Smoothing parameter selection and outliers accommodation for smoothing splines

Osorio, Felipe
Universidad de Valparaíso, Department of Statistics
Av. Gran Bretaña 1091, 4to piso, Playa Ancha
Valparaíso (2360102), Chile
E-mail: felipe.osorio@uv.cl

Introduction

Nonparametric regression methods using splines are attractive because they offer a flexible approach to curve fitting and frequently are used to highlight the tendencies underlying to the data. Discussions about smoothing and nonparametric regression methods can be found in Ruppert, Wand and Carroll (2003).

In order to evaluate the model assumptions and for assessing the impact of each observation on the parameter estimates the development of diagnostic procedures has become important in the context of nonparametric and semiparametric regression models (see, for instance, Eubank, 1985). This has motivated the development of robust methodologies in order to attenuate the effect of outlying and/or extreme observations. For instance, Utreras (1981) and Wei (2004) proposed robust smoothing estimation procedures considering M-estimators, whereas Cantoni and Ronchetti (2001) focused on the robust selection of the smoothing parameter. Recently, Staudenmayer, Lake and Wand (2009) described an approach for accommodating outliers in semiparametric regression based on using the Student-t distribution.

The aim of this work is to propose an alternative to outlier accommodation in the context of smoothing splines under heavy-tailed distributions. Specifically, we consider the class of scale mixture of normal distributions (Andrews and Mallows, 1974), which has been suggested in the literature as an alternative to statistical modeling when the normality assumption is violated due to the presence of outliers and/or extreme observations. In smoothing spline the selection of the smoothing parameter is crucial, and several works (Thomas, 1991) have described that the choice can be strongly affected by the presence of outliers. We believe that the formulation presented in the current work can be useful for the resistant selection of the smoothing parameter and therefore may be competitive with robust procedures that have extended the generalized cross-validation method (see, for instance, Cantoni and Ronchetti, 2001).

Model description

This work focuses on the study of the model

$$(1) \quad Y_i = g(t_i) + \epsilon_i, \qquad i = 1, \dots, n,$$

where the observations Y_i are measured at the design points t_i, with g a smooth function defined in $[a, b]$, is also assumed that the design points satisfy $a \leq t_1 < \cdots < t_n \leq b$ and $\{\epsilon_i\}$ are random variables with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \phi$. Nonparametric regression using splines, provides an estimate of g by minimizing the penalized least squares criterion,

$$(2) \quad S(g) = \sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 + \lambda \int_a^b \{g''(t)\}^2 \, \mathrm{d}t.$$

over the class of all twice differentiable functions g, here $\lambda > 0$ represents a smoothing parameter. For simplicity, assume that $g(t) = \sum_{j=1}^{p} a_j B_j(t)$, where p is an appropriately chosen number of basis functions, typically $p = n + 2$. A common choice for the basis functions $B_j(t)$ correspond to B-splines. Thus, the equation (2) can be written as:

$$(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a})^T(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a}) + \lambda \boldsymbol{a}^T\boldsymbol{P}\boldsymbol{a}$$

where $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{B} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)^T = (b_{ij}) = (B_j(t_i))$ is a $n \times p$ matrix, $\boldsymbol{a} = (a_1, \dots, a_p)^T$ and $\boldsymbol{P} = (p_{rs})$ with

$$p_{rs} = \int B_r''(t)B_s''(t)\,\mathrm{d}t, \qquad r, s = 1, \dots, p.$$

The estimator of the coefficients $\boldsymbol{a}$ is given by

$$\widehat{\boldsymbol{a}}(\lambda) = (\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{P})^{-1}\boldsymbol{B}^T\boldsymbol{Y},$$

Therefore, we can obtain an explicit expression for the *spline smoother*, as

$$\widehat{\boldsymbol{g}}(\lambda) = (\widehat{g}_\lambda(t_1), \dots, \widehat{g}_\lambda(t_n))^T = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{P})^{-1}\boldsymbol{B}^T\boldsymbol{Y} = \boldsymbol{H}(\lambda)\boldsymbol{Y},$$

where $\boldsymbol{H}(\lambda) = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{P})^{-1}\boldsymbol{B}^T$ is called the matrix of prediction or influence. That is key for choosing the smoothing parameter via the generalized cross-validation criterion (Craven and Wahba, 1979; Golub, Heath and Wahba, 1979),

$$(3) \quad GCV(\lambda) = \frac{1}{n}\frac{\|(\boldsymbol{I} - \boldsymbol{H}(\lambda))\boldsymbol{Y}\|^2}{\{1 - \operatorname{tr}(\boldsymbol{H}(\lambda))/n\}^2}.$$

There has been considerable interest in extending this kind of estimation procedures to more general settings. For instance, the penalized maximum likelihood estimation for non-gaussian data has been proposed focusing mainly in the exponential family (Wahba et al., 1995; Xiang and Wahba, 1996).

Preliminaries

We propose an alternative for accommodation of outliers in smoothing spline based on the class of scale mixture of normal distributions. Specifically, consider a random variable Z with standard normal distribution, whose density function takes the form

$$p(z) = (2\pi)^{-1/2}\exp\{-\tfrac{1}{2}z^2\}, \qquad z \in \mathbb{R}$$

and let τ a positive random variable independent of Z, with distribution function $\mathcal{H}(\tau; \boldsymbol{\theta})$.Then, the scaled random variable $U = \tau^{-1/2}Z$ is said to have a scale mixture of normal distribution (Andrews and Mallows, 1974) with density function given by

$$(4) \quad f(u) = \int_0^\infty (2\pi)^{-1/2}\tau^{1/2}\exp\{-\tfrac{1}{2}\tau z^2\}\,\mathrm{d}H(\tau).$$

When $Y \overset{\mathrm{d}}{=} \mu + \phi^{1/2}U$, where U is a random variable with density given by (4) we shall use the notation $Y \sim \mathcal{SMN}(\mu, \phi; \mathcal{H})$. Some of the most popular examples of distributions in the class defined by Equation (4), correspond to the Student-t, slash, contaminated normal and exponential power. The random variable Y can be alternatively written using the following hierarchical representation

$$(5) \quad Y|\tau \sim \mathcal{N}(\mu, \tau^{-1}\phi), \qquad \tau \sim \mathcal{H}(\tau; \boldsymbol{\nu}).$$

This kind of distributions represents an interesting alternative to the normal distribution when we are in presence of outliers and has been quite useful to statistical modeling as evidenced by a series of papers, among which it is possible to cite Lange and Sinsheimer (1993).

The EM algorithm (Dempster et al., 1977) is a popular iterative procedure for calculating parameter estimates via maximum likelihood in models with missing data or in models that can be formulated as such. This approach maximizes the log-likelihood function of observed data $\boldsymbol{Y}_{\text{obs}}$, denoted as $\ell_{\text{o}}(\boldsymbol{\theta};\boldsymbol{Y}_{\text{obs}}) = \log p(\boldsymbol{\theta};\boldsymbol{Y}_{\text{obs}})$ by considering a data augmentation scheme $\boldsymbol{Y}_{\text{aug}}$ defined such as $\boldsymbol{Y}_{\text{obs}} = \mathcal{M}(\boldsymbol{Y}_{\text{aug}})$ for some many-to-one mapping $\mathcal{M}$. The E-step of the algorithm requires the calculation of the conditional expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \text{E}\{\ell_{\text{a}}(\boldsymbol{\theta};\boldsymbol{Y}_{\text{aug}})|\boldsymbol{Y}_{\text{obs}},\boldsymbol{\theta}^{(k)}\}$, where $\ell_{\text{a}}(\boldsymbol{\theta};\boldsymbol{Y}_{\text{aug}}) = \log p(\boldsymbol{Y}_{\text{aug}};\boldsymbol{\theta})$ represents the log-likelihood function for the augmented data model and $\boldsymbol{\theta}^{(k)}$ denotes the estimate of $\boldsymbol{\theta}$ at the kth iteration; the M-Step update $\boldsymbol{\theta}^{(k+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$. Each iteration of the EM algorithm increases the likelihood function $\ell_{\text{o}}(\boldsymbol{\theta};\boldsymbol{Y}_{\text{obs}})$ and the EM sequence $\{\boldsymbol{\theta}^{(k)}\}$ converges to a stationary point of the observed likelihood under mild regularity conditions (Vaida, 2005).

The nested EM algorithm (van Dyk, 2000) is based on that under certain circumstances, it is possible to nest one EM algorithm inside another. Suppose that there are two nested data augmentation schemes $\boldsymbol{Y}_{\text{aug}_1}$ and $\boldsymbol{Y}_{\text{aug}_2}$ such as $\boldsymbol{Y}_{\text{obs}} = \mathcal{M}_1(\boldsymbol{Y}_{\text{aug}_1})$ and $\boldsymbol{Y}_{\text{aug}_1} = \mathcal{M}_2(\boldsymbol{Y}_{\text{aug}_2})$, for two many-to-one mappings $\mathcal{M}_1$ and $\mathcal{M}_2$. For computational efficiency van Dyk (2000) suggests to merge the two stages of nested EM algorithm. Thus, the kth iteration of algorithm cycle T times, following the steps:

E-step: Compute the conditional expectations

$$Q_{21}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k+\frac{t}{T})},\boldsymbol{\theta}^{(k)}) = \text{E}[\text{E}\{\ell_{\text{a}}(\boldsymbol{\theta};\boldsymbol{Y}_{\text{aug}_2})|\boldsymbol{Y}_{\text{aug}_1},\boldsymbol{\theta}^{(k+\frac{t}{T})}\}|\boldsymbol{Y}_{\text{obs}},\boldsymbol{\theta}^{(k)}].$$

M-step: Update $\boldsymbol{\theta}^{(k+\frac{t+1}{T})}$ by maximizing $Q_{21}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k+\frac{t}{T})},\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$.

Once it has finished executing the Tth cycle, let $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k+\frac{T}{T})}$ and start a new EM iteration.

Selection of the smoothing parameter

Consider the following model

$$Y_i|\boldsymbol{a},\tau_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{b}_i^T\boldsymbol{a},\phi/\tau_i), \quad \tau_i \overset{\text{ind}}{\sim} \mathcal{H}(\tau_i;\boldsymbol{\nu}), \quad \boldsymbol{a} \sim \mathcal{N}_p(\boldsymbol{0},\frac{\phi}{\lambda}\boldsymbol{P}^-), \quad i = 1,\dots,n, \tag{6}$$

where τ_i is a positive random variable with distribution function $\mathcal{H}(\tau_i;\boldsymbol{\nu})$. In order to use a nested EM algorithm we consider the augmented data vector $\boldsymbol{Y}_{\text{aug}} = (\boldsymbol{Y}^T,\boldsymbol{\tau}^T,\boldsymbol{a}^T)^T$, where $\boldsymbol{\tau} = (\tau_1,\dots,\tau_n)^T$ and $\boldsymbol{a}$ are assumed missing variables. For the model defined in (6) we consider the following nested augmentation schemes, $\boldsymbol{Y} = \mathcal{M}_1(\boldsymbol{Y}_{\text{aug}_1})$ and $\boldsymbol{Y}_{\text{aug}_1} = \mathcal{M}_2(\boldsymbol{Y}_{\text{aug}_2})$, where $\boldsymbol{Y}_{\text{aug}_2} = (\boldsymbol{Y}^T,\boldsymbol{\tau}^T,\boldsymbol{a}^T)^T$ and $\boldsymbol{Y}_{\text{aug}_1} = (\boldsymbol{Y}^T,\boldsymbol{\tau}^T)^T$. In this case, the log-likelihood function for the augmented data model, dropping out all the terms that are not functions of $\boldsymbol{\theta} = (\phi,\lambda)^T$, takes the form

$$\begin{aligned}\ell(\boldsymbol{\theta};\boldsymbol{Y}_{\text{aug}}) &= -\frac{n+p}{2}\log\phi - \frac{1}{2\phi}\sum_{i=1}^n \tau_i(Y_i - \boldsymbol{b}_i^T\boldsymbol{a})^2 + \frac{p}{2}\log\lambda - \frac{\lambda}{2\phi}\boldsymbol{a}^T\boldsymbol{P}\boldsymbol{a}\\ &= -\frac{n+p}{2}\log\phi + \frac{p}{2}\log\lambda - \frac{1}{2\phi}\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^T\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a}) + \lambda\boldsymbol{a}^T\boldsymbol{P}\boldsymbol{a}\}, \qquad (7)\end{aligned}$$

where $\boldsymbol{W} = \text{diag}(\tau_1,\dots,\tau_n)$. Following West (1984), we can show that

$$\boldsymbol{a}|\boldsymbol{Y},\boldsymbol{\tau} \sim \mathcal{N}_p(\boldsymbol{a}_W(\lambda),\phi(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda\boldsymbol{P})^{-1}), \qquad \boldsymbol{a}_W(\lambda) = (\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda\boldsymbol{P})^{-1}\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{Y}.$$

Consider $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(k+\frac{t}{T})}$, using basic results about the expectations of quadratic forms, we obtain the conditional expectations required by the internal stage of the nested EM algorithm, given by

$$\text{E}\{\boldsymbol{a}^T\boldsymbol{P}\boldsymbol{a}|\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\theta}^*\} = \phi^*\,\text{tr}(\boldsymbol{P}(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda^*\boldsymbol{P})^{-1}) + \boldsymbol{a}_W^T(\lambda^*)\boldsymbol{P}\boldsymbol{a}_W(\lambda^*),$$

and

$$\mathrm{E}\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^T\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})|\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\theta}^*\} = \phi^*\operatorname{tr}(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda^*\boldsymbol{P})^{-1}) + S_W(\boldsymbol{a}_W(\lambda^*)),$$

where $S_W(\boldsymbol{a}) = (\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^T\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})$, and the Q-function associated with the internal EM algorithm assumes the form

$$\begin{aligned}\mathrm{E}\{\ell_{\mathrm{a}}(\boldsymbol{\theta};\boldsymbol{Y}_{\mathrm{aug}_2})|\boldsymbol{Y}_{\mathrm{aug}_1},\boldsymbol{\theta}^*\} = -\frac{n+p}{2}\log\phi + \frac{p}{2}\log\lambda - \frac{1}{2\phi}\{S_W(\boldsymbol{a}_W(\lambda^*)) + \lambda\boldsymbol{a}_W^T(\lambda^*)\boldsymbol{P}\boldsymbol{a}_W(\lambda^*)\} \\ -\frac{\phi^*}{2\phi}\operatorname{tr}(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda\boldsymbol{P})(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda^*\boldsymbol{P})^{-1}.\end{aligned}$$

Finally, the conditional expectation required in the E-step of the nested EM algorithm is given by

$$\begin{aligned}Q_{21}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k+\frac{t}{T})},\boldsymbol{\theta}^{(k)}) &= \mathrm{E}[\mathrm{E}\{\ell_{\mathrm{a}}(\boldsymbol{\theta};\boldsymbol{Y}_{\mathrm{aug}_2})|\boldsymbol{Y}_{\mathrm{aug}_1},\boldsymbol{\theta}^{(k+\frac{t}{T})}\}|\boldsymbol{Y},\boldsymbol{\theta}^{(k)}] \\ &= -\frac{n+p}{2}\log\phi + \frac{p}{2}\log\lambda - \frac{1}{2\phi}\mathrm{E}\{[S_W(\boldsymbol{a}_W(\lambda^{(k+\frac{t}{T})})) + \lambda\boldsymbol{a}_W^T(\lambda^{(k+\frac{t}{T})})\boldsymbol{P}\boldsymbol{a}_W(\lambda^{(k+\frac{t}{T})})]|\boldsymbol{Y},\boldsymbol{\theta}^{(k)}\} \\ &\quad -\frac{1}{2\phi}\phi^{(k+\frac{t}{T})}\operatorname{tr}\mathrm{E}\{(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda\boldsymbol{P})(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda^{(k+\frac{t}{T})}\boldsymbol{P})^{-1}|\boldsymbol{Y},\boldsymbol{\theta}^{(k)}\}. \qquad (8)\end{aligned}$$

Fix $\lambda = \lambda^{(k+\frac{t}{T})}$, and update $\phi^{(k+\frac{t+1}{T})}$ as

$$\phi^{(k+\frac{t+1}{T})} = \frac{1}{n+p}\Big[p\phi^{(k+\frac{t}{T})} + \mathrm{E}\{S_W(\boldsymbol{a}_W(\lambda^{(k+\frac{t}{T})}))|\boldsymbol{Y},\boldsymbol{\theta}^{(k)}\} + \lambda^{(k+\frac{t}{T})}\,\mathrm{E}\{\|\boldsymbol{D}\boldsymbol{a}_W(\lambda^{(k+\frac{t}{T})})\|^2|\boldsymbol{Y},\boldsymbol{\theta}^{(k)}\}\Big], \qquad (9)$$

where $\boldsymbol{D}$ is a matrix satisfying $\boldsymbol{P} = \boldsymbol{D}^T\boldsymbol{D}$. Fixing $\phi = \phi^{(k+\frac{t+1}{T})}$, we have that $\lambda^{(k+\frac{t+1}{T})}$ is given by

$$\lambda^{(k+\frac{t+1}{T})} = \frac{p\phi^{(k+\frac{t+1}{T})}}{\mathrm{E}\{\|\boldsymbol{D}\boldsymbol{a}_W(\lambda^{(k+\frac{t}{T})})\|^2|\boldsymbol{Y},\boldsymbol{\theta}^{(k)}\} + \phi^{(k+\frac{t+1}{T})}\operatorname{tr}\boldsymbol{P}\,\mathrm{E}\{(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B}+\lambda^{(k+\frac{t}{T})}\boldsymbol{P})^{-1}|\boldsymbol{Y},\boldsymbol{\theta}^{(k)}\}}. \qquad (10)$$

The kth iteration of the EM algorithm cycle a fixed number of times T between the equations (8)-(10). Note that the expectations required to update the estimates of ϕ and λ does not have an explicit form and can be approximated using simulation. Furthermore, as suggested by van Dyk (2000), the simulation stage can be simplified performing this stage only once at the beginning of each EM iteration and reuse the generated values in each inner loop of the algorithm.

The matrix $\boldsymbol{W} = \operatorname{diag}(\tau_1,\dots,\tau_n)$ that appears in the definition of the conditional expectation $\boldsymbol{a}_W(\lambda)$ may be interpreted as a weight. In the simulation step required at the kth iteration of the nested EM, we need to draw $\boldsymbol{W}^{(r)} = \operatorname{diag}(\tau_1^{(r)},\dots,\tau_n^{(r)}), r = 1,\dots,M$, from the conditional distribution $p(\tau_i|Y_i,\boldsymbol{\theta}^{(k)})$ whose expectation tends to be inversely proportional to the distance $D_i^2(\boldsymbol{\theta}) = (Y_i - \boldsymbol{b}_i^T\boldsymbol{a})^2/\phi$, and the estimation procedure (8)-(10) tends to give smaller weight to outlying observations.

An aplication: Ridge regression

To illustrate the results we consider an experimental study on heat emission during production and hardening of 13 samples of Portland cement. Woods, Steinour and Howard (1932) considered four compounds for the clinkers from which cement is produced. The response (Y) is the emission of heat, after 180 days of curing, measured in calories per gram of cement. The predictors are the percentages of four compounds: tricalcium aluminate (X_1), tricalcium silicate (X_2), tetracalcium aluminate ferrite (X_3) and dicalcium silicate (X_4). Thus, the following model is considered

$$Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta},\phi), \quad i = 1,\dots,13; \qquad \boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{0},\frac{\phi}{\lambda}\boldsymbol{I}).$$

Woods et al. (1932) consider a linear regression model without intercept (homogeneous model). The scaled condition number is $\kappa(\boldsymbol{X}) = 9.432$, i.e. $\boldsymbol{X}$ is well conditioned. On the other hand, several authors have adopted a model with intercept (non-homogeneous model); see Kaçıranlar et al. (1999) for some references. In this case, $\kappa(\boldsymbol{X}) = 249.578$, suggesting the presence of collinearity. The increase in the scaled condition number is due to that there is an approximate linear relationship, in fact, $X_1 + X_2 + X_3 + X_4 \approx 100$. Thus, to include the intercept causes a severe collinearity.

Proceeding with the estimation of parameters using ordinary least squares (LS) and the ridge estimator. We obtain the results summarized in the table below

Parameter estimates for the Portland dataset considering different procedures

Estimator	$\widehat{\lambda}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\phi}$
LS (homogeneous)	–	–	2.193	1.153	0.759	0.486	4.047
LS (non homogeneous)	–	62.405	1.551	0.510	0.102	-0.144	3.682
Ridge (HKB)	0.0077	8.583	2.105	1.065	0.668	0.400	4.001
Ridge (GCV)	1.9716	0.085	2.165	1.159	0.738	0.490	5.090
Ridge (PML)	0.2989	0.300	2.186	1.151	0.752	0.484	0.523

For a comparison we present the parameter estimation considering the ridge estimator using GCV (Golub et al., 1979) and the HKB estimator $\widehat{\lambda}_{HKB} = ps^2/\|\widehat{\boldsymbol{\beta}}_{LS}\|^2$ proposed by Hoerl, Kennard and Baldwin (1975). As expected, the parameter estimates obtained using the proposed model, i.e., penalized maximum likelihood (PML) via the nested EM algorithm are very similar to those of the homogeneous model, in addition these estimates are more accurate.

Concluding Remarks

Gu (1992) and Xiang and Wahba (1996) among others, have extended the generalized cross validation criterion given in (3) to modelling problems that can be viewed as penalized log likelihood under non-gaussian distributions, those studies indicate that several alternatives may be considered for the selection of the smoothing parameter in semiparametric regression considering generalized linear models. Motivated by the iterative procedure that arises in the estimation in generalized linear models, several authors have proposed alternate a step in order to select the smoothing parameter by minimizing the criterion in (3). Although this procedure can be quite effective in practice, unfortunately it is not guaranteed that this strategy reaches convergence (Gu, 1992; Xiang and Wahba, 1996). On the other hand, for the procedure of smoothing parameter selection described in this work, we can ensure the convergence by using the results derived in Vaida (2005).

We notice that the proposed procedure is still valid under the normality assumption for the errors, and in our knowledge has not been previously proposed. In addition, for this case the procedure described in (8)-(10) is greatly simplified and can be seen as an interesting alternative to generalized cross validation. Currently, the author works in the influence diagnostic related to the smoothing parameter and develop a small simulation study.

REFERENCES

Andrews, D.F., and Mallows, C.L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B* **36**, 99-102.

Cantoni, E., and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing* **11**, 141-146.

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377-403.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Eubank, R.L. (1985). Diagnostics for smoothing splines. *Journal of the Royal Statistical Society, Series B* **47**, 332-341.

Golub, G.H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as method for choosing a good ridge parameter. *Technometrics* **21**, 215-223.

Green, P.J., and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman & Hall, London.

Gu, C. (1992). Cross-validating non-gaussian data. *Journal of Computational and Graphical Statistics* **1**, 169-179.

Kaçıranlar, S., Sakallioğlu, S., Akdeniz, F., Styan, G.P.H., and Werner, H.J. (1999). A new biased estimator in linear regression and a detailed analysis of the widely-analyzed dataset on Portland cement. *Sankhyā, Series B* **61**, 443-459.

Lange, K., and Sinsheimer, J.S. (1993). Normal/Independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* **2**, 175-198.

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

Staudenmayer, J., Lake, E.E., and Wand, M.P. (2009). Robustness for general design mixed models using the t-distribution. *Statistical Modeling* **9**, 235-255.

Thomas, W. (1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing. *Journal of the American Statistical Association* **86**, 693-698.

Utreras, F.I. (1981). On computing robust splines and applications. *SIAM Journal on Scientific and Statistical Computing* **2**, 153-163.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statistica Sinica* **15**, 831-840.

van Dyk, D.A. (2000). Nesting EM algorithms for computational efficiency. *Statistica Sinica* **10**, 203-225.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline ANOVA for exponential families with applications to the study of diabetic retinopathy. *The Annals of Statistics* **23**, 1865-1895.

Wei, W.H. (2004). Derivatives diagnostics and robustness for smoothing splines. *Computational Statistics & Data Analysis* **46**, 335-356.

West, M. (1984). Outliers models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B* **46**, 431-439.

Woods, H., Steinour, H.H., Starke, H.R. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Industrial and Engineering Chemistry* **24**, 1207-1214.

Xiang, D., and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica* **6**, 675-692.

ABSTRACT

It is well documented that the presence of outliers and/or extreme observations can have a strong impact on smoothing spline. This has motivated the development of robust procedures. In particular, some studies have focused on the robust selection of smoothing parameter proposing extensions of the generalized cross-validation method. In this work we consider an alternative for accommodation of outliers in spline smoothing. Our proposal is based in to consider the penalty introduced in smoothing splines as a random effect. We use a nested EM algorithm to perform the parameter estimation under distributions with tails heavier than normal. Numerical example and a simulation study illustrate the technique. We expect that this approach allows to us choose the smoothing parameter automatically and can be seen as an alternative to cross-validation.