Parametric density estimation

Quick note

Abdelwahid Benslimane

wahid.benslimane@gmail.com

When a priori knowledge about the data indicates that some probability density is among a set of parametric density functions (e.g. multidimensional normal laws), it is then efficient to look directly for the optimal solution among this set, i.e. to estimate the parameters that define it among this set. If an elementary law from the set does not explain the data well enough, it is then possible to consider an additive mixture of several laws from the same set. Contrary to the kernel method, here the mixture contains a small number of laws but these laws can have different parameters and different weights in the mixture.

The term "efficient" is used above for two very distinct aspects:

- once the mixture model is estimated, the computation of the density at a point x is much faster because the mixture contains m laws and not N, with $m << N$ (N = number of examples in the sample data),
- for the same number N of observations in D_N, the parametric estimation has in general a better precision than the nonparametric estimation by kernels **(please take a look at my explanation of the kernel density estimation to understand this part. You can contact me if you you want to get it)**.

If we consider a set of density functions F parameterized by a vector $\theta \in \Omega$ (parameter space), then parametric estimation consists of finding the optimal parameter vector $\hat{\theta}^*$ that defines the best estimate $f_{\hat{\theta}^*}$ of the sought-after density.

Each vector θ defines a density f_θ which explains more or less well the observations of D_N.

To measure the adequacy of the density f_θ (belonging to the set F and defined by the parameter vector θ) to the observations of D_N, we use $p(D_N|\theta)$. If the observations $x_1, \ldots, x_N \in D_N$ correspond to independent and identically distributed draws (according to the desired density), then:

$$p(D_N|\theta) = \prod_{i=1}^{N} f_\theta(x_i).$$

In the formula above, $p(D_N|\theta)$ is the likelihood of θ with respect to the sample D_N. The parameter vector $\hat{\theta}^*$ which defines the density function $f_{\hat{\theta}^*}$ that is most consistent with the observations in D_N is the one that maximizes the likelihood. In order to find it, it is necessary to use an optimization procedure.

When the density functions are defined with an exponential it is generally preferable to work with the log-likelihood as the application of the logarithm allows an easier handling:

$$L(\theta) \equiv ln[p(D_N|\theta)] = \sum_{i=1}^{N} ln[f_\theta(x_i)].$$

The logarithm in base > 1 being a strictly increasing function, the argument $\hat{\theta}^*$ that maximizes the likelihood also maximizes the log-likelihood and vice versa.

A very simple case is when the observations are one-dimensional (d=1), $D_N \subset \mathbb{R}$, and have been generated according to a density function f which is part of the set of (one-dimensional) normal laws, $f \in F = \mathcal{N}(\mu, \sigma)$. Such a density function is defined by the expectation μ and the standard deviation σ, so the parameter vector is $\theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}$. We seek to estimate this density function by $\hat{f} \in F = \mathcal{N}(\mu, \sigma)$, with parameters $\hat{\theta} = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix}$.

In order to find the optimal parameter vector $\hat{\theta}^* = \begin{pmatrix} \hat{\mu}^* \\ \hat{\sigma}^* \end{pmatrix}$ it is necessary to write the expression for the (log-)likelihood and determine its maximum. The (one-dimensional) normal distribution of parameter θ is:

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Consequently:

$$ln[p(D_N|\theta)] = \sum_{i=1}^{N} ln[f_\theta(x_i)] = \sum_{i=1}^{N} [-ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}(x_i - \mu)^2].$$

The expression above is indefinitely derivable with respect to both variables. We obtain the following first partial derivatives:

$$\frac{\partial ln[p(D_N|\theta)]}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)$$

$$\frac{\partial ln[p(D_N|\theta)]}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)^2$$

These partial derivatives become zero for:

- $\hat{\mu}^* = \frac{1}{N} \sum_{i=1}^{N} x_i$ which is the empirical mean of the observations of D_N

and

- $\hat{\sigma}^{*2} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}^*)^2$ which is the empirical biased variance of the observations.

By calculating the second derivatives we can verify that they are negative, therefore $\hat{\theta}^* = \begin{pmatrix} \hat{\mu}^* \\ \hat{\sigma}^* \end{pmatrix}$ is indeed the argument that maximizes the log-likelihood $ln[p(D_N|\theta)]$.

Two observations should be made:

- The estimate $\hat{\mu}^*$ is unbiased, i.e., its expectation over samples of size N is equal to the true value of μ.
- The estimate $\hat{\sigma}^{*2}$ on the other hand, is biased, although asymptotically unbiased. The following unbiased estimate is preferred: $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}^*)^2$.