Polynomial smoothing of discrete sparsely observed distribution

Paulo Eduardo Oliveira

CMUC, University of Coimbra, Portugal, email: paulo@mat.uc.pt

Keywords: discrete data, local polynomial, sparse observations.

Let $\mathbf{P}$ be a probability distribution on a support with N cells arranged, for simplicity, in a table $\mathbf{C} = (C_{i,j})$, where $i = 1 \ldots, K$, $j = 1, \ldots L$. Observation counts are described by $\mathbf{N} = (N_{i,j})$, or equivalently, by the empirical probability distribution $\overline{\mathbf{P}} = (\overline{P}_{i,j} = N_{i,j}/n)$, where $n = \sum_{i,j} N_{i,j}$. Rearranging the rows in order to have a N-dimensional vector, $\mathbf{N}$ is multinomially distributed.

This talk is concerned with the estimation of $\mathbf{P} = (P_{i,j})$ with a special interest when the sample size n is small. Moreover, having in mind a few applications, some partial knowledge of the distribution might be available and we should integrate this into the estimation. We will assume the knowledge of the marginal distribution of $\mathbf{P}$, that is, for some given Π_i, $i = 1, \ldots, K$, $\Pi_i = \sum_{j=1}^{L} P_{i,j}$. The general idea in constructing estimators is to adapt polynomial smoothing to this framework.

To avoid computational difficulties with border and edge effects, we consider a replication of the tables $\mathbf{C}$, $\mathbf{P}$ and $\mathbf{N}$, enlarging them by reflecting cells with respect to each one of the four borders and edges. Defining the appropriate matrices in the usual way, local polynomials can be represented as the minimizers of $H_{i,j} = \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right)^t \mathbf{W}_{i,j} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right)$, for each i and j. This procedure does not integrate knowledge of the marginal distribution. Moreover, it may produce non-acceptable estimates, especially when n is small, as is our case of interest. We propose to correct this in two different ways. The first one is to introduce in the minimization problem a constraint, forcing the solution to agree with the marginal distribution:

$$\begin{aligned} &\text{minimize} \sum_{\ell=1}^{L} H_{i,\ell} \\ &\text{subject to } \textstyle\sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i, \quad i = 1, \ldots, K. \end{aligned}$$

The second one changes the minimizing function by considering relative errors:

$$\begin{aligned} &\text{minimize } H_i^* = \sum_{\ell=1}^{L} \frac{1}{\beta_{0,i,\ell}} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,\ell}\beta_{i,\ell}\right)^t \mathbf{W}_{i,\ell} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,\ell}\beta_{i,\ell}\right), \\ &\text{subject to } \sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i, \quad i = 1, \ldots, K. \end{aligned}$$

We characterize the obtained estimators, describe their behaviour and relations, and present some numerical results showing their performance. Finally, we note that this approach is easily extended to higher dimensional supports.