Chapter G10

Smoothing in Statistics

Contents

1   Scope of the Chapter

This chapter is concerned with methods for smoothing data. Included are methods for density estimation, smoothing time series data, and statistical applications of splines. These methods may also be viewed as nonparametric modelling.

2   Background to the Problems

2.1   Smoothing Methods

Many of the methods used in statistics involve fitting a model, the form of which is determined by a small number of parameters, for example, a distribution model like the gamma distribution, a linear regression model or an autoregression model in time series. In these cases the fitting involves the estimation of the small number of parameters from the data. In modelling data with these models there are two important stages in addition to the estimation of the parameters; these are: the identification of a suitable model, for example, the selection of a gamma distribution rather than a Weibull distribution, and the checking to see if the fitted model adequately fits the data. While these parametric models can be fairly flexible, they will not adequately fit all data sets, especially if the number of parameters is to be kept small.

Alternative models based on smoothing can be used. These models will not be written explicitly in terms of parameters. They are sufficiently flexible for a much wider range of situations than parametric models. The main requirement for such a model to be suitable is that the underlying models would be expected to be smooth, so excluding those situations where, for example, a step function would be expected.

These smoothing methods can be used in a variety of ways, for example:

(1)   producing smoothed plots to aid understanding;
(2)   identifying of a suitable parametric model from the shape of the smoothed data;
(3)   eliminating complex effects that are not of direct interest so that attention can be focused on the effects of interest.

Several smoothing techniques make use of a smoothing parameter which can be either chosen by the user or estimated from the data. The smoothing parameter balances the two criterion of smoothness of the fitted model and the closeness of the fit of the model to the data. Generally, the larger the smoothing parameter is, the smoother the fitted model will be, but for small values of the smoothing parameter the model will closely follow the data, and for large values the fit will be poorer.

The smoothing parameter can either be chosen using previous experience of a suitable value for such data, or estimated from the data. The estimation can be either formal, using a criterion such as the cross-validation, or informal by trying different values and examining the result by means of suitable graphs.

Smoothing methods can be used in three important areas of of statistics: regression modelling, distribution modelling and time series modelling.

2.2   Smoothing Splines and Regression Models

For a set of n observations (y_i, x_i), $i = 1, 2, \ldots, n$, the spline provides a flexible smooth function for situations in which a simple polynomial or nonlinear regression model is not suitable.

Cubic smoothing splines arise as the function, f, with continuous first derivative which minimizes

$$\sum_{i=1}^{n} w_i \{y_i - f(x_i)\}^2 + \rho \int_{-\infty}^{\infty} (f''(x))^2 dx,$$

where w_i is the (optional) weight for the ith observation and ρ is the smoothing parameter. This criterion consists of two parts: the first measures the fit of the curve and the second the smoothness of the curve. The value of the smoothing parameter, ρ, weights these two aspects: larger values of ρ give a smoother fitted curve but, in general, a poorer fit.

Splines are linear smoothers since the fitted values, $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$, can be written as a linear function of the observed values $y = (y_1, y_2, \ldots, y_n)^T$, that is,

$$\hat{y} = Hy$$

for a matrix H. The degrees of freedom for the spline is $\text{trace}(H)$ giving residual degrees of freedom

$$\text{trace}(I - H) = \sum_{i=1}^{n}(1 - h_{ii}).$$

The diagonal elements of H, h_{ii}, are the leverages.

The parameter ρ can be estimated in a number of ways.

(1) The degrees of freedom for the spline can be specified, i.e., find ρ such that $\text{trace}(H) = \nu_0$ for given ν_0.

(2) Minimize the cross-validation (CV), i.e., find ρ such that the CV is minimized, where

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i}{1 - h_{ii}}\right)^2.$$

(3) Minimize generalised cross-validation (GCV), i.e., find ρ such that the GCV is minimized, where

$$\text{GCV} = n\left(\frac{\sum_{i=1}^{n} r_i^2}{\left(\sum_{i=1}^{n}(1 - h_{ii})\right)^2}\right).$$

2.3 Density Estimation

The object of density estimation is to produce from a set of observations a smooth nonparametric estimate of the unknown density function from which the observations were drawn. That is, given a sample of n observations, $x_1, x_2, \ldots, x_n$, from a distribution with unknown density function, $f(x)$, find an estimate of the density function, $\hat{f}(x)$. The simplest form of density estimator is the histogram; this may be defined by:

$$\hat{f}(x) = \frac{1}{nh}n_j; \quad a + (j-1)h < x < a + jh; \quad j = 1, 2, \ldots, n_s,$$

where n_j is the number of observations falling in the interval $a + (j-1)h$ to $a + jh$, a is the lower bound of the histogram and $b = n_s h$ is the upper bound. The value h is known as the window width. A simple development of this estimator would be the running histogram estimator

$$\hat{f}(x) = \frac{1}{2nh}n_x; \quad a \leq x \leq b,$$

where n_x is the number of observations falling in the interval $[x - h : x + h]$. This estimator can be written as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right)$$

for a function w where

$$
\begin{aligned}
w(x) &= \tfrac{1}{2} \quad \text{if } -1 < x < 1 \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

The function w can be considered as a kernel function. To produce a smoother density estimate, the kernel function, $K(t)$, which satisfies the following conditions can be used:

$$\int_{-\infty}^{\infty} K(t)dt = 1 \text{ and } K(t) \geq 0.0.$$

The kernel density estimator is therefore defined as:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

The choice of $K(\cdot)$ is usually not important but to ease computational burden, use can be made of Gaussian kernel defined as:

$$K(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

The smoothness of the estimator, $\hat{f}(x)$, depends on the window width, h. In general, the larger the value h is, the smoother the resulting density estimate is. There is, however, the problem of oversmoothing

when the value of h is too large and essential features of the distribution function are removed. For example, if the distribution was bimodal, a large value of h may result in a unimodal estimate. The value of h has to be chosen such that the essential shape of the distribution is retained while effects due only to the observed sample are smoothed out. The choice of h can be aided by looking at plots of the density estimate for different values of h, or by using cross-validation methods; see Silverman [2].

Silverman [2] shows how the Gaussian kernel density estimator can be computed using a fast Fourier transform (FFT).

2.4   Smoothers for Time Series

If the data consists of a sequence of n observations recorded at equally spaced intervals, usually a time series, several robust smoothers are available. The fitted curve is intended to be robust to any outlying observations in the sequence, hence the techniques employed primarily make use of medians rather than means. These ideas come from the field of exploratory data analysis (EDA); see Tukey [3] and Velleman and Hoaglin [4]. The smoothers are based on the use of running medians to summarize overlapping segments; these provide a simple but flexible curve.

In EDA terminology, the fitted curve and the residuals are called the smooth and the rough respectively, so that

$$\text{Data} = \text{Smooth} + \text{Rough}.$$

Using the notation of Tukey, one of the smoothers commonly used is 4253H, twice. This consists of a running median of 4, then 2, then 5, then 3. This is then followed by what is known as Hanning. Hanning is a running weighted mean, the weights being 1/4, 1/2 and 1/4. The result of this smoothing is then 'reroughed'. This involves computing residuals from the computed fit, applying the same smoother to the residuals and adding the result to the smooth of the first pass.

3   Recommendations on Choice and Use of Available Routines

Note. Refer to the Users' Note for your implementation to check that a routine is available.

The following routines fit smoothing splines:

G10ABF   computes a cubic smoothing spline for a given value of the smoothing parameter. The results returned include the values of leverages and the coefficients of the cubic spline. Options allow only parts of the computation to be performed when the routine is used to estimate the value of the smoothing parameter or as when it is part of an iterative procedure such as that used in fitting generalized additive models; see Hastie and Tibshrani [1].

G10ACF   estimates the value of the smoothing parameter using one of three criteria and fits the cubic smoothing spline using that value.

G10ABF and G10ACF require the x_i to be strictly increasing. If two or more observations have the same x_i-value then they should be replaced by a single observation with y_i equal to the (weighted) mean of the y-values and weight, w_i, equal to the sum of the weights. This operation can be performed by G10ZAF.

The following routine produces an estimate of the density function:

G10BAF   computes a density estimate using a Normal kernel.

The following routine produces a smoothed estimate for a time series:

G10CAF   computes a smoothed series using running median smoothers.

The following service routine is also available:

G10ZAF   orders and weights the (x, y) input data to produce a data set strictly monotonic in x.

4   Routines Withdrawn or Scheduled for Withdrawal

None since Mark 13.

5   References

[1]   Hastie T J and Tibshirani R J (1990) *Generalized Additive Models* Chapman and Hall

[2]   Silverman B W (1990) *Density Estimation* Chapman and Hall

[3]   Tukey J W (1977) *Exploratory Data Analysis* Addison–Wesley

[4]   Velleman P F and Hoaglin D C (1981) *Applications, Basics, and Computing of Exploratory Data Analysis* Duxbury Press, Boston, MA