Penalized Splines with an Application in Economics

Chaojiang Wu
Drexel University, USA

INTRODUCTION

Nonparametric smoothing has become the norm of statistics research in the last few decades. The advantage of nonparametric smoothing is its data driven nature. It does not assume any predetermined functional form but allow data to speak for themselves. One particularly popular nonparametric smoothing technique is penalized splines or P-splines. This technique is popular partly because it is computationally expedient and flexible enough to be adapted in complex models. Despite of its popularity in statistics methodological research, its applications in other areas such as finance and economics are limited. One reason for such limited applications is that P-splines remain unknown to most economics researchers. This chapter introduces the general idea of penalized splines and illustrates its application in forecasting US real GNP. Some references and resources are pointed out for future researches.

BACKGROUND

Nonparametric smoothing or nonparametric regression is a technique for fitting curves, where the form of the curve is not predetermined but estimated through data. In its simplest form, given data pairs (x_i , y_i), $i = 1,\ldots, n$, the classical simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ,$$

where β_0 is the intercept and β_1 is the slope parameter, ε_i is error term, usually assumed to be i.i.d. $N(0,\sigma^2)$. In the above model, a predetermined linear functional form describes the dependence of response variable y on predictor x; the parameters β_0 and β_1 are to be estimated. However, in many cases, it is inappropriate to fit a straight line. Alternatively, response variable y takes the following form

$$y_i = g\left(x_i\right) + \varepsilon_i ,$$

where $g(\cdot)$ is an unspecified function and ε_i is again i.i.d. $N(0,\sigma^2)$. No parameters are explicitly present in the model. In other words, we are trying to estimate the function rather than estimating the parameters in any predetermined form.

The most popular ways of estimating $g(\cdot)$ are local polynomial methods (local methods) and spline methods (global methods). Some excellent references on local methods can be found in Wand and Jones (1995) and Fan and Gijbels (1996). The method introduced in this chapter (P-spline) is one type of global methods. There are generally three types of spline methods: smoothing splines, regression splines and penalized splines. Penalized splines or P-splines can be viewed as a generalization of smoothing splines with a more flexible choice of bases and penalties. Alternatively, P-splines can be viewed as least square regression splines with a roughness penalty. Some discussions on smoothing spline methods can be seen in Wahba (1990) and Eubank (1988). General discussions on penalized splines can be found in Ruppert, Wand and Carroll (2003). Applications of penalized splines have gained popularity only recently. Some examples can be found in Jarrow, Ruppert and Yu (2004) on estimating the term

DOI: 10.4018/978-1-4666-5202-6.ch163

structure of corporate bonds, Brunauer, Lang, Wechselberger and Bienert (2010) on estimating the nonlinear effects and smooth time trend in modeling the rent in Vienna, and Çetinyürek and Lambert (2011) on estimating the survival function and hazard ratios for interval-censored data in Medicine.

MAIN FOCUS

This section introduces P-spline estimation ideas, computational issues (bases choice, smoothing parameter selection and computing resource), and illustrate the method by one application in forecasting real United States GNP.

Penalized Splines Estimation

The basic idea of splines estimation of an unknown smooth function *g(x)* is to express the smooth function as a linear combination of piece-wise polynomial functions, or splines. For mathematical convenience, the function is expressed as a linear combination of the spline basis. That is

$$g\left(x\right) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - x_k)_+^p,$$

where p is the degree of spline (polynomials), $\left(x - x_k\right)_+^p = \max\{\left(x - x_k\right)^p, 0\}$, and $\boldsymbol{\beta} = \left(\beta_0, \dots, \beta_{p+k}\right)^T$ is the spline coefficients to be estimated. Define $\boldsymbol{X} = \left(1, x, \dots, x^p, \left(x - x_1\right)_+^p, \dots, \left(x - x_k\right)_+^p\right)$ as the spline basis (truncated power basis in this case), with break points at x_k, *k=1,...,K*. ***X*** is the "design matrix" similar to that of linear regressions. Then the unknown smooth function can be written as $g\left(x\right) = \boldsymbol{X\beta}$.

The series of break points $x_1 < \ldots < x_k$ are known as knots of the splines. The choice of number (*K*) and locations (x_k) of knots is crucial and usually the process is complicated and computationally intensive. One school of thought is the regression splines, which select the knots and locations carefully. Such knots selection usually requires sophisticated algorithm such as Multivariate Adaptive Regression Splines (MARS) algorithm (Friedman, 1991), which selects knots by stepwise type regressions. Smoothing splines, on the other hand, place knots at each distinct observation of *x*. Smoothing splines are theoretically appealing because of its mathematical nice properties. Specifically, the natural cubic spline with knots placed in distinct observations of *x* is the solution to the minimizer of roughness penalized least square deviance, over the class of twice differentiable functions. However, the computation of smoothing splines is intensive for such knots placement, especially when the application is complex beyond univariate smoothing. In between the two types of splines, penalized splines use sufficiently large number of knots, often equally placed in the equidistant quantiles of *x* and shrink unnecessary knots by adding penalties (Eilers & Marx, 1996). The number of knots is no longer crucial as long as the smoothing parameter is properly chosen (Ruppert, 2002).

Consider finding the smoothing estimator for *g(x)=**Xβ*** using spline model with *K* knots (*K* sufficiently large, say 40), one way of formulating the fit is

$$\underset{\boldsymbol{\beta}}{argmin} \left\| \boldsymbol{y} - \boldsymbol{X\beta} \right\|^2 + \lambda \sum_{k=1}^{K} \beta_{p+k}^2,$$

where $\boldsymbol{y}=(y_1,y_2,\ldots,y_n)^T$ is the vector of response variable, and $\lambda>0$ is the smoothing parameter that controls the smoothness of the fitting curve. The first term of the objective function is the usual Ordinary Least Squares (OLS) goodness-of-fit term while the second term is the roughness penalty term which constrains the influence of unnecessary knots. The smoothing parameter λ balances the goodness-of-fit and roughness of the curve. Very large λ results in a linear least square