A Note on Penalized Spline Smoothing with Correlated Errors

Tatyana Krivobokova*
Katholieke Universiteit Leuven

Göran Kauermann†
Universität Bielefeld

9th July 2007

Abstract

This note investigates the behavior of data driven smoothing parameters for penalized spline regression in the presence of correlated data. It has been shown for other smoothing methods before, that mean squared error minimizers, such as (generalized) cross validation or Akaike criterion, are extremely sensitive to misspecifications of the correlation structure over or (under) fitting the data. In contrast to this, we show that a maximum likelihood based choice of the smoothing parameter is more robust and for moderately misspecified correlation structure over or (under) fitting does not occur. This is demonstrated in simulations, data examples and supported by theoretical investigations.

Keywords: Correlation structure misspecification; Smoothing parameter selection; Linear mixed model.

*ORSTAT, K.U. Leuven, Naamsestraat 69 - bus 3555, B-3000 Leuven, Belgium. The work was conducted while the first author was affiliated to the University Bielefeld. She wishes to thank for a productive working environment. The authors are also indebted to the Deutsche Forschungsgemeinschaft (DFG) for partial support.

†Department of Economics, University of Bielefeld, Postfach 100131, D-33501 Bielefeld, Germany.

1 INTRODUCTION

Smooth nonparametric regression has achieved a considerable standard over the last two decades. In its simplest form the objective is to estimate the mean response $E(y|x) = \mu(x)$ where x is a univariate continuous covariate and function $\mu(x)$ is assumed to be smooth but otherwise unspecified. The toolbox for estimating $\mu(x)$ based on data pairs (x_i, y_i), $i = 1, \ldots, n$ is large including for instance local approaches (see Fan & Gijbels, 1996) or spline methods (see Wahba, 1990 or Eubank, 1999). In the mid nineties Eilers & Marx (1996) introduced smoothing with penalized splines (P-splines), extending an original idea of O'Sullivan (1986). This powerful and applicable technique has been explored and exploited further in a recent book by Ruppert, Wand & Carroll (2003). Regardless of the method used, a bandwidth or smoothing parameter has to be chosen to compromise goodness of fit with complexity of the estimated function. This can be done by minimizing the mean squared error (MSE) where in practice an empirical version is employed. Unfortunately, in the presence of correlated errors, that is if $\varepsilon_i = y_i - \mu(x_i)$ and ε_j for $i \neq j$ are (positively) correlated, standard smoothing parameter selectors fail to work and overfit the data. This has been nicely espoused in Opsomer, Wang & Yang (2001) for a number of smoothing techniques. Hart & Lee (2005) show that overfitting is less dominant if one-sided instead of standard (two-sided) cross validation is used. Overfitting can be generally avoided by taking the correlation structure explicitly into account for smoothing parameter selection. This has been demonstrated among others in Wang (1998) for spline smoothing and in Altman (1990), Hart (1991), Beran & Feng (2001) or Ray & Tsay (1997) for local smoothing. For penalized spline fitting Currie & Durban (2002) and Durban & Currie

(2003) present a strategy for smoothing with correlated errors and selecting the correlation structure based on the likelihood. A Bayesian approach for fitting models with correlated errors is found, for instance, in Smith, Wong & Kohn (1998).

In general, the correlation structure is unknown in advance and estimation of the correlation structure requires a sufficiently good fit of the mean function. Hence, one is faced with a dilemma in practice. In fact, even smallest misspecifications of the correlation structure can result in serious over (or under) fitting as demonstrated in Opsomer, Wang & Yang (2001). This exhibits an undesirable sensitivity of MSE-based smoothing parameter selectors. The problem of smoothing with correlated errors is most prominent in a time series setting where $x = t$ gives the time and adjacent observations y_t and y_{t+1} are correlated. Typical examples are macroeconomic time series like inflation or price indices. In this case $\mu(t)$ gives the (long term) trend which has to be estimated in the presence of correlated residuals. An overview about common trend estimates is provided, for instance, in Fan & Yao (2003). A traditional method for long term trend estimation in time series is the Hodrick & Prescott (1997) (HP) filter, which also makes use of a penalized approach. To our knowledge, however, no data driven routine for choosing the penalty parameter in the HP filter has been suggested yet and instead the choice "$\lambda = 1600$" as heuristically suggested by Hodrick & Prescott (1997) is usually used.

In this paper we investigate penalized spline smoothing using two different smoothing parameter selectors. First, a classical MSE minimizer, based on the Akaike criterion is used. Secondly, a restricted maximum likelihood (REML) smoothing parameter estimate is used by considering the smoothing model as a linear mixed model with random spline coefficient (see for instance Wand, 2003 or Kauermann, 2005). It is shown in theory and simulations that the latter approach is more recom-

mendable, since REML based smoothing parameter selection is less sensitive towards misspecifications of the correlation structure compared to MSE based choices. This means, for instance, if data have been mistakenly considered as independent while they are (not too strongly) positively correlated, this shows in a inevitable overfit using a MSE smoothing parameter selector, while the REML estimate is robust and features a satisfactory behavior. This performance is demonstrated for simulated data in Figure 1, where both smoothing parameter selectors are applied to auto-correlated errors while mistakenly assuming uncorrelated errors for the fitting. Of course, any fit using a misspecified correlation structure is inferior to one which considers the true correlation, regardless of the smoothing parameter selection being used. However, the true correlation is typically unknown (unless in simulations) so that the reported superiority of the REML provides a practical advantage when the correlation is not known. Additionally the REML based fit is available using standard software for fitting mixed models, e.g. the `lme(.)` routine in R or Splus (see Pinheiro & Bates, 2002), as demonstrated in the Appendix A.2.

The paper is organized as follows. In Section 2 we explore the beneficial behavior of the REML based smoothing parameter in theory and application. Section 3 gives some examples and extensions. Section 4 provides a discussion, while technical details and guidelines for the numerical realization are collected in the Appendix.

2 SMOOTHING PARAMETER SELECTION

2.1 Akaike and REML

We consider the smoothing model

$$\mathbf{Y} \sim N(\mu(\mathbf{x}), \sigma_\varepsilon^2 \mathbf{R}), \tag{1}$$

with $\mathbf{Y} = (y_1, \ldots, y_n)$ and $\mu(\mathbf{x}) = (\mu(x_1), \ldots, \mu(x_n))$, where $\mu(.)$ is a smooth but unknown function. The correlation matrix $\mathbf{R}$ is, like σ_ε^2, unknown. Estimation of $\mu(\mathbf{x})$ can be carried out by penalized spline smoothing, that is replacing $\mu(\mathbf{x})$ in model (1) by some high dimensional parametric structure $\mu(\mathbf{x}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$. Here $\mathbf{X}$ is a low dimensional basis, e.g. with rows $X_i^T = (1, x_i)$, while $\mathbf{Z}$ is high dimensional, e.g. truncated lines with rows $Z_i = [(x_i - \tau_1)_+, \ldots, (x_i - \tau_K)_+]$, where τ_j are fixed knots, $j = 1, \ldots, K$ and $(x)_+ = \max\{x, 0\}$. Alternatively, one can work with B-splines (de Boor, 1978) as suggested in Eilers & Marx (1996). For theoretical investigation the use of truncated polynomials proves, however, to be simpler and is therefore preferred here. With respect to the dimension K we follow Ruppert (2002) who has shown, that the actual choice of K and the location of knots have little influence on the resulting penalized fit as long as K is large, e.g. $K = \min(n/4, 40)$. A more theoretical exercise is to let the number of knots grow with the sample size. Some first results are found in Cardot (2002) and Hall & Opsomer (2005). Regardless of these new theoretical development the practical implication is that K is far less than n.

Coefficients $\boldsymbol{\beta}$ and $\mathbf{u}$ are estimated from the penalized log likelihood function

$$\begin{aligned} l_p(\boldsymbol{\beta}, \mathbf{u}; \sigma_\varepsilon^2, \mathbf{R}, \lambda) &= -\frac{1}{2}\left\{n \log(\sigma_\varepsilon^2) + \log|\mathbf{R}| + (\mathbf{Y} - \mathbf{C}\boldsymbol{\theta})^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{C}\boldsymbol{\theta})/\sigma_\varepsilon^2\right\} \\ &\quad -\frac{\lambda}{2\sigma_\varepsilon^2}\mathbf{u}^T \tilde{\mathbf{D}}\mathbf{u}, \end{aligned} \tag{2}$$

where $\mathbf{C} = (\mathbf{X}, \mathbf{Z})$, coefficient $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \mathbf{u}^T)^T$ and $\tilde{\mathbf{D}}$ is an appropriately chosen penalty matrix. A conventional choice for truncated polynomial basis is the identity matrix, i.e. $\tilde{\mathbf{D}} = \mathbf{I}_K$. Other choices related to B-splines are suggested in Eilers & Marx (1996). The penalty term $\lambda \mathbf{u}^T \tilde{\mathbf{D}}\mathbf{u}$ works as a ridge or shrinkage effect and

penalizes the coefficients of basis $\mathbf{Z}$ only. Note, that the estimate $\hat{\mu}(\mathbf{x})$ results as

$$\hat{\mu}_\lambda(\mathbf{x}) = \mathbf{C}\hat{\boldsymbol{\theta}} = \mathbf{C}\left(\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C} + \lambda\mathbf{D}\right)^{-1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{Y} =: \mathbf{S}_{R,\lambda}\mathbf{Y}, \tag{3}$$

with $\mathbf{D}$ as a block diagonal matrix built from $\mathbf{0}$ and $\tilde{\mathbf{D}}$ with matching dimensions. With $\mathbf{S}_{R,\lambda}$ we denote the resulting smoothing matrix. The penalty parameter λ thereby steers the amount of smoothness. A data driven choice for λ is available by minimizing the Akaike criterion

$$AIC(\mathbf{R},\lambda) = n\log\{RSS(\mathbf{R},\lambda)\} + 2df(\mathbf{R},\lambda), \tag{4}$$

where $RSS(\mathbf{R},\lambda) = \{\mathbf{Y} - \hat{\mu}_\lambda(\mathbf{x})\}^T\,\mathbf{R}^{-1}\,\{\mathbf{Y} - \hat{\mu}_\lambda(\mathbf{x})\}$ and $df(\mathbf{R},\lambda) = \mathrm{tr}(\mathbf{S}_{R,\lambda})$, with tr(.) as trace of the matrix. Alternatively, a modified version of the criterion suggested by Simonoff & Tsay (1999) can be used.

The penalized fit can also be motivated by treating $\mathbf{u}$ as random coefficient leading to the linear mixed model

$$\mathbf{Y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2\mathbf{R}), \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\tilde{\mathbf{D}}^-), \tag{5}$$

where $\tilde{\mathbf{D}}^-$ is the (generalized) inverse of $\tilde{\mathbf{D}}$. In this case, $\hat{\mu}_\lambda(\mathbf{x})$ as given in (3) results as posterior Bayes estimate or Best Linear Unbiased Predictor (BLUP) (see Searle, Casella, & McCulloch, 1992) with $\lambda = \sigma_\varepsilon^2/\sigma_u^2$. Model (5) affords an estimate of the smoothing parameter λ by maximizing the likelihood resulting from the linear mixed model. In practice, an adjusted restricted maximum likelihood (REML, see Harville, 1977) shows advantages. In this case λ is chosen by minimizing the negative REML function.

$$-2\,REML(\mathbf{R},\lambda) = (n-p)\log(\hat{\sigma}^2_{\varepsilon,MM}) + \log|\mathbf{V}_{R,\lambda}| + \log|\mathbf{X}^T\,\mathbf{V}^{-1}_{R,\lambda}\,\mathbf{X}|, \tag{6}$$

with p as dimension of $\boldsymbol{\beta}$, $\hat{\sigma}^2_{\varepsilon,MM} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{V}_{R,\lambda}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$ as variance estimate in the mixed model (5) and $\mathbf{V}_{R,\lambda} = \mathbf{R} + \mathbf{Z}\tilde{\mathbf{D}}^{-}\mathbf{Z}^T/\lambda$.

2.2 Smoothing with misspecified correlation

The true correlation structure of the data is typically unknown and for estimation some "working correlation" $\tilde{\mathbf{R}}$, which is supposed to be close to $\mathbf{R}$, has to be used in the smoothing parameter selection step. Our objective is to explore which of the two above smoothing parameter selectors is more sensitive with respect to misspecifications of such working correlation. Without loss of generality we explore this point using working zero correlation, that is $\tilde{\mathbf{R}} = \mathbf{I}_n$, with $\mathbf{I}_n$ as identity matrix. Note that if a different working correlation $\tilde{\mathbf{R}}$ is used, then observations $\mathbf{Y}^* = \tilde{\mathbf{R}}^{-1/2}\mathbf{Y}$ show working zero correlation with mean function $\mu^*_\lambda(\mathbf{x}) = \tilde{\mathbf{R}}^{-1/2}\mu_\lambda(\mathbf{x})$. This implies that the results derived for the zero working correlation can be directly transferred to more general settings. For our further investigation we make the following assumptions:

(A1) The values of x_i, $i = 1, ..., n$ are ordered and equidistant with $x_i \in [0, 1]$, for simplicity.

(A2) We denote with $\mathbf{R}$ the true (unknown) correlation of the residuals and assume that $R_{ij} = r(|i-j|)$ with r as some stationary positive correlation function, descending to zero for $|i-j|$ growing. Note that this implies that the correlation between two fixed points in $[0, 1]$ is decreasing as $n \to \infty$. We parameterize $\mathbf{R} = \mathbf{R}(\boldsymbol{\varrho})$ with some vector $\boldsymbol{\varrho} = (\varrho_1, ..., \varrho_{n-1})$, so that $\varrho_i = r(i)$, $i = 1, ..., n-1$ and $\mathbf{R}(\boldsymbol{\varrho} = 0) = \mathbf{I}_n$.

(A3) For the mixed model (5) we assume $\sigma_u^2 > 0$.

(A3$'$) In model (1) we assume $\mu(\mathbf{x}) = X\boldsymbol{\beta} + \mathbf{Zu}$ and in particular $\|\mathbf{u}\| > 0$.

Note that assumptions (A3) and (A3$'$) guarantee that the corresponding model does not collapse to a simple parametric model. We will now explore the performance of the smoothing parameters by making use of the following notation:

- $s_{REML}(\lambda) := -2\lambda\, \hat{\sigma}^2_{\varepsilon,MM}\, \partial REML(\lambda)/\partial\lambda$ is the estimating equation for the REML based smoothing parameter defined in the model (5) with $\mathbf{R} = \mathbf{I}_n$.
- $s_{AIC}(\lambda) := \lambda/2\, \hat{\sigma}^2_{\varepsilon}\, \partial AIC(\lambda)/\partial\lambda$ is the estimating equation for the AIC based smoothing parameter defined in the model (1) with $\mathbf{R} = \mathbf{I}_n$.
- λ^0_{REML} and λ^0_{AIC} denote the "true" values of the smoothing parameters under the according models with independent residuals, implicitly defined through $\mathrm{E}_{Y,u}\{s_{REML}(\lambda^0_{REML})|\mathbf{R} = \mathbf{I}_n\} = 0$ and $\mathrm{E}_{Y|u}\{s_{AIC}(\lambda^0_{AIC})|\mathbf{R} = \mathbf{I}_n\} = 0$, respectively. Note that with assumptions (A3) and (A3$'$) we have that $\lambda^0_{REML} > 0$ and $\lambda^0_{AIC} > 0$.
- λ^{ϱ}_{REML} and λ^{ϱ}_{AIC} are the "true" smoothing parameters under the misspecified correlation model, i.e. they satisfy the equations $\mathrm{E}_{Y,u}\{s_{REML}(\lambda^{\varrho}_{REML})|\mathbf{R}\} = 0$ and $\mathrm{E}_{Y|u}\{s_{AIC}(\lambda^{\varrho}_{AIC})|\mathbf{R}\} = 0$, respectively.

Note that the smoothing parameter estimates are implicitly defined through $s_{AIC}(\hat{\lambda}_{AIC}) = 0$ and $s_{REML}(\hat{\lambda}_{REML}) = 0$, respectively. To this end it is important to reflect that both smoothing parameters are defined in different stochastic frameworks. While we take expectation with respect to $\mathbf{Y}$ and $\mathbf{u}$ for the REML estimate, for the Akaike based smoothing parameter coefficient vector $\mathbf{u}$ is treated as given. This is also

indicated with subscripts at the expectation symbols above. For our theoretical investigation we now have to assume one of the two models. It has been shown in Kauermann (2005) that the MSE based smoothing parameter does not equal the smoothing parameter based on REML for fixed $\mathbf{u}$ and no general theoretical finite sample comparison is possible since the behavior depends primarily on the unknown function $\mu(\mathbf{x})$. In contrast assuming the mixed model (5) we can compare the two approaches on a theoretical ground. We pursue both frameworks here. First, we assume model (5) and show why the REML smoothing parameter choice is less sensitive towards correlation misspecification. Second, we exemplify the theoretical findings for particular functions $\mu(\mathbf{x})$ using model (1) and a finite sample situation. The second part is for reasons noted above more heuristic but concludes with the same superiority of the REML based choice.

We start by defining $\bar{\lambda}^0_{AIC} := \mathrm{E}_u(\lambda^0_{AIC})$ as mean value of λ^0_{AIC} taking $\mathbf{u}$ as random. As shown in the Appendix we obtain $\bar{\lambda}^0_{AIC} = \lambda^0_{REML}$. The objective is now to investigate the performance of λ^{ϱ}_{REML} and $\bar{\lambda}^{\varrho}_{AIC} := \mathrm{E}_u(\lambda^{\varrho}_{AIC})$.

Theorem. Under assumptions (A1) to (A3) we find

$$\begin{aligned} &\lambda^{\varrho}_{REML} - \bar{\lambda}^{\varrho}_{AIC} \qquad (7)\\ &= 2\lambda \sum_{i=1}^{n-1} \varrho_i \left[\frac{\mathrm{tr}\left\{\mathbf{A}_i \mathbf{S}_\lambda (\mathbf{I}_n - \mathbf{S}_\lambda)^2\right\}}{\mathrm{tr}\{\mathbf{S}^2_\lambda (\mathbf{I}_n - \mathbf{S}_\lambda)\}} - \frac{\mathrm{tr}\left\{\mathbf{A}_i \mathbf{S}_\lambda (\mathbf{I}_n - \mathbf{S}_\lambda)\right\}}{\mathrm{tr}(\mathbf{S}^2_\lambda) - p} \right] + O(\boldsymbol{\varrho}^T \boldsymbol{\varrho} + n^{-1}), \end{aligned}$$

with $\lambda = \bar{\lambda}^0_{AIC} = \lambda^0_{REML}$, $\mathbf{S}_\lambda = \mathbf{S}_{R=I,\lambda}$ and $\mathbf{A}_i$ as a lower shift matrix with ones on i-th sub-diagonal.

The proof is provided in the Appendix. Formula (7) shows how the difference between the smoothing parameters changes with the misspecified correlation $\boldsymbol{\varrho}$, which we now want to explore further. It seems hardly possible to do this for any general

correlation structure, so that we restrict the investigation to correlation matrices $\mathbf{R}$ of the following special structure.

(A2′) Additional to assumption (A2) we postulate that $\varrho_i = r(i) = O(\delta^i)$, $i = 1, ..., n-1$ for some $0 < \delta < 1$.

Assumption (A2′) is apparently not too restrictive. It holds for instance, for stationary Markov processes as can be seen as follows. Consider a stationary process of order d, say, so that ε_i given $(\varepsilon_{i-1}, ...\varepsilon_{i-d})$ is independent of $(\varepsilon_{i-d-1}, \varepsilon_{i-d-2}, ...)$. In this case $\mathbf{R}^{-1}$ is of block band diagonal structure with bandwidth d. Let M be the maximum element of the diagonal of $\mathbf{R}$. Assuming that M does not depend on n and $\|\mathbf{R}/M\| \leq 1$ for some matrix norm we find with the results in Demko (1977) that $\varrho_i \leq c\,\delta^i$, with $0 < \delta < 1$ and the constant c depending on M and d only.
We are particularly interested in small misspecifications, which is mirrored in small values of δ, and with $\delta \to 0$ we get $\mathbf{R} \to \mathbf{I}_n$. This allows us to formulate the following Corollary.

Corollary. Assuming (A1), (A2′) and (A3) we find

$$\begin{aligned} \lambda^{\varrho}_{REML} &= \lambda + \Delta_{REML}\,\varrho_1 + O(\delta^2 + n^{-1}) \\ \bar{\lambda}^{\varrho}_{AIC} &= \lambda + \Delta_{AIC}\,\varrho_1 + O(\delta^2 + n^{-1}), \end{aligned}$$

with $\lambda = \bar{\lambda}^0_{AIC} = \lambda^0_{REML}$, $\Delta_{REML} = \partial\lambda^{\varrho}_{REML}/\partial\varrho_1|_{\varrho_1=0}$, $\Delta_{AIC} = \partial\bar{\lambda}^{\varrho}_{AIC}/\partial\varrho_1|_{\varrho_1=0}$ and it holds

$$\Delta := |\Delta_{AIC}| - |\Delta_{REML}| = 2\lambda \frac{\sum_{j=1}^{K}\sum_{j<l\leq K} b_j b_l (b_l - b_j)^2}{\sum_{j=1}^{K}\sum_{l=1}^{K} b_j^2 b_l^2 (1 - b_j)} \left\{1 - O(n^{-1})\right\} > 0, \quad (8)$$

with $b_j = (1 + \lambda e_j)^{-1}$ where $e_j \geq 0$ are the eigenvalues of the singular value decomposition $\mathbf{B}^T\mathbf{D}\mathbf{B}^{-1} = \mathbf{U}\mathrm{diag}(e_j)\mathbf{U}^T$ for a square and invertible matrix $\mathbf{B}$ obtained from a Cholesky decomposition $\mathbf{B}^T\mathbf{B} = \mathbf{C}^T\mathbf{C}$.

The proof is provided in the Appendix. Note also the following remarks, which relate to the above results.

1. Since Δ_{REML} and Δ_{AIC} measure changes in the corresponding smoothing parameters if the correlation parameter changes, the positive value of Δ as obtained in (8) reflects the stronger sensitivity of the AIC based smoothing parameter towards misspecified correlation.

2. Assuming ϱ_1 to be small, λ^{ϱ}_{REML} as well as $\bar{\lambda}^{\varrho}_{AIC}$ are nearly linear functions of ϱ_1 which is also visualized in simulations in the next section. Moreover,

$$\lambda^{\varrho}_{REML} - \bar{\lambda}^{\varrho}_{AIC} = \Delta\varrho_1 + O(\delta^2 + n^{-1}).$$

3. The result can be extended to different empirical MSE minimizer like the modified Akaike criterion suggested by Simonoff & Tsay (1999) or generalized cross validation. Details are found in the supplementary material to this paper provided under www.amstat.org/publications/jasa/supplemental_materials.

Before demonstrating the above results in simulations we want to visualize the different performance of both smoothing parameters using the theoretical grounds from above. For some fixed $\lambda_0 = \sigma^2_\varepsilon/\sigma^2_u$ we take the mixed model (5) and calculate the expected estimating function $\mathrm{E}_{Y,u}\{s_{REML}(\lambda)|\mathbf{R}\}$ for different values of λ. The shape of the function is shown in Figure 2 (top plot) as solid line, where it is calculated also under the misspecified models, an AR(1) model with first order correlation $\varrho = 0.1$ and $\varrho = 0.2$. We thereby used truncated basis of order two based on 40 knots. For $\varrho = 0$ we see the expected estimated score function which cuts the x axis at the true parameter value λ_0. Small changes of the correlation have only a minor impact on the root of the function. This is in contrast to the course of $\mathrm{E}_{Y,u}\{s_{AIC}(\lambda)|\mathbf{R}\}$ also shown

in Figure 2 top plot as dashed line, again for correlations $\varrho = 0$, $\varrho = 0.1$ and $\varrho = 0.2$. Here the root of the function depends quite strongly on the misspecified correlation which mirrors in apparent overfitting of the smooth estimate. We can visualize the same mean estimating functions by considering $\mathbf{u}$ as fixed, i.e. $\mathrm{E}_{Y|u}\{s_{REML}(\lambda)|\mathbf{R}\}$ and $\mathrm{E}_{Y|u}\{s_{AIC}(\lambda)|\mathbf{R}\}$. This means we use the smoothing model (1) with a penalized estimate instead of the mixed model (5) for the calculation of the mean. As true function $\mu(\mathbf{x})$ we use the sine curve as already seen in Figure 1. In this case we obtain the graphs of two functions shown in Figure 2 bottom plot. In this sine curve example the two plots in Figure 2 look much alike, that is whether we condition on coefficients $\mathbf{u}$ or consider them as random makes no difference on the shape of the estimating functions. Fixing $\mathbf{u}$ at different values, that is taking different mean functions $\mu(\mathbf{x})$, can impose differences between the two curves. The overall performance based on the numerous functional examples we investigated remains however the same, i.e. the root of the estimating equation changes more severely for $s_{AIC}(\lambda)$ than for $s_{REML}(\lambda)$. We refer to the supplementary material provided with this paper under www.amstat.org/publications/jasa/supplemental_materials.

2.3 Simulation

To illustrate the theoretical findings we ran a number of simulation studies some of which are reported here. Following Wang (1998) and Currie & Durban (2002), we generate $n = 300$ data points with $y_i = \sin(2\pi i/n) + 0.3\varepsilon_i$, where ε_i, $i = 1, ..., n$ are drawn from a first-order autoregressive process with mean zero, standard deviation one and first-order autocorrelation equal to 0.4. Figure 1 shows exemplary one simulation. The smooth fit is based on a quadratic truncated polynomial basis with $K = 40$ knots placed equidistantly over the observed x values. The smoothing

parameters are selected assuming independence. Clearly, the AIC based choice fails to estimate the function properly while the REML estimated smoothing parameter behaves well.

We rerun the simulation with different values for the autocorrelation, ranging from 0 to 0.4 with step size 0.05. The top plot in Figure 3 shows the average of the simulated smoothing parameters $\hat{\lambda}_{AIC}$ and $\hat{\lambda}_{REML}$ on a log scale based on 100 simulations. The vertical lines correspond to the interquartile range (note that the REML estimate clearly exhibits less variability, which is however not further discussed in this paper, see also Kauermann, 2005). It appears that $\hat{\lambda}_{AIC}$ reacts stronger on the misspecified correlation structure. In contrast the REML based smoothing parameter behaves clearly more inertially. Note that the behavior of both smoothing parameter estimates is almost linear in the correlation parameter, reproducing our theoretical findings. Even though our focus is on the behavior of the smoothing parameters λ_{AIC} and λ_{REML}, from a practical viewpoint there is greater interest in the effect on the Mean Squared Error of the resulting estimates. This is visualized for the above simulations in the bottom plot of Figure 3, where we show the term $\sum_{i=1}^{n}\{\hat{\mu}(x_i)-\mu(x_i)\}^2/n$ for the different smoothing parameters. Clearly, even small omitted correlations among the residuals have an impact on the Mean Squared Error of the AIC based fit, while the REML based choice performs more stable.

We ran a number of other simulations with (i) different numbers of knots, (ii) different functional forms, (iii) different residual variability (i.e signal to noise ratio), (iv) different basis functions (e.g. B-splines) and (v) different MSE-based smoothing parameter selectors (e.g. GCV and modified AIC, suggested by Simonoff & Tsay, 1999). The findings were the same as those reported here and these factors did hardly change the general behavior. The superiority of the REML approach is always

clearly seen. The interested reader may consult our supplementary material to this paper provided under www.amstat.org/publications/jasa/supplemental_materials. The behavior of the REML estimate is also superior in more complex correlation scenarios. To demonstrate this we simulated data from an AR(2) process with first and second order autocorrelation 0.4 and 0.3, respectively. For fitting we employed a (misspecified) AR(1) correlation structure with first order autocorrelation estimated from the data. Note, that this is easily accommodated in the linear mixed model framework and implemented for instance in the Splus or R `lme(.)` function (see Appendix A.2 and Pinheiro & Bates, 2002). The resulting fit is shown in Figure 4, top row plots. The data clearly exhibit an AR(2) process in the correlation structure. If this is however not correctly specified, the AIC smoothing parameter choice suffers from overfitting. This is in contrast to the REML selected λ which works fine even for misspecified correlation. We rerun the simulation for different values of the second order autocorrelation, ranging from 0 (which equals AR(1)) to 0.4. The plots of Figure 5 show the resulting smoothing parameter estimates and Mean Squared Errors, respectively, if the data are in fact fitted with a misspecified AR(1) structure. The weak dependence of the REML estimate on the correlation structure is again visible and confirms our theoretical findings.

3 EXAMPLES AND EXTENSIONS

3.1 Examples

To demonstrate the applicability of the described property of the REML estimator for smoothing parameter selection we first analyze some data obtained from the International Statistical Yearbook. Average monthly data on electricity consumption

in Italy in the period January 1986 to May 2003 are presented in Figure 6. Estimating the data assuming independent residuals yields the overfitting AIC based estimate (left hand side) and the satisfactory REML fit (right hand side), both seen as solid lines. Examination of the partial autocorrelation function of the residuals corresponding to the REML fit suggests that the data are AR(1) correlated with the first order correlation about 0.4. Refitting the data taking an AR(1) structure into account leads to the dashed line fits shown in Figure 6. Note that for the REML based fit both estimates (without and with accounting for correlation) are nearly indistinguishable, although the correlation structure is wrongly specified in the first case.

In our second example we consider data obtained from International Financial Statistics (IFS) - service of the International Monetary Fund. We analyze 184 average monthly observations of import prices in Germany in the period January 1990 to April 2005 (basis year 2000), see Figure 7. The data clearly exhibit correlation and we start fitting the data by using a REML based smoothing parameter (upper right plot) assuming an AR(1) correlation structure for the residuals (solid line). The plot of the partial autocorrelation functions (bottom right plot) provides evidence that in fact an AR(2) structure looks more suitable. We refitted the model with a REML based smoothing parameter but now assuming an AR(2) structure. The resulting fit is shown in the top right plot of Figure 7 as dashed line. It looks more appropriate and does not change the autocorrelation function in a notable amount. In contrast to the REML estimate the AIC based fit with an AR(1) correlation structure on the left hand side plots is clearly not appropriate and does not even help to discover the underline correlation structure as can be seen from the autocorrelation function shown in the bottom row.

These examples suggest a simple strategy for the mean estimation of correlated data. First, fit the model with a mixed model software, assuming the most probable correlation structure and inspect whether the residuals behave in accordance with this assumption. If the correlation structure is only moderately misspecified, the mean estimate with the REML based smoothing parameter will still be appropriate and examination of the (partial) autocorrelation functions could help to determine the true correlation structure of the data.

For further application we also refer to Krivobokova, Kauermann & Archontakis (2006), where a two dimensional fit of the term structure of interest rates with non-standard correlation structure was performed. Again, it was the REML based smoothing parameter which made the fit possible.

3.2 Extensions

The advantages of REML based estimates of smoothing parameters extend to additive and generalized response models. We sketch the ideas for the latter here. We assume that $g\{E(y_i|x_i)\} = g\{\mu(x_i)\} = \eta(x_i)$ with $g(.)$ as known link function and y given x distributed according to an exponential family distribution. We replace $\eta(\mathbf{x}) = (\eta(x_1), ..., \eta(x_n))$ like above with $\mathbf{C}\boldsymbol{\theta}$ and impose a penalty on coefficient vector $\mathbf{u}$. Formulating the latter as a priori normal distribution leads to the Generalized Linear Mixed Models (GLMM)

$$g\{E(\mathbf{Y}|\mathbf{x}, \mathbf{u})\} = \mathbf{C}\boldsymbol{\theta}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \tilde{\mathbf{D}}^-). \tag{9}$$

Estimation in (9) can be carried out using a Laplace approximation to achieve the marginal likelihood. This approach is better known under the phrase penalized quasi likelihood (PQL) and is extensively discussed in Breslow & Clayton

(1993). The Laplace approach is asymptotically justifiable if we assume the number of basis functions K to be bounded for growing sample size. Practically this means that $K \ll n$. Particularly, this setting circumvents the problems listed for instance in Breslow & Lin (1995) or more generally in Shun & McCullagh (1995). In particular, the PQL approach provides a REML based estimation for σ_u^2 which proves to behave satisfactory if the data are in fact correlated. An illustration of the procedure is provided with the supplementary material to the paper (www.amstat.org/publications/jasa/supplemental_materials).

4 DISCUSSION

We investigated the sensitivity to misspecified correlation of two data-driven smoothing parameter selectors for penalized spline smoothing - Akaike and REML. It has been shown that the AIC based smoothing parameter is (on average) more affected by the presence of correlated errors than the REML based smoothing parameter. The difference of both smoothing parameters obtained under misspecified correlation was evaluated and the stronger dependence of the AIC based smoothing parameter on the misspecified correlation was quantified. The findings were supplemented by real data examples.

A APPENDIX

A.1 Technical details

In the subsequent proofs we make use of the following relationships:

$$\begin{aligned}
\frac{\partial(\log|\mathbf{X}^T\mathbf{V}_{R,\lambda}^{-1}\mathbf{X}|)}{\partial\lambda} &= \frac{1}{\lambda}\mathrm{tr}\{\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\lambda\mathbf{D})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}_{R,\lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_{R,\lambda}^{-1}\},\\
\frac{\partial(\log|\mathbf{V}_{R,\lambda}|)}{\partial\lambda} &= -\frac{1}{\lambda}\mathrm{tr}\{\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\lambda\mathbf{D})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\},\\
\frac{\partial\mathrm{tr}(\mathbf{S}_{R,\lambda})}{\partial\lambda} &= -\frac{1}{\lambda}\mathrm{tr}(\mathbf{S}_{R,\lambda}-\mathbf{S}_{R,\lambda}\mathbf{S}_{R,\lambda}),\\
\hat{\boldsymbol{\theta}}^T\mathbf{D}\hat{\boldsymbol{\theta}} &= \frac{1}{\lambda}\mathbf{Y}^T(\mathbf{S}_{R,\lambda}-\mathbf{S}_{R,\lambda}\mathbf{S}_{R,\lambda})\mathbf{Y},\\
\hat{\boldsymbol{\theta}}^T\mathbf{D}(\mathbf{I}_{K+p}-\tilde{\mathbf{S}}_{R,\lambda})\hat{\boldsymbol{\theta}} &= \frac{1}{\lambda}\mathbf{Y}^T(\mathbf{I}_n-\mathbf{S}_{R,\lambda})\mathbf{S}_{R,\lambda}(\mathbf{I}_n-\mathbf{S}_{R,\lambda})\mathbf{Y},\\
\hat{\sigma}^2_{\varepsilon,MM} &= \frac{\mathbf{Y}^T(\mathbf{I}_n-\mathbf{S}_{R,\lambda})\mathbf{Y}}{n-p},
\end{aligned}$$

where $\tilde{\mathbf{S}}_{R,\lambda} = (\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}+\lambda\mathbf{D})^{-1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}$.

Proof of Theorem.

Differentiation of (6) results in the score equation

$$-2\frac{\partial REML(\lambda)}{\partial\lambda} = \frac{\hat{\boldsymbol{\theta}}^T\mathbf{D}\hat{\boldsymbol{\theta}}}{\hat{\sigma}^2_{\varepsilon,MM}} - \frac{1}{\lambda}\{\mathrm{tr}(\mathbf{S}_{R,\lambda})-p\} = 0. \tag{10}$$

Accordingly, differentiating (4) we obtain the estimating equation

$$\frac{\partial AIC(\lambda)}{\partial\lambda} = \frac{2\hat{\boldsymbol{\theta}}^T\mathbf{D}(\mathbf{I}_{K+p}-\tilde{\mathbf{S}}_{R,\lambda})\hat{\boldsymbol{\theta}}}{\hat{\sigma}^2_{\varepsilon}} - \frac{2}{\lambda}\mathrm{tr}\{S_{R,\lambda}(\mathbf{I}_n-\mathbf{S}_{R,\lambda})\} = 0, \tag{11}$$

with $\hat{\sigma}^2_{\varepsilon} = RSS(\mathbf{R},\lambda)/\{n-\mathrm{tr}(\mathbf{S}_{R,\lambda})\}$. Setting in (10) and (11) $\mathbf{R}=\mathbf{I}_n$ we find the corresponding estimating functions $s_{REML}(\lambda) = \lambda\,\hat{\boldsymbol{\theta}}\mathbf{D}\hat{\boldsymbol{\theta}} - \hat{\sigma}^2_{\varepsilon,MM}\{\mathrm{tr}(\mathbf{S}_\lambda)-p\}$ and $s_{AIC}(\lambda) = \lambda\,\hat{\boldsymbol{\theta}}\mathbf{D}(\mathbf{I}_{K+p}-\tilde{\mathbf{S}}_\lambda)\hat{\boldsymbol{\theta}} - \hat{\sigma}^2_{\varepsilon}\mathrm{tr}\{\mathbf{S}_\lambda(\mathbf{I}_n-\mathbf{S}_\lambda)\}$. Under the mixed model (5) the data are distributed according to $\mathbf{Y}|\mathbf{R}\sim N(\mathbf{X}\boldsymbol{\beta},\sigma^2_\varepsilon\mathbf{R}+\sigma^2_\varepsilon\mathbf{Z}\tilde{\mathbf{D}}^-\mathbf{Z}^T/\lambda)$ with

$\lambda = \bar{\lambda}^0_{AIC} = \lambda^0_{REML}$. Simple calculations now yield the implicit definition of λ^{ϱ}_{REML} through

$$\begin{aligned} 0 &= \mathrm{E}_{Y,u}\left\{s_{REML}(\lambda^{\varrho}_{REML})|\mathbf{R}\right\} \\ &= \sigma^2_{\varepsilon}\left[\mathrm{tr}\{(\mathbf{R}-\mathbf{I}_n)\mathbf{S}_{\lambda^{\varrho}_{REML}}(\mathbf{I}_n-\mathbf{S}_{\lambda^{\varrho}_{REML}})\} - \frac{\lambda-\lambda^{\varrho}_{REML}}{\lambda}\{\mathrm{tr}(\mathbf{S}^2_{\lambda^{\varrho}_{REML}})-p\}\right] \\ &+ \sigma^2_{\varepsilon}\left\{\mathrm{tr}(\mathbf{S}_{\lambda^{\varrho}_{REML}})-p\right\}\frac{\mathrm{tr}(\mathbf{R}\mathbf{S}_{\lambda^{\varrho}_{REML}})-p-\lambda^{\varrho}_{REML}/\lambda\{\mathrm{tr}(\mathbf{S}_{\lambda^{\varrho}_{REML}})-p\}}{n-p}. \quad (12) \end{aligned}$$

Accordingly, $\bar{\lambda}^{\varrho}_{AIC}$ is defined through

$$\begin{aligned} 0 &= \mathrm{E}_{Y,u}\left\{s_{AIC}(\bar{\lambda}^{\varrho}_{AIC})|\mathbf{R}\right\} \\ &= \sigma^2_{\varepsilon}\left[\mathrm{tr}\{(\mathbf{R}-\mathbf{I}_n)\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}}(\mathbf{I}_n-\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}})^2\} - \frac{\lambda-\bar{\lambda}^{\varrho}_{AIC}}{\lambda}\mathrm{tr}\{\mathbf{S}^2_{\bar{\lambda}^{\varrho}_{AIC}}(\mathbf{I}_n-\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}})\}\right] \\ &+ \sigma^2_{\varepsilon}\,\mathrm{tr}\{\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}}(\mathbf{I}_n-\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}})\} \\ &\cdot \quad \frac{\mathrm{tr}\{\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}}(2\mathbf{R}-\mathbf{R}\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}}-\mathbf{I}_n)\}-\bar{\lambda}^{\varrho}_{AIC}/\lambda\,\mathrm{tr}\{\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}}(\mathbf{I}_n-\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}})\}}{n-\mathrm{tr}(\mathbf{S}_{\bar{\lambda}^{\varrho}_{AIC}})}. \quad (13) \end{aligned}$$

The idea of our proof is now to expand λ^{ϱ}_{REML} and $\bar{\lambda}^{\varrho}_{AIC}$ around $\mathbf{R}=\mathbf{I}_n$, that is

$$\lambda^{\varrho} = \lambda + \left.\frac{\partial\lambda^{\varrho}}{\partial\boldsymbol{\varrho}}\right|_{\boldsymbol{\varrho}=0}\boldsymbol{\varrho} + \frac{1}{2}\boldsymbol{\varrho}^T\left.\frac{\partial^2\lambda^{\varrho}}{\partial\boldsymbol{\varrho}\partial\boldsymbol{\varrho}^T}\right|_{\boldsymbol{\varrho}=0}\boldsymbol{\varrho} + \ldots$$

Using the derivative rule for implicit functions we easily find

$$\lambda^{\varrho}_{REML} = \lambda\left[1-\frac{\mathrm{tr}\{\frac{\partial\mathbf{R}}{\partial\boldsymbol{\varrho}}|_{\boldsymbol{\varrho}=0}\mathbf{S}_{\lambda}(\mathbf{I}_n-\mathbf{S}_{\lambda})\}\boldsymbol{\varrho}}{\mathrm{tr}(\mathbf{S}^2_{\lambda})-p}\right] + O(\boldsymbol{\varrho}^T\boldsymbol{\varrho}+n^{-1}), \quad (14)$$

$$\bar{\lambda}^{\varrho}_{AIC} = \lambda\left[1-\frac{\mathrm{tr}\{\frac{\partial\mathbf{R}}{\partial\boldsymbol{\varrho}}|_{\boldsymbol{\varrho}=0}\mathbf{S}_{\lambda}(\mathbf{I}_n-\mathbf{S}_{\lambda})^2\}\boldsymbol{\varrho}}{\mathrm{tr}\{\mathbf{S}^2_{\lambda}(\mathbf{I}_n-\mathbf{S}_{\lambda})\}}\right] + O(\boldsymbol{\varrho}^T\boldsymbol{\varrho}+n^{-1}). \quad (15)$$

Note that the last components in (12) and (13) are of order $O(n^{-1})$ and so are their derivatives. Based on the assumed parametrization of $\mathbf{R}$ its derivative becomes $\partial\mathbf{R}/\partial\varrho_i = \mathbf{A}_i + \mathbf{A}_i^T$ and we immediately obtain equation (7). We stop our Taylor expansion at the first order already, since we are interested in the moderate correlation misspecification that is for small $\boldsymbol{\varrho}$, so that $O(\boldsymbol{\varrho}^T\boldsymbol{\varrho})$ is of negligible order.

Relation of λ^0_{REML} and λ^0_{AIC}

Setting in (13) $\mathbf{R} = \mathbf{I}_n$ one gets that $\mathrm{E}_{Y,u}\{s_{AIC}(\bar{\lambda}^0_{AIC})|\mathbf{R} = \mathbf{I}_n\} = 0$ only if $\bar{\lambda}^0_{AIC} = \lambda = \lambda^0_{REML}$.

Proof of Corollary

Note that using a Demmler-Reinsch decomposition (see e.g. Green & Silverman, 1994 or Ruppert, Wand & Carroll, 2003) we can rewrite (14) and (15) in terms of eigenvalues of the smoothing matrix. Namely, writing $\mathbf{C}^T\mathbf{C} = \mathbf{B}^T\mathbf{B}$, where $\mathbf{B}$ is a square and invertible matrix obtained by a Cholesky decomposition and applying a singular value decomposition $\mathbf{B}^{-T}\mathbf{D}\mathbf{B}^{-1} = \mathbf{U}\mathrm{diag}(e_j)\mathbf{U}^T$, with $\mathbf{U}$ as a matrix of eigenvectors and e_j as corresponding eigenvalues, allows to represent the smoothing matrix as $\mathbf{S}_\lambda = \mathbf{L}\mathrm{diag}(b_j)\mathbf{L}^T$, with $b_j = 1/(1+\lambda e_j)$ and $\mathbf{L} = \mathbf{C}\mathbf{B}^{-1}\mathbf{U}$, so that $\mathbf{L}^T\mathbf{L} = \mathbf{I}_{K+p}$. In this notations we get from (14) and (15) together with assumption (A2$'$)

$$\lambda^{\varrho}_{REML} = \lambda + \underbrace{(-2)\lambda\frac{\sum_{l=1}^{K+p} a_l^1 b_l(1-b_l)}{\sum_{l=1}^{K+p} b_l^2 - p}}_{\Delta_{REML}}\varrho_1 + O(\delta^2 + n^{-1}), \tag{16}$$

$$\bar{\lambda}^{\varrho}_{AIC} = \lambda + \underbrace{(-2)\lambda\frac{\sum_{j=1}^{K+p} a_j^1 b_j(1-b_j)^2}{\sum_{j=1}^{K+p} b_j^2(1-b_j)}}_{\Delta_{AIC}}\varrho_1 + O(\delta^2 + n^{-1}), \tag{17}$$

where p is the dimension of $\boldsymbol{\beta}$ and $a_j^1 = (\mathbf{L}^T\mathbf{A}_1\mathbf{L})_{jj}$. The terms Δ_{REML} and Δ_{AIC} give the changes in the smoothing parameter if correlation increases, that is the

slope at $\boldsymbol{\varrho} = 0$, if we consider λ^{ϱ}_{REML} and $\bar{\lambda}^{\varrho}_{AIC}$ as functions of $\boldsymbol{\varrho}$. We can now find

$$
\begin{aligned}
& |\Delta_{AIC}| - |\Delta_{REML}| \\
= \; & 2\lambda \left\{ \frac{\sum_{j=1}^{K+p} a_j^1 b_j (1-b_j)^2}{\sum_{j=1}^{K+p} b_j^2 (1-b_j)} - \frac{\sum_{l=1}^{K+p} a_l^1 b_l (1-b_l)}{\sum_{l=1}^{K+p} b_l^2 - p} \right\} \varrho_1 \qquad (18) \\
= \; & 2\lambda \frac{\sum_{j=1}^{K} \sum_{j<l\leq K} b_j b_l (b_l - b_j) \{a_j^1 b_l (1-b_j) - a_l^1 b_j (1-b_l)\}}{\sum_{j=1}^{K} \sum_{l=1}^{K} b_j^2 b_l^2 (1-b_j)} \varrho_1. \qquad (19)
\end{aligned}
$$

To get from (18) to (19) we used the fact that due to the structure of $\mathbf{D}$ the last p eigenvalues $e_{K+1} = ... = e_{K+p} = 0$, implying $b_{K+1} = ... = b_{K+p} = 1$. We now represent $a_j^1 = \sum_{i=1}^{n-1} L_{ij} L_{(i+1)j}$ for all $j = 1, ..., K$. Based on the orthogonality of matrix L we get that $\sum_{i=1}^{n} L_{ij}^2 = 1$, implying $L_{ij}^2 = O(n^{-1})$ and $L_{ij} = O(n^{-1/2})$. Note that elements L_{ij} can be negative or positive. However, it is known that due to the oscillation property of matrix L (see Demmler & Reinsch, 1975 and Gantmacher, 1960) in each column j of matrix L there are exactly $(K + p - j)$ sign changes, providing that $L_{ij} L_{(i+1)j}$ has negative sign only for $(K + p - j)$ indices i. Moreover, $\sum_{i=1}^{n} L_{ij} = 0$. Thus, $a_j^1 = \sum_{i=1}^{n-1} L_{ij} \{L_{ij} + O(n^{-1/2})\} - (K+p-j) O(n^{-1}) = 1 - (K + p - j) O(n^{-1})$. Since we assumed K to be fixed and far less compared to n we obtain $a_j^1 = 1 - O(n^{-1})$ for $j = 1, ..., K$, which together with (19) implies (8).

A.2 Computational Issues

To demonstrate the simplicity and numerical feasibility of the REML estimate we present the implementation in R (www.r-project.org, R Development Core Team, 2005) using the example of the German import prices data. We take advantage of the `lme` function of package *nlme* (see also Pinheiro & Bates, 2002 or Ngo & Wand, 2004 for more details in smoothing using `lme(.)`). The dataset has 184 observations for `time` and `price`. First we define our model matrices for `k=90` knots. We use truncated lines here.

```
> step <- 183/(k+1)
> knots <- seq(min(time)+step,max(time)-step,by=step)
> Z <- outer(time,knots,"-")
> Z <- Z*(Z>0)
```

With this spline matrix we can call the `lme` function, using

```
> library(nlme)
> all <- rep(1,184)
> price.fit <- fitted(lme(price~time,random=list(all=pdIdent(~Z-1)),
              correlation=corAR1()))
> plot(time,price)
> lines(time,price.fit)
```

Here we allow for an AR(1) process in the residuals. The plot of the partial autocorrelation function in Figure 7 suggests refitting the model with an AR(2) correlation structure as follows

```
> price.fit1 <- fitted(lme(price~time,random=list(all=pdIdent(~Z-1)),
               correlation=corARMA(p=2)))
```

The confidence bands can be obtained according to e.g. Ruppert, Wand & Carroll (2003) from $\mathrm{Var}\{\hat{\mu}_\lambda(\mathbf{x})\} = \hat{\sigma}_\varepsilon^2 \mathrm{diag}\{\mathbf{C}(\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C} + \lambda\tilde{\mathbf{D}})^{-1}\mathbf{C}^T\}$, with $\mathbf{R}$ as estimated correlation matrix.

For an efficient algorithm for the penalized smoothing using MSE based criteria we refer to Ruppert, Wand & Carroll (2003), Appendix A.2.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**, 749–759.

Beran, J. and Feng, Y. (2001). Local polynomial estimation with FARIMA-GARCH error process. *Bernoulli* **7**, 733–750.

de Boor, C. (1978). *A practical guide to splines*. New York: Springer.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association.* **88**, 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Cardot, H. (2002). Local roughness penalties for regression splines. *Computational Statistics* **17**, 89–102.

Currie, I. and Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling* **2**, 333–349.

Demko, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM J. Numer. Anal* **14(4)**, 616–619.

Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.* **24**, 375–382.

Durban, M. and Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics* **18**, 251–262.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Science* **11(2)**, 89–121.

Eubank, R. (1999). *Nonparametric regression and spline smoothing*. New York: Dekker.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Fan, J. and Yao, Q. (2003). *Nonlinear time series*. New York: Springer.

Gantmacher, F. R. (1960). *The theory of matrices*. New York: Chelsea.

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and generalized linear models*. London: Chapman & Hall.

Hall, P. and Opsomer, J. (2005). Theory for penalized regression. *Biometrika* **92**, 105–118.

Hart, J. D. (1991). Kernel regression estimation with time series error. *Journal of the Royal Statistical Society, Series B* **53**, 173–187.

Hart, J. D. and Lee, C.-L. (2005). Robustness of one-sided cross-validation to autocorrelation. *Journal of Multivariate Analysis* **92**, 77–96.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association.* **72**, 320–338.

Hodrick, R. and Prescott, E. (1997). Postwar U.S. business cycles: An empirical approach. *Journal of Money, Credit ad Banking* **29**[**1**], 1–16.

Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planing and Inference* **127**, 53–69.

Krivobokova, T., Kauermann, G., and Archontakis, T. (2006). Estimating the term structure of interest rates using penalized splines. *Statistical pa-*

pers **47(3)**, 443–459.

Ngo, L. and Wand, M. (2004). Smoothing with mixed model software. *Journal of statistical software* **9(1)**.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science* **1**, 502–518.

Pinheiro, J. and Bates, D. (2002). *Mixed-Effects Models in S and Splus*. New York: Springer Verlag.

R Development Core Team (2005). *R: A Language and Enviroment for Statistical Computing*. R foundation for statistical computing, Wien.

Ray, B. and Tsay, R. (1997). Bandwidth selection for kernel regression with long-range dependent errors. *Biometrika* **84(4)**, 791–802.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

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

Searle, S., Casella, G., , and McCulloch, C. (1992). *Variance Components*. New York: Wiley.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B* **57**, 749–760.

Simonoff, J. and Tsay, C. (1999). Semiparametric and additive model selection using an improved Akaike criterion. *Journal of Computational and Graphical*

Statistics **8**, 22–40.

Smith, M., Wong, C., and Kohn, R. (1998). Additive nonparametric regression with autocorrelated errors. *Journal of the Royal Statistical Society, Series B* **60**, 311–331.

Wahba, G. (1990). *Spline Models for observational data*. Philadelphia: SIAM.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association.* **93**, 34–348.

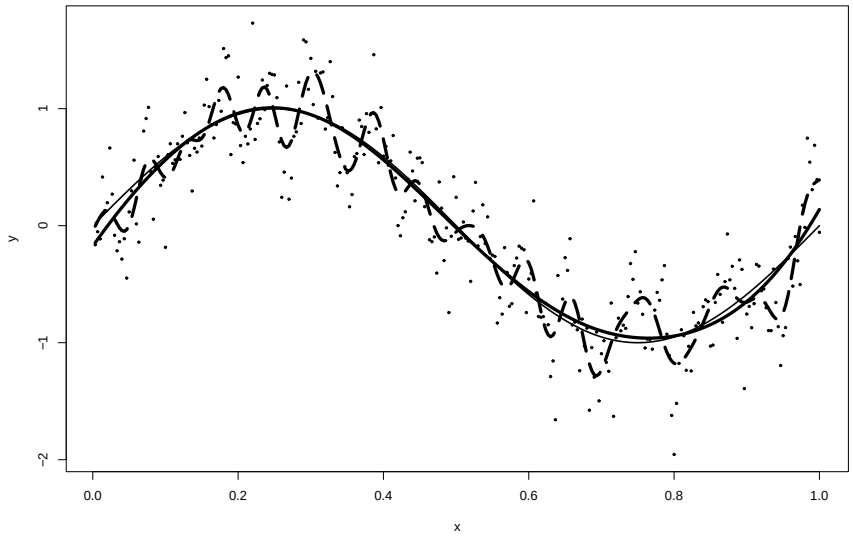

Figure 1: Estimated curves with REML (bold) and AIC (dashed) based smoothing parameter choice obtained under the assumption of independent residuals. The errors are simulated from an AR(1) process with first-order correlation 0.4.

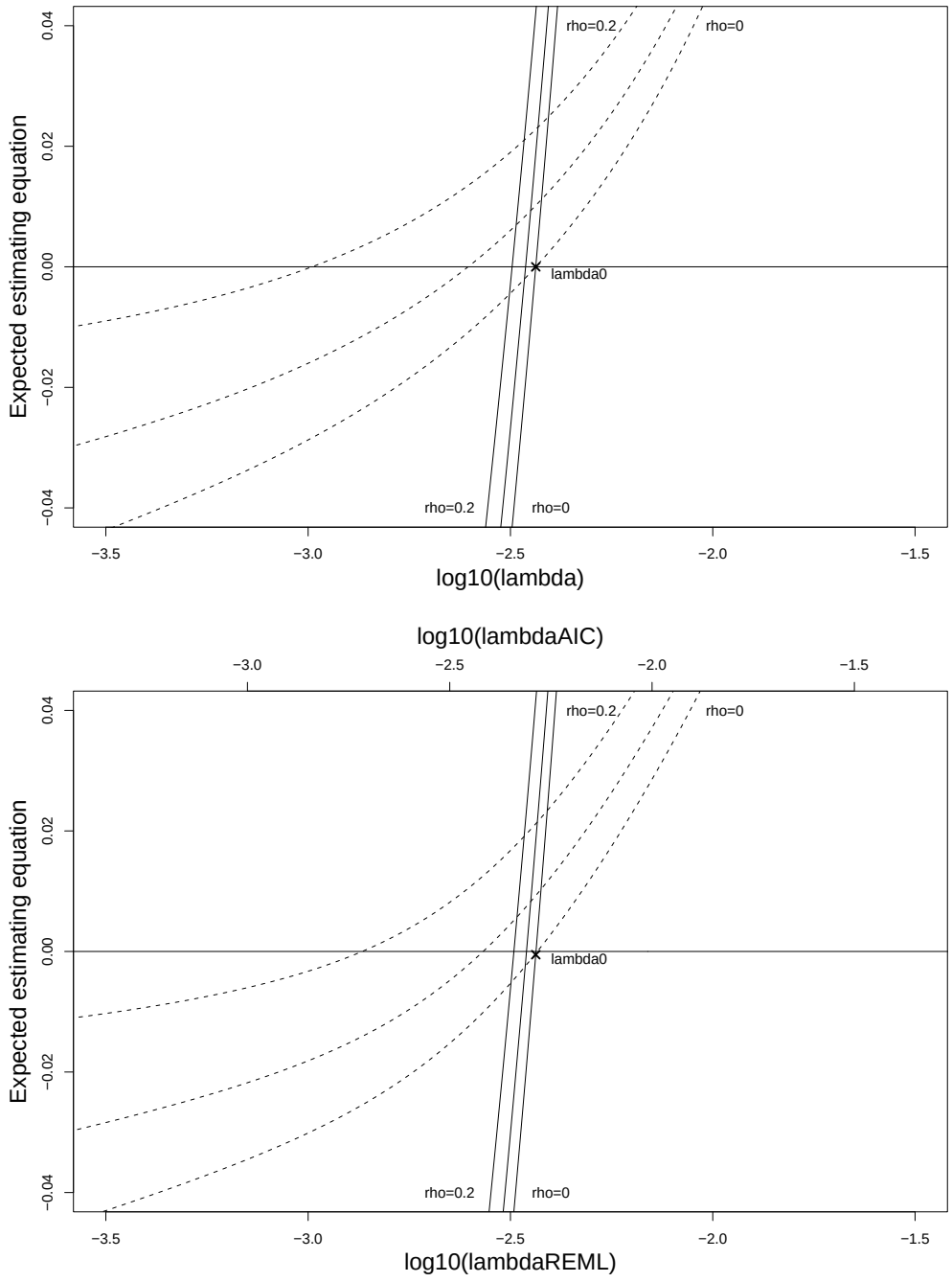

Figure 2: REML (solid) and AIC (dashed) expected estimating equations under the appropriate and misspecified models.

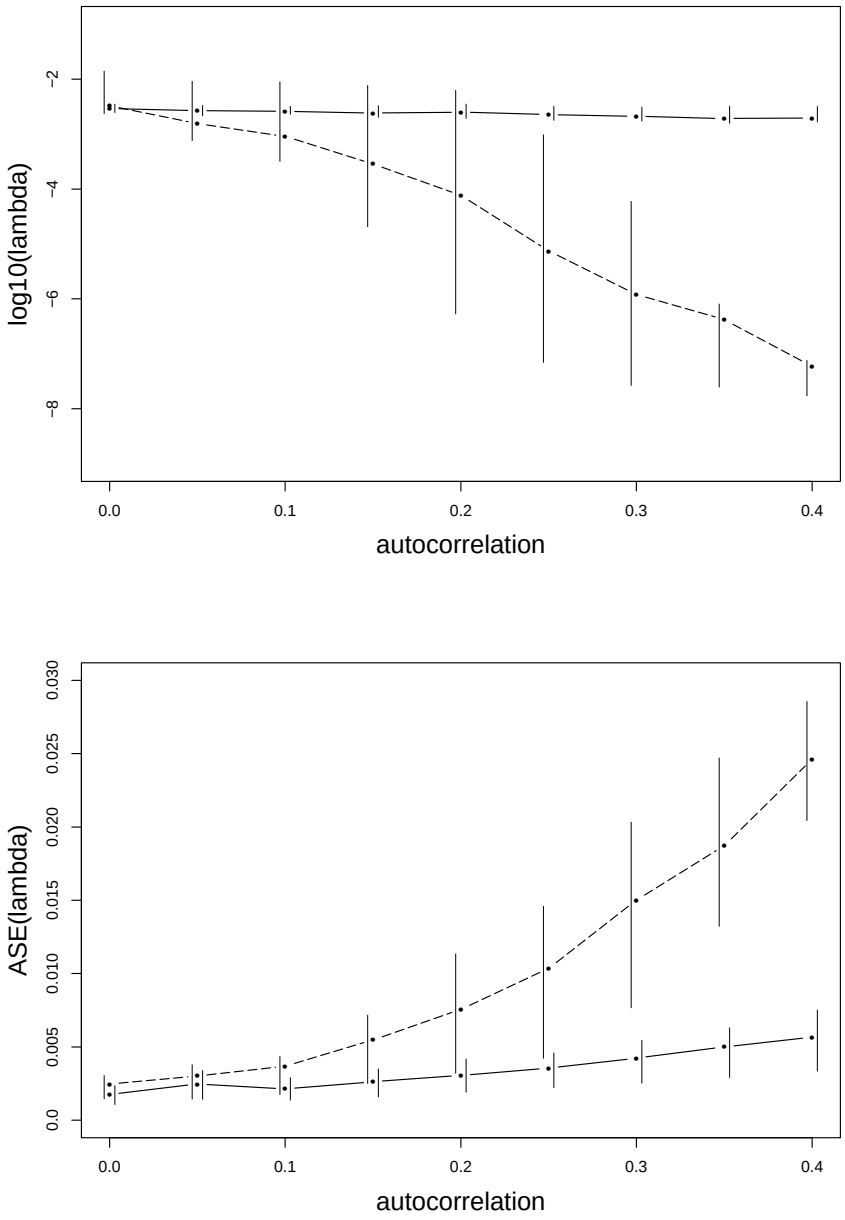

Figure 3: Average log-transformed smoothing parameters $\log_{10}(\hat{\lambda}_{REML})$ (solid) and $\log_{10}(\hat{\lambda}_{AIC})$ (dashed) in 100 simulations (upper plot) and corresponding average squared error of the fit (lower plot). Simulations are drawn from an AR(1) process with different autocorrelations ranging from 0 (independence) to 0.4. The estimates are obtained under the assumption of independent residuals. The vertical lines correspond to the interquartile range.

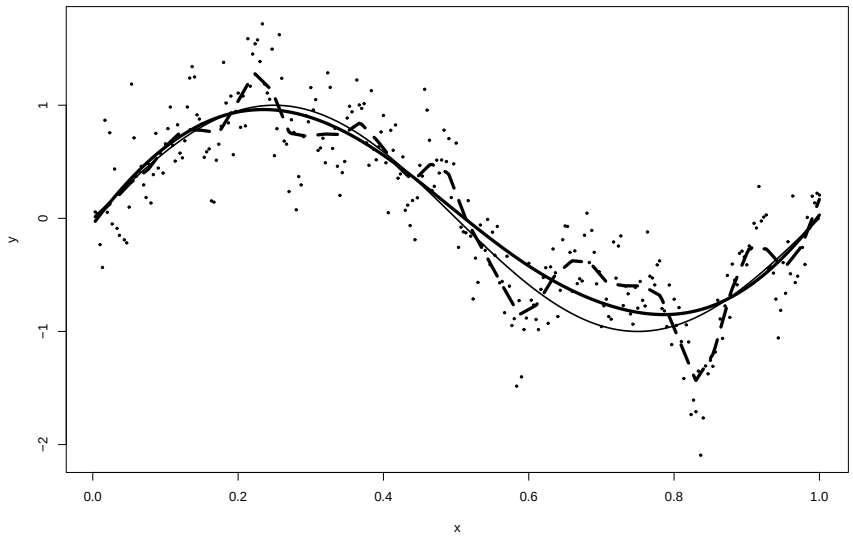

Figure 4: Estimated curves with REML (bold) and AIC (dashed) based smoothing parameter choice obtained under the assumption of an AR(1) correlation structure of the residuals. The errors are simulated from an AR(2) process with first and second order correlation 0.4 and 0.3, respectively.

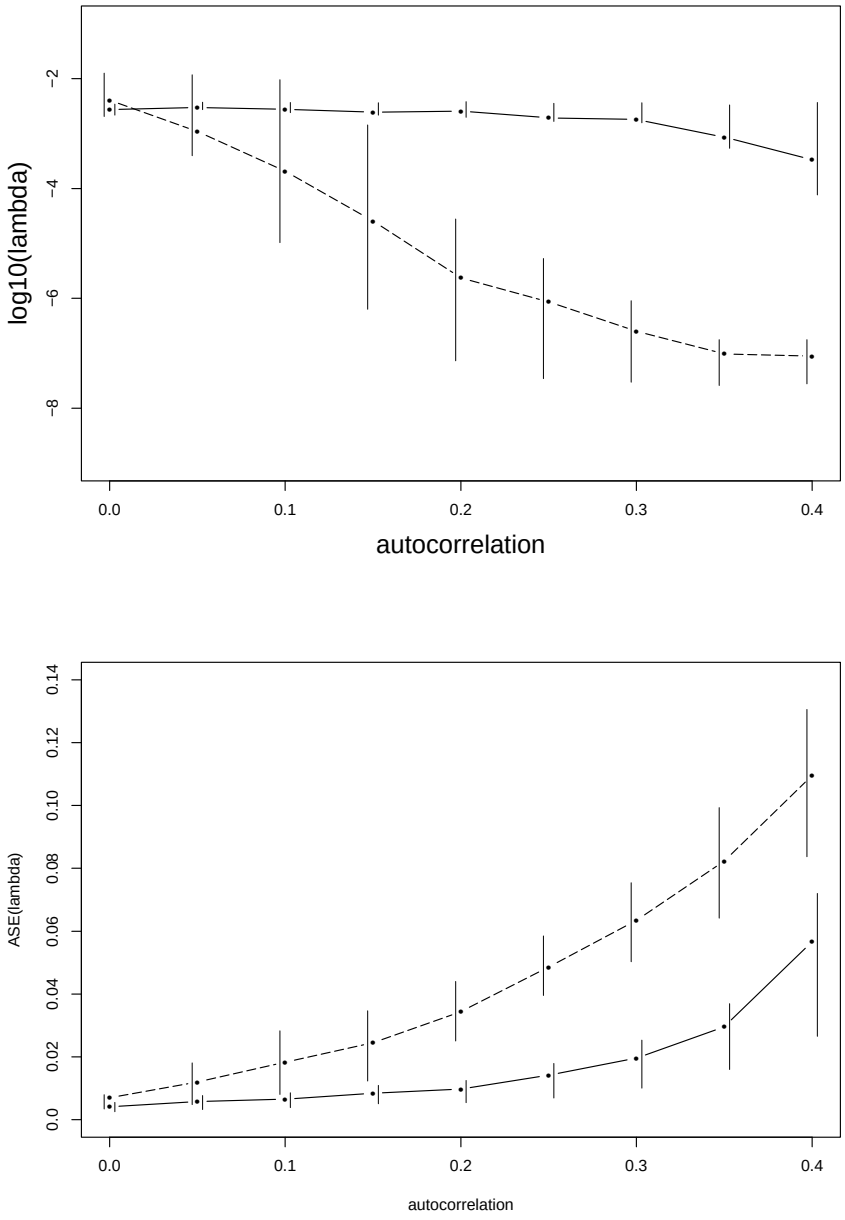

Figure 5: Average log-transformed smoothing parameters $\log_{10}(\hat{\lambda}_{REML})$ (solid) and $\log_{10}(\hat{\lambda}_{AIC})$ (dashed) in 100 simulations (upper plot) and corresponding average squared error of the fit (lower plot). Simulations are drawn from an AR(2) process with first order autocorrelation equal 0.4 and different second order autocorrelations ranging from 0 (independence) to 0.4. The estimates are obtained under the assumption of AR(1) correlated residuals. The vertical lines correspond to the interquartile range.

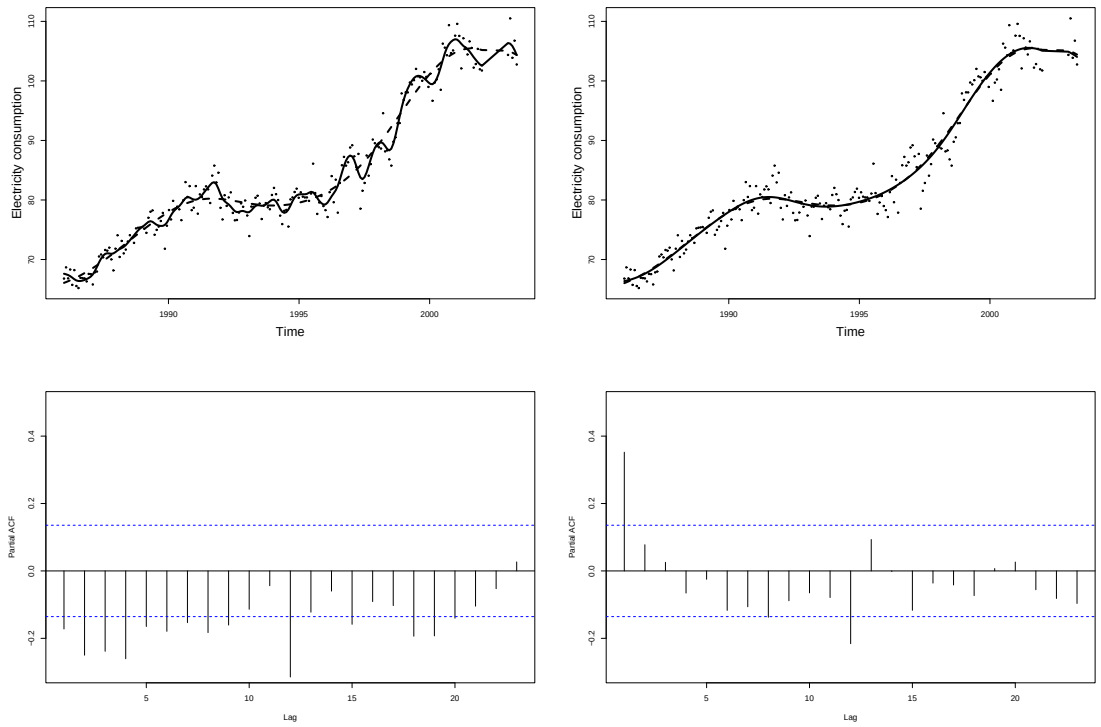

Figure 6: Top: Estimated curves with AIC (left) and REML (right) based smoothing parameter choice. Dashed lines show fits with an AR(1) correlation structure taken into account, while solid lines are fits with independent residuals assumed. Bottom: Partial autocorrelation function corresponding to the AIC (left) and REML (right) estimates obtained under the assumption of independent residuals.

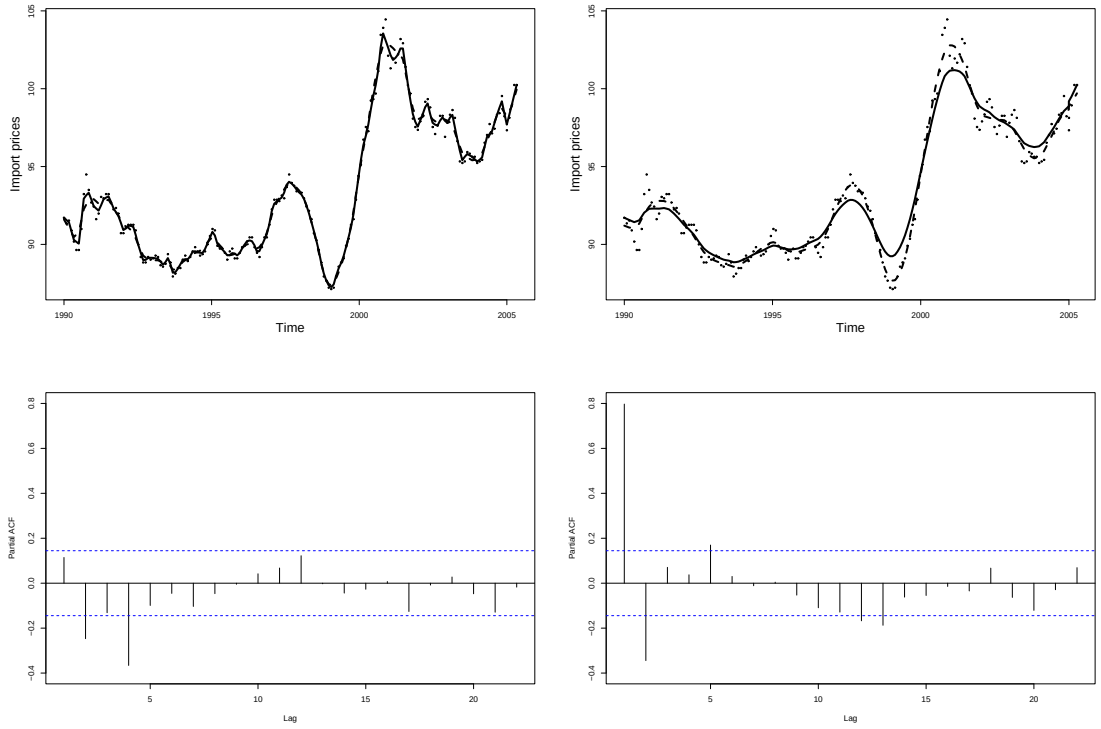

Figure 7: Top: Estimated curves with AIC (left) and REML (right) based smoothing parameter choice. Dashed lines show fits with an AR(2) correlation structure taken into account, while solid lines are fits with an AR(1) structure. Bottom: Partial autocorrelation function corresponding to the AIC (left) and REML (right) estimates with an AR(1) correlation structure taken into account.