CEM

Center for Empirical Macroeconomics

Working Paper No. 85

Estimating penalized spline regressions: Theory and application to economics

by

Alfred Greiner

University of Bielefeld
Department of Economics
Center for Empirical Macroeconomics
P.O. Box 100 131
33501 Bielefeld, Germany

Estimating penalized spline regressions: Theory and application to economics*

Alfred Greiner

Department of Business Administration and Economics

Bielefeld University, P.O. Box 100131

33501 Bielefeld, Germany

e-mail: agreiner@wiwi.uni-bielefeld.de

Abstract

In this paper we give a brief survey of penalized spline smoothing. Penalized spline smoothing is a general non-parametric estimation technique which allows to fit smooth but else unspecified functions to empirical data. While penalized spline regressions are quite popular in natural sciences only few applications can be found in economics. We present an example demonstrating how this non-parametric estimation technique can help to gain insights into economics.

JEL: C140

Keywords: Non-parametric estimation, penalized spline smoothing, Generalized Additive Models

*I thank Uwe Köller for providing me with the data.

1 Introduction

Models with smooth functions which are not specified parametrically have become more and more popular over the last two decades in statistics. One reason for this process is to be seen in the development of capable computers and in the design of statistical software which allow to fit highly structured models.

A first milestone in the development of non-parametric and semi-parametric estimation was set by Hastie and Tibshirani (1990) who introduced Generalized Additive Models as a new flexible class of regression model. The theoretical results were complemented by the development of numerical algorithms which led to software packages such as S-PLUS and later on R (see Venables and Ripley, 2003).

The main idea behind this class of models is that the effect of an explanatory variable on some dependent variable of interest is not modelled as a parametric, usually linear, function but kept flexible. The only assumption needed is that the effect of the explanatory variables are modelled as smooth, i.e. differentiable, functions. The functional shape, then, is to be estimated from the data by either using local kernel based methods or by spline smoothing.

Available software readily allows to estimate such models. The contributions by Wood (2000, 2001) allowed to also settle the disputable point of choosing the right amount of smoothing in practice. An overview of the state of art in this field can be found in Ruppert, Wand and Carroll (2003).

In this paper we give a brief introduction to non-parametric and semi-parametric estimation based on penalized spline estimation and we present an example from economics in order to highlight the advantage of this estimation technique.

The rest of the paper is organized as follows. In the next section, we give an outline of spline smoothing and of penalized splines. Section 3 presents an example from economics and section 4, finally, concludes.

2 Spline smoothing and penalized splines

Assume that we have n data points for the dependent variable y which is explained by the independent variable x. Thus, we have observations (y_i, x_i), $i = 1,...,n$, and the regression model we want to estimate is

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_t \sim iid(0, \sigma^2). \tag{1}$$

$f(\cdot)$ is an unknown function which is not specified further, except that we require $f(\cdot)$ to be continuous and sufficiently differentiable. The idea behind spline estimation, then, is to find the function $f(x)$ such that the following minimization problem is solved (cf. Hasti and Tibshirani, 1990),

$$\min_{f(\cdot)} \left\{ \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 \, dx \right\}. \tag{2}$$

(2) shows that the function to be minimized consists of two components: first, the deviation of the fitted function from the observed values should be minimized which as usual gives the goodness of the fit. Second, complex functions are penalized by the second term in (2), measured by the second order derivative.

Reinsch (1967) demonstrated that $f = (f(x_1),..., f(x_n))$ in (2) can be written as $f = C\alpha$, with C as cubic spline basis and α the spline coefficient. Thus, (2) can be rewritten in the following form,

$$\min_{\alpha} ||y - C\alpha||^2 + \lambda\alpha^T H\alpha, \tag{3}$$

with $|| \cdot ||^2$ the usual Euclidian norm and H a penalty matrix.[1] λ in (2) and (3) is a smoothing parameter which controls the trade-off between closely matching the data and having a linear model. For $\lambda \to \infty$ the minimization of (3) gives a linear fit whereas letting $\lambda \to 0$ gives a wiggly function.

[1] In Wood and Augustin (2002), for example, it is demonstrated in detail how this matrix can be constructed.

These considerations demonstrate that the choice of λ plays an important role in the estimation. In principle, λ can be set by hand, but it is also possible to choose λ data driven. One possibility to do so is to resort to cross-validation. Cross validation works as follows: leave out one observation and fit the model to the rest of the data. Then, compute the squared difference between the observation point that was left out and the value for this observation predicted by the estimated model. This procedure is repeated for each data point in the data set and the following cross-validation sum of squares is computed,

$$CV(\lambda) = (1/n) \sum_{i=1}^{n} \left(y_i - \hat{f}_{-i,\lambda}(x_i) \right)^2, \tag{4}$$

with $\hat{f}_{-i,\lambda}(x_i)$ the estimate for $f(x_i)$ based on data points (x_j, y_j), $j = 1,..., i-1, i+1,..., n$, and computed with the smoothing parameter λ. The minimization of $CV(\lambda)$ with respect to λ then gives a data driven value for λ. In practical applications one replaces the cross validation criterion by the generalized cross validation (GCV) criterion which is easier to compute (for details see Hastie and Tibshirani, 1990, ch. 3.4).

One problem associated with solving (3) is that the spline basis C grows with the order of the size of the sample. So, for large samples the smoothing spline estimation would lead to the problem of inverting an $n \times n$ matrix, where n gives the size of the sample. A modification to smoothing spline estimation results by reformulating (3) such that $f = D\alpha$ with D a high-dimensional basis function (conventionally, D is 10 to 40 dimensional). The difference to smoothing splines is that the number of basis function is fixed and does not grow with the number of observations.

To fit a model the spline coefficients are penalized in the same way as in (3) with an appropriate penalty matrix H giving the minimization problem,

$$\min_{\alpha} ||y - D\alpha||^2 + \lambda \alpha^T H \alpha. \tag{5}$$

This approach is referred to as penalized spline smoothing (for more details see Ruppert, Wand and Carroll, 2003).

Writing the objective in (5) as

$$(y - D\alpha)^T (y - D\alpha) + \lambda\alpha^T H\alpha = y^T y - 2\alpha^T D^T y + \alpha^T \left[D^T D + \lambda H\right] \alpha \tag{6}$$

and minimizing by differentiating with respect to α and setting the result equal to zero gives $\hat{\alpha} = \left(D^T D + \lambda H\right)^{-1} D^T y$.

The fitted function, then, is obtained as

$$\hat{f}(x) = A(\lambda)\, y, \quad \text{with} \quad A(\lambda) = D\left(D^T D + \lambda H\right)^{-1} D^T. \tag{7}$$

The matrix $A(\lambda)$ is called the smoothing matrix and the trace of this matrix are the estimated degrees of freedom of the model. These reflect the degree of complexity of the fitted model. For $\lambda \to \infty$ the trace of $H(\lambda)$ equals 1, giving a linear fit, while for $\lambda = 0$ the trace of $H(0)$ is $p + 1$, with p as dimension of the matrix D. Setting $\lambda = 0$ implies that the curvature is not punished and, consequently, yields a very wiggly function.[2]

Our considerations above dealt with a model where the dependent variable y was a function of one explanatory variable x as modelled in (1). One extension is obtained when we generalize the assumption of normality and write the equation as

$$g(\mu) = f(x), \tag{8}$$

with $\mu = E(y|\, x)$ the expected value, $g(\cdot)$ a monotonic and differentiable link function which is known and y belonging to an exponential family distribution. Another extension is obtained by allowing y to depend on more than one explanatory variable, for example x and z, in a nonlinear way. This gives rise to a Generalized Additive Model (GAM) which can be written as follows,

$$g(\mu) = f(x) + h(z), \tag{9}$$

where $\mu = E(y|\, x, z)$ is again the expected value and $f(\cdot)$ and $h(\cdot)$ are unknown but smooth functions which are to be estimated from empirical data.

---

[2]We do not discuss question of statistical inference. For a description of how to obtain the variance associated with the smooth term, see e.g. Wood and Augustin (2002) or Hasti and Tibshirani (1990).

With more than one function the estimation of the model can be done by using a backfitting strategy. This means that all terms are kept fixed except for the smooth which is fitted by following smoothing as described above. One then circles over the different smooth components by fitting just one of the smooth functions. For further details we again refer to Hastie and Tibshirani (1990). The major advantage of the additive model is that the curse of dimensionality, stating that the required sample size grows exponentially with the dimension of the fitted function, can be avoided.

Of course, one can also fit a semi-parametric model, i.e. a model which consists of a parametric and of a non-parametric part at the same time. For example, such a model could be written as

$$g(\mu) = f(x) + \beta z, \tag{10}$$

where β is the parameter and $f(\cdot)$ is the nonlinear function to be estimated.

Another type of models, the Varying Coefficient Models (see Hasti and Tibshirani, 1993), is obtained when metrically scaled and nominal variables are used as explanatory variables. However, we will not go into the details of this class of models. In statistics a lot of work has been devoted to find efficient ways to fit those models as well as models (9) and (10) (see e.g. Kauermann and Tutz, 2000, and Wood, 2003).

3 Applying penalized spline estimation: An example from economics

One possible application of penalized spline smoothing is to study the relation which exists between the primary surplus to GDP ratio and the debt ratio. The motivation for this is that a given fiscal policy is sustainable if the primary surplus to GDP ratio is a linear or convex function of the debt ratio as demonstrated by Bohn (1998). For the US, Bohn estimated the relationship by OLS and found a positive and statistically significant reaction of the primary surplus-GDP ratio to the debt ratio implying that US debt is sustainable. Greiner and Kauermann (2005) showed that the relation between the surplus and the debt ratio is characterized by nonlinearities and that the response of the primary surplus tends to be the larger the higher the debt ratio is, suggesting a convex curve.

Here, we want to apply this test to Germany. The equation we estimate is as follows

$$s_t = \alpha + f(d_{t-1}) + \beta_1 \, YVAR_t + \beta_2 \, int_t + \epsilon_t \,. \tag{11}$$

s_t is the ratio of the primary surplus to GDP in period t and d_{t-1} is the debt ratio of period $t-1$.[3] It should be noted that we include the lagged debt ratio, d_{t-1}, because budget plans are usually made one year in advance. $YVAR_t$ is a variable which reflects business cycles and is computed by applying the HP-filter to the GDP series and int_t is the real interest rate. All data are annual and cover the time period 1960-2003.

$f(\cdot)$ is a smooth function and β_i, $i = 1, 2$, are the coefficients associated with the variables $YVAR_t$ and int_t, respectively. (11) implies that we estimate a semi-parametric model. We do this because pre-estimations showed that there is no indication that any variable other than the debt ratio d_{t-1} enters the equation nonlinearly. In these pre-estimations we first estimated (11) with all variables entering the equation in a nonlinear way showing that only the variable d_{t-1} has a nonlinear effect while the estimated degrees of freedom (edf) for all other variables is 1 which suggests a linear relationship. Next, we estimated equation (11) with all variables entering the equation linearly except for one. The results demonstrated again that the relationship is linear for all variables with the exception of d_{t-1} which has a nonlinear effect on the primary surplus ratio.

As a benchmark we use the result obtained by first performing a linear estimation, i.e. we assume $f(d_{t-1}) = \beta_0 \, d_{t-1}$. Table (1) gives the estimation result.

---

[3] As to the data source see the Appendix.

Table 1: Estimated parameters for model (11) with $f(d_{t-1}) = \beta_0 \cdot d_{t-1}$.

coefficient	estimate	std dev	t-value
Intercept	-0.001	0.01	-0.1
d_{t-1}	-0.017	0.014	-1.19
$YVAR_t$	0.507	0.117	4.33
int_t	0.276	0.227	1.21
DW = 0.74	R^2(adj) = 0.32		

Table (1) suggests that the parameter β_0 is negative but not statistically significant. A negative sign of the coefficient would imply that the government does not react to higher debt ratios implying that a given fiscal policy would not be sustainable. However, looking at the relationship between the primary surplus ratio and the lagged debt ratio shows that this relationship is not well described by a linear curve as figure 1 suggests.

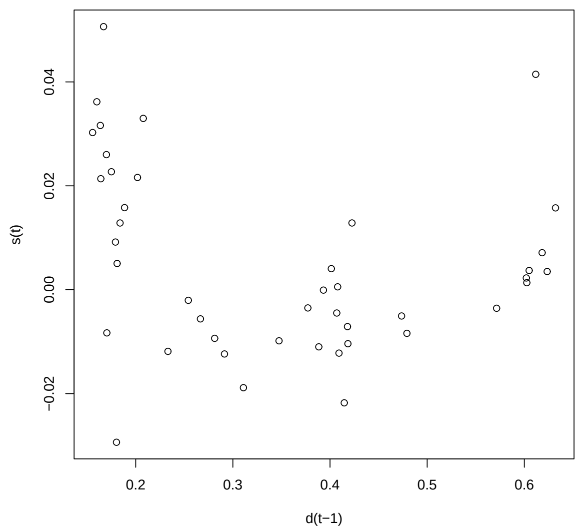

Figure 1: Surplus s_t and lagged debt ratio d_{t-1}.

Therefore, we next assume that the function $f(\cdot)$ is smooth but otherwise unspecified. Table (2) gives the penalized spline estimation result for the parametric part of equation (11) in this case.[4]

Table 2: Estimates for the parametric part of model (11) with $f(d_{t-1})$ obtained by penalized spline estimation.

coefficient	estimate	std dev	t-value
Intercept	-0.018	0.008	-2.23
$YVAR_t$	0.382	0.102	3.76
int_t	0.56	0.196	2.85
DW = 1.35	R^2(adj) = 0.55		

Comparing the Durbin Watson statistic and R^2(adj) in tables (1) and (2), one realizes that the serial correlation of the residuals declines and R^2(adj) rises. As to the non-parametric part of (11) the estimated degrees of freedom associated with the smooth term are about 2.39 and the approximate p-value[5], reflecting the significance of d_{t-1}, is about $8 \cdot 10^{-3}$.

Figure 2 shows the estimated function $f(d_{t-1})$ where the dotted lines give 95% confidence intervals. The function $f(\cdot)$ is such that its average value is equal to zero. It should be recalled that the smoothing parameter λ is hereby chosen data driven such that the GCV criterion, which is similar to (4), is minimized.

[4] All estimations were performed using the mgcv (version 1.2-3) package in R (version 2.1.0), cf. Wood (2001).

[5] It must be underlined that the p-value of the smooth term is only an approximate value (for more details see Ruppert et al., 2003, ch. 4.8).

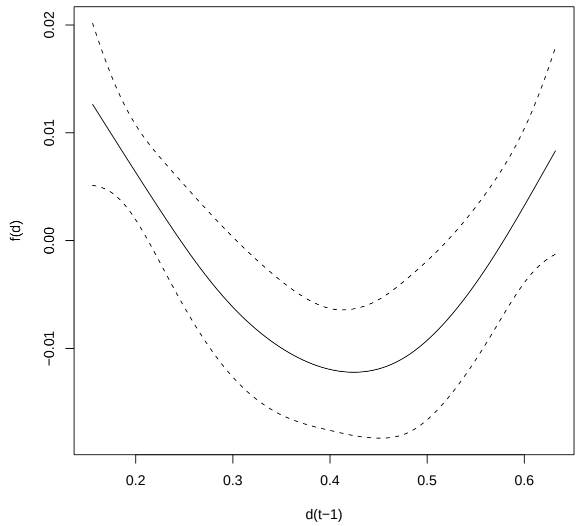

Figure 2: The estimated function $f(d_{t-1})$ with λ chosen according to the GCV criterion.

As mentioned in the last section, the smoothing parameter λ could also be chosen by hand which, then, gives different values for the GCV score as well as different degrees of freedom for the nonlinear part. Table (3) shows the relation between the value of the smoothing parameter λ and the estimated degrees of freedom as well as the GCV score. The minimum of the GCV score is obtained for $\lambda \approx 1.76 \cdot 10^{-4}$ giving about 2.39 degrees of freedom, as already mentioned in the last paragraph.

Table 3: Estimated degrees of freedom (edf) and GCV score associated with different values for the smoothing parameter λ.

λ	1	$5 \cdot 10^{-3}$	$1.76 \cdot 10^{-4}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-6}$
edf	1	1.23	2.39	4.62	7.16
GCV score	$2.43 \cdot 10^{-4}$	$2.14 \cdot 10^{-4}$	$1.67 \cdot 10^{-4}$	$1.78 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$

The next figure, figure 3, gives the estimated function $f(d_{t-1})$ depending on the value of the smoothing parameter λ where we delete the confidence interval. Panel (a) gives a linear fit which is obtained by setting $\lambda = 1$ which gives edf=1. Panel (b) and (c) show the curves obtained with $\lambda = 5 \cdot 10^{-3}$ and $\lambda = 1 \cdot 10^{-5}$, respectively. Panel (d), finally, gives the estimated curve for $\lambda = 1 \cdot 10^{-6}$. One clearly realizes how the complexity of the fitted curve increases the smaller the value of λ becomes.

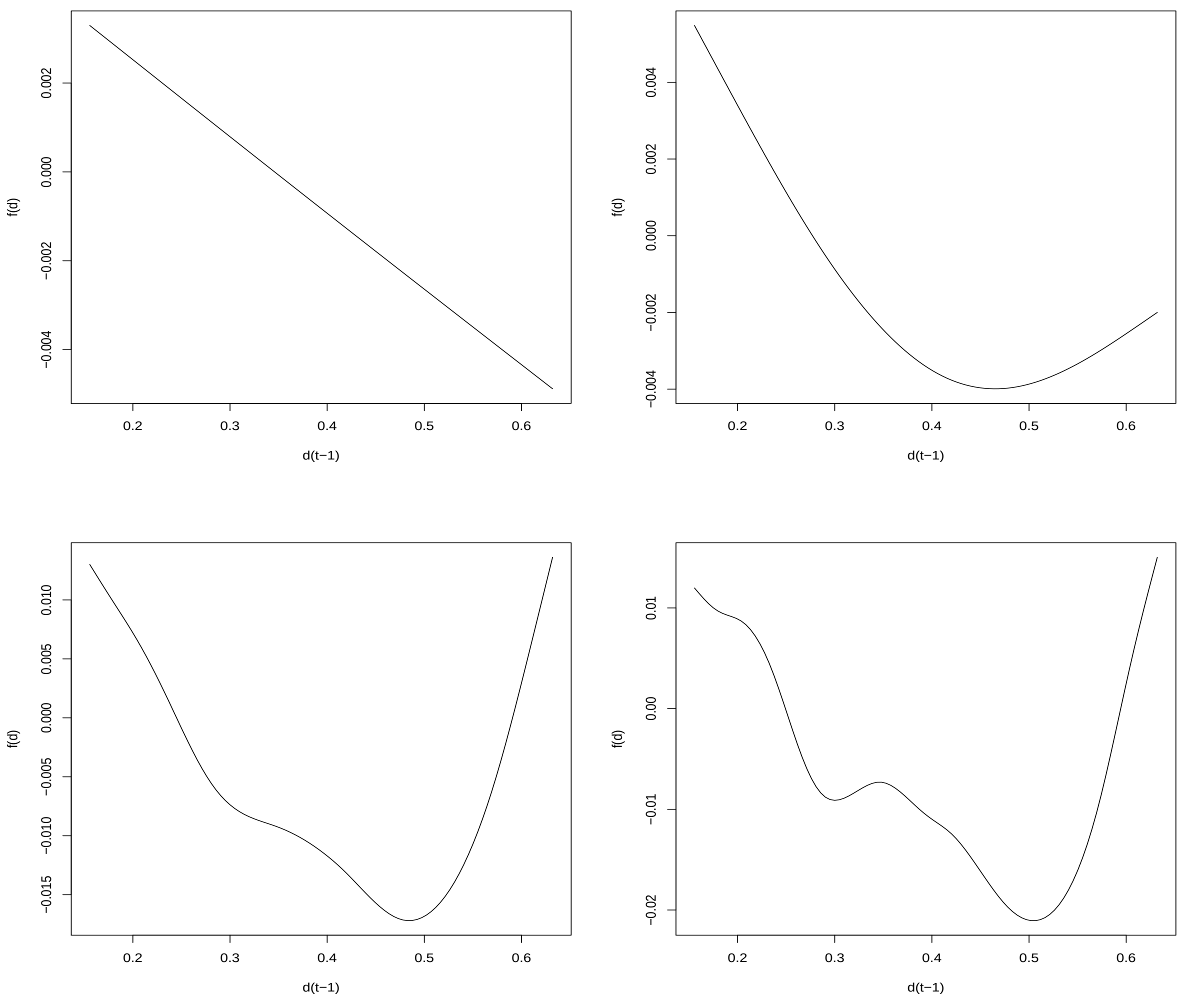

Figure 3: The estimated function $f(d_{t-1})$ for $\lambda = 1$ (a), $\lambda = 5 \cdot 10^{-3}$ (b), $\lambda = 1 \cdot 10^{-5}$ (c) and for $\lambda = 1 \cdot 10^{-6}$ (d).

4 Conclusion

This paper has given a short introduction into smoothing and penalized spline estimation. This non-parametric estimation method is more flexible than e.g. OLS estimation and may yield insights into economics which are difficult to detect with OLS or other conventional estimation techniques. The latter was demonstrated using as an example the relation between the primary surplus to GDP ratio and the debt ratio for Germany, which is characterized by a U-shaped function. The software needed to apply penalized spline estimation is open source software and can be downloaded from http://www.r-project.org/ so that its application to real-world phenomena is straightforward.

A Data source

Source: OECD Economic Outlook Statistics and Projections

We use the Data Set corresponding to those published in the June 2003 issue of the OECD Economic Outlook. Especially, we take the entire Data set for the Government Account and the series for Gross Domestic Product at Market prices (GDP). The data are available from the author upon request.

References

[1] Bohn, H. (1998), "The behaviour of U.S. public Debt and deficits" *Quarterly Journal of Economics*, Vol. 113: 949-963.

[2] (2005) Greiner, A. and G. Kauermann "Sustainability of US public debt: Estimating smoothing spline regressions." *Center for Empirical Macroeconomis*, Working Paper, No. 83, Bielefeld University, http://www.wiwi.uni-bielefeld.de/∼cem/.

[3] Hastie, T.J. and R.J. Tibshirani (1990) *Generalized Additive Models.* Chapman and Hall, London.

[4] Hastie, T.J. and R.J. Tibshirani (1993) "Varying Coefficient Models." *Journal of the Royal Statistical Society, Series B*, Vol. 55: 757-796.

[5] Kauermann, G. and G. Tutz (2000) "Local likelihood estimation in Varying–Coefficient Models including additive bias correction." *Journal of Nonparametric Statistics*, Vol. 12: 343-371.

[6] Reinsch C.H. (1967) "Smoothing by spline functions." *Numerische Mathematik*, Vol. 10: 177-183.

[7] Ruppert R., Wand M.P. and R.J. Carroll (2003) *Semiparametric Regression.* Cambridge University Press, Cambridge.

[8] Venables, W.N. and B.D. Ripley (2002) *Modern Applied Statistics with S.* (4th edition) Springer, New York.

[9] Wood, S.N. (2000) "Modelling and smoothing parameter estimation with multiple quadratic penalties." *Journal of the Royal Statistical Society*, Series B, Vol. 62: 413-428.

[10] Wood, S.N. (2001) "mgcv: GAM's and Generalized Ridge Regression for R." *R News*, Vol. 1(2): 20-25.

[11] Wood, S.N. and N.H. Augustin (2002) "GAMs with integrated model selection using penalized regression splines and applications to environmental modelling." Working paper, University of Glasgow http://www.stats.gla.ac.uk/∼simon/simon/papers/wagam.pdf. A slightly different version is published in *Ecological Modelling*, Vol. 157: 157-177.

[12] Wood, S.N. (2003) "Thin plate regression splines." *Journal of the Royal Statistical Society*, Series B, Vol. 65: 95-114.