2.$$

If  $m$  is unknown, it can be estimated by minimizing the complexity-regularization criterion

$$\mathcal{E}(m, \theta) = -\ln p(\underline{y}|\theta) + C_n m \quad (4)$$

$$= C + \frac{1}{2\sigma^2} \sum_{i=1}^n \left( Y_i - \sum_{j=1}^m \theta_j Y_{i-j} \right)^2 + C_n m. \quad (5)$$

It has been shown [4] that  $\hat{m}$  is a consistent estimator of the true model order provided that  $C_n \geq \ln \ln n$ .

**Regression.** This is analogous to the signal estimation problem, except that the “sampling times” are random. The observations are pairs  $(X_i, Y_i)$ ,  $1 \leq i \leq n$ , where  $Y_i = s(X_i) + W_i$  and  $\{X_i\}, \{W_i\}$  are drawn iid from distributions  $p_X$  and  $p_W$ , respectively. The function  $s$  is the regression function, to be estimated from the data.

## 2 Nonparametric Density Estimation

Estimation of pdf’s is another problem that can be addressed either in a parametric framework (e.g., estimate the mean and variance of a Gaussian distribution; this is a two-dimensional parametric family) or in a nonparametric framework (by viewing the pdf as an unknown function [6, 7]). We are given iid data  $Y_i$ ,  $1 \leq i \leq n$  drawn from  $p$  which is to be estimated. Note this is much more tricky than the cdf estimation problem for which the empirical estimator

$$\hat{P}(y) = \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i \leq y\}}$$

converges a.s. in the sup norm to the true cdf  $P(y)$ . Indeed, while the pdf is just the derivative of the cdf, differentiation of  $\hat{P}(y)$  yields a sum of Dirac impulses located at the data points and weighted by  $\frac{1}{n}$ :

$$\hat{p}(y) = \frac{1}{n} \sum_{i=1}^n \delta(y - Y_i).$$

This so-called empirical pdf estimator is very noisy and generally considered unacceptable. For instance, the  $L^1$  norm  $\int |\hat{p}(y) - p(y)| dy$  does not converge to 0 as  $n \rightarrow \infty$ , and the  $L^2$  norm  $\int |\hat{p}(y) - p(y)|^2 dy$  is infinite for all  $n$ !

### 2.1 Histogram Estimator

Partition the domain of  $Y$  into a collection  $\mathcal{B} = \{\mathcal{B}_j, 1 \leq j \leq m\}$  of bins, and count how many samples  $Y_i$  land in each bin. The histogram estimator is defined as

$$\hat{p}^{(m)}(y) = \frac{1}{n} \sum_i 1_{\{Y_i \in \mathcal{B}_j\}}, \quad 1 \leq j \leq m, y \in \mathcal{B}_j.$$

For instance, if the domain of  $Y$  is  $[0, 1]$ , it is customary to choose equal-width bins:  $\mathcal{B}_j = [j/m, (j+1)/m)$ . The bin width should be neither too small (in which case  $\hat{p}^{(m)}$  would resemble the noisy empirical pdf) nor too large (in which case resolution is poor). For any given  $m$ , note that  $\hat{p}^{(m)}$  is the ML estimator of  $p$  in the class of piecewise-constant pdf's with bins  $\{\mathcal{B}_j, 1 \leq j \leq m\}$ . The MDL pdf estimator minimizes

$$\mathcal{E}(m) = - \sum_{i=1}^n \ln \hat{p}^{(m)}(y_i) - C_n m,$$

and it may be shown that it is a consistent estimator for “well-behaved”  $p$ . However piecewise-constant approximations are undesirable for smooth pdf's, and unsurprisingly the convergence rate of the histogram estimator is suboptimal.

## 2.2 Kernel Density Estimator

To obtain a smoother pdf estimator, it is customary to use kernel estimators. A kernel is a smooth, symmetric pdf  $K(y)$  with zero mean, e.g., a Gaussian distribution. The kernel estimator is defined as the convolution of the empirical pdf estimator with the kernel:

$$\hat{p}_K(y) = (\hat{p} \star K)(y) = \frac{1}{n} \sum_{i=1}^n K(y - Y_i).$$

The notion of choosing the histogram bin width is replaced here with the similar notion of choosing the *kernel bandwidth* (r.m.s. value of  $K(y)$ , which is equal to  $\sigma$  if  $K = \mathcal{N}(0, \sigma^2)$ ).

## References

- [1] H. Akaike, “A New Look at the Statistical Model Identification,” *IEEE Trans. Automatic Control*, Vol. 19, No. 6, pp. 716—723, 1974.
- [2] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1992.
- [3] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, Vol. 6, No. 2, pp. 461—464, 1978.
- [4] E. J. Hannan and B. G. Quinn, “The Determination of the Order of an Autoregression,” *Journal of the Royal Statistical Society, B*, Vol. 41, pp. 190—195, 1979.
- [5] A. R. Barron, L. Birgé and P. Massart, “Risk bounds for model selection via penalization,” *Probability Theory and Related Fields*, Vol. 113, 301—415, 1999.
- [6] A. R. Barron and T. M. Cover, “Minimum Complexity Density Estimation,” *IEEE Trans. IT*, Vol. 37, No. 4, pp. 1034—1054, 1991.
- [7] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, UK, 1986.