Data-driven Selection of the Spline Dimension in Penalized Spline Regression

Göran Kauermann
Bielefeld University

J. D. Opsomer
Colorado State University

19th August 2009

Abstract

A number of criteria exist to select the penalty in penalized spline regression, but the selection of the number of spline basis functions has received much less attention in the literature. We propose to use a maximum likelihood-based criterion to select the number of basis functions in penalized spline regression. The criterion is easy to apply and we describe its theoretical and practical properties. The criterion is also extended to the generalized regression case.

Key words: nonparametric regression, mixed model, maximum likelihood, knot selection.

1 Introduction

Penalized spline smoothing has become a popular smoothing technique during the last decade. Originally introduced by O'Sullivan (1986), the practical aspects of the method have been demonstrated convincingly in Eilers and Marx (1996) who introduced the term "P-spline" smoothing. The idea was picked up and extended by Ruppert, Wand, and Carroll (2003), who took advantage of the link between

spline smoothing and mixed models, see also Wand (2003). An overview of recent developments in this field is provided by Ruppert, Wand, and Carroll (2009). In penalized spline regression, an unknown smooth function is estimated by least squares using a high dimensional spline basis and a penalty is imposed on the spline coefficients to achieve a smooth fit. The penalty can either be treated as a fixed smoothing parameter that balances the bias and variance properties of the estimator, or it can be obtained more naturally by specifying a prior distribution on the spline coefficients. An advantage of the latter approach is that the penalty is now the ratio of the error variance and the "a priori" variance of the spline coefficients, which can be estimated following Maximum Likelihood theory. Practically, available software for fitting Mixed Models can be used for simultaneously selecting the amount of smoothing and obtaining the model fit, see Ngo and Wand (2004) or Wood (2006). There are further benefits resulting from the connection with mixed models, such as robustness towards correlated residuals (see Krivobokova and Kauermann, 2007) or generalization to more complex models (see Opsomer, Claeskens, Ranalli, Kauermann, and Breidt, 2008). In the current paper, we will demonstrate that the connection with mixed models can also be used to determine the dimension of the spline basis.

The usual approach to penalized spline regression is to fix the dimension of the spline basis in advance and let the penalty adjust for the amount of smoothness of the curve, under either of the two paradigms described above. This means that the set of knots denoted by, say, $\tau_1 < \tau_2 < \ldots \tau_K$, is specified in advance and kept fixed. The knots are commonly placed to achieve a suitable spread over the design space of the covariates, for instance by placing them at empirical quantiles or equally spaced over the covariate design region. From a theoretical viewpoint, Li and Ruppert (2008) were the first to derive the asymptotic order at which K should increase as the sample size n increases. Claeskens, Krivobokova, and Opsomer (2009) showed that there are two asymptotic scenarios for K, either leading to classical smoothing spline or regression

spline asymptotics, respectively. In Kauermann, Krivobokova, and Fahrmeir (2009) the asymptotic behavior of K and n was tackled by making use of the link between penalized spline smoothing and mixed models. While these articles give asymptotic rates for K based on sample size n, the results do not address the issue of how to select K for a given fixed sample size n.

In practical terms, the choice of the dimension K has an effect of the properties of the nonparametric function estimator. Too few knots can result in biased estimation, because features of the mean function are missed when the spline basis is not sufficiently large. Conversely, too many knots increase the risk of overfitting. This is at least partly offset by the selection of an appropriate penalty, but a more parsimonious spline basis will generally result in a more stable fit even after allowing for the effect of the penalization. In an extensive simulation study, Ruppert (2002) demonstrated the severe bias caused by values of K that are too small, and the fact that once K is sufficiently large, the effects of further increasing K are relatively modest. Specifically, he noted that increasing K beyond what is required to capture the features of the underlying mean function caused a modest increase in mean squared error as well as an increase in computational complexity. Ruppert (2002) investigated a GCV-based criterion to select K, and also suggested a heuristic rule of thumb $K = \min(40, n/4)$, which works well in practice for many datasets. The GCV criterion only applies under the framework of treating the penalty as a tuning constant. In the current article, we consider the mixed model specification of the penalized spline regression and now treat K as a "parameter" in the model likelihood and we choose K such that the likelihood function is maximized. This provides a natural way to integrate the selection of K within the mixed model framework, for which no method is currently available. Interestingly enough, the results will show that the maximum of the likelihood function considered as a function of K increases steeply for small values of K and then reaches a plateau at a value K_0, say, after which it flattens out. Figure 1, graph (b) gives a typical shape. This matches the above discussion of the effect

of K and suggests that using the "inflection point" K_0 is a reasonable dimension for the spline basis, as it follows the principle of parsimony. We will provide some theoretical justification for this proposal, by considering a scenario in which the sample size n is large but fixed and K is increasing.

The remainder of the paper is structured as follows. Section 2 briefly reviews the mixed model specification of the penalized spline regression, describes the likelihood-based method for selecting the spline dimension and provides a theoretical justification. Section 3 gives simulation results and an example. In Section 4, we extend the results to the generalized regression case.

2 Spline Dimension and Mixed Models

We consider the standard nonparametric model with independent observations

$$Y_i = \mu(x_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{1}$$

where $\mu(.)$ is an unknown smooth function, the ϵ_i are independent and identically distributed errors, i.e. $\epsilon_i \sim (0, \sigma_\epsilon^2)$, and the x_i take values in $[0, 1]$, for simplicity. We estimate $\mu(.)$ using penalized splines. To do so, we replace $\mu(x)$ in (1) by some high dimensional spline of the form $\mu(x) = X(x)\beta + Z(x)u$. Here, $X(.)$ is a low dimensional basis while $Z(.)$ is high dimensional. Using truncated polynomials, we set $X(x) = (1, x, x^2/2!, \ldots, x^q/q!)$ and $Z(x) = \left((x-\tau_1)_+^q/q!, \ldots, (x-\tau_K)_+^q/q!\right)$ where $(t)_+^q = t^q$ for $t > 0$ and 0 otherwise, and the knots $\tau_1 < \tau_2 < \ldots < \tau_K$ cover the range of x values. Alternatively, one can transform the truncated polynomials to q-th order B-splines (de Boor, 1978), which exhibit a numerically more stable behavior. Although the two bases are equivalent in theoretical terms, the truncated polynomials allow for an easier theoretical insight, which is why this representation is preferred here. If K is large, an unconstrained Least Squares estimate $X(x)\hat{\beta} + Z(x)\hat{u}$ is likely to be unacceptably variable due to overfitting. This can be controlled by imposing a penalty on u of the form $\lambda u^T D^* u$, where

D^* is an appropriately chosen penalty matrix and λ is the penalty or smoothing parameter. Using truncated polynomials, the choice of D^* as identity matrix has shown to work well in practice and will be used here for simplicity.

Under the mixed model framework for fitting penalized splines, we assume the model

$$Y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n), \qquad u \sim N(0, \sigma_u^2 I_{K-1}), \tag{2}$$

where $Y = (Y_1, \ldots, Y_n)$, X and Z are matrices with i-th row $X(x_i)$ and $Z(x_i)$, respectively, and I_n and I_K are n and K dimensional identity matrices, respectively. Under model (2), the loglikelihood function can be written as

$$l_K(\beta, \sigma_\epsilon^2, \lambda) = -\frac{n}{2}\log(\sigma_\epsilon^2) - \frac{1}{2}\log|V_\lambda| - \frac{1}{2\sigma_\epsilon^2}(Y - X\beta)^T V_\lambda^{-1}(Y - X\beta) \tag{3}$$

with $V_\lambda = I_n + ZZ^T/\lambda$ where $\lambda = \sigma_\epsilon^2/\sigma_u^2$. The likelihood depends on the parameters $\beta, \sigma_\epsilon^2, \lambda$ as well as on spline dimension K, written as subscript. In this mixed model representation of the penalized spline, the log-likelihood is maximized with respect to the parameters, but K is kept fixed. Our proposal in this paper is to maximize $l_K(\beta, \sigma_\epsilon^2, \lambda)$ with respect to both the parameters and K.

Considering first maximization of (3) with respect to the parameters, we obtain $\hat{\theta} = (\hat{\beta}^T, \hat{u}^T)^T = (P^T P + \lambda D)^{-1} P^T Y$, with $P = (X, Z)$ and D a diagonal matrix with $(0_{(q+1)}, 1_K)$ on the diagonal, where $0_{(q+1)}$ is the zero vector of dimension $q + 1$ and corresponding definition for 1_K. Moreover, it is not difficult to show that

$$\hat{\sigma}_\epsilon = \frac{(Y - X\hat{\beta})^T V_\lambda^{-1}(Y - X\hat{\beta})}{n} \tag{4}$$

and

$$\hat{\sigma}_u^2 = \frac{\hat{\sigma}_\epsilon^2}{\hat{\lambda}} = \frac{\hat{u}^T\hat{u}}{tr(S_\lambda)} \tag{5}$$

are the Maximum Likelihood estimates with $S_\lambda = P(P^T P + \lambda D)^{-1} P^T$ (see also Searle, Casella, and McCulloch, 1992). As an alternative to (4) and (5), we could consider Restricted Maximum Likelihood (REML, see Harville, 1977) estimators,

which correct the bias due to the estimation of β. This bias is potentially significant if the number of parameters is large compared to the number of random components. In our scenario the opposite happens, i.e. the dimension of β is small compared to u. Therefore, and also for simplicity and ease of notation, we do not pursue REML estimation but concentrate on the Maximum Likelihood estimates.

While it might in principle be possible to maximize l_K directly and simultaneously with respect to all the parameters and K, the discrete nature of K and the fact that the structure of V_λ changes with K makes this impractical. Instead, we therefore consider the following "pre-maximized" likelihood

$$l_K(\hat{\beta}, \hat{\sigma}_\epsilon^2, \hat{\lambda}) = -\frac{n}{2}\log(\hat{\sigma}_\epsilon^2) - \frac{1}{2}\log|V_{\hat{\lambda}}|, \tag{6}$$

where we plugged in the estimators for $\sigma_\epsilon^2, \lambda$ and β and removed terms not depending on K. This quantity is only a function of K, and it can now be maximized with respect to K (we note that for each evaluation at a value for K, the quantities $\hat{\sigma}_\epsilon^2, \hat{\lambda}$ and $\hat{\beta}$ need to be recomputed). Figure 1 shows an example. Plot (a) gives $n = 300$ data points drawn from a smooth function, shown as solid line. The dashed line is the best fit using truncated linear polynomials when K is chosen to maximize the likelihood (6). The latter is plotted against K in plot (b) in Figure 1. We see that the likelihood increases for small K but once it reaches its weakly exposed maximum at K_0, it completely flattens out. Using the maximum likelihood principle, we therefore propose to select the number of knots K as K_0, the maximum of the curve $l_K(\hat{\beta}, \hat{\sigma}_\epsilon^2, \hat{\lambda})$ against K.

We now provide some theoretical motivations for the behavior shown in these plots and for using K_0 as the "optimal" number of knots. In particular, the behavior seen does not seem to depend on the functional form being used. We derive our asymptotic results by letting K increase under the following assumptions.

- **Assumption 1** *We assume that* $P(x) = \big(X(x), Z(x)\big)$ *is a truncated polynomial basis of order* q *and that* $\mu(x)$ *is* $q+1$ *times differentiable except at*

a finite number of isolated points (which may be an empty set).

- **Assumption 2** *We use model (2) to fit a smooth function $\mu(x)$ which is approximated by a spline of the form $P(x)\theta = X(x)\beta + Z(x)u$, with K as the number of knots in $Z(x)$. We assume that covariate x is uniformly distributed on $[0,1]$ and knots $0 < \tau_1 < \ldots < \tau_K = 1$ are placed such that*

$$\tau_j - \tau_{j-1} = O(K^{-1}) \quad , j = 2, 3, \ldots, K \tag{7}$$

- **Assumption 3** *We assume that $\sigma_u^2 = cK^{-1}$ for some constant c.*

Assumption 3 can be motivated as follows. With $Z(x)$ as truncated polynomial basis of order q, we get for $h > 0$

$$P(x+h)\theta - P(x)\theta = O(h) + 1_{\{q>0\}} \sum_{j \leq j(x)} \left\{ h \frac{(x-\tau_j)}{(q-1)!} \right\} u_j \{1 + O(h)\} + \sum_{j(x) < j \leq j(x+h)} \frac{(x+h-\tau_j)^q}{q!} u_j \tag{8}$$

where $j(x) = \arg\max\{j : \tau_j \leq x\}$. Taking the variance of the above difference, using the prior distribution for u, gives

$$Var[\{P(x+h) - P(x)\}\theta] = 1_{\{q>0\}} O(Kh^2\sigma_u^2) + O(Kh^{2q+1}\sigma_u^2)$$

with $1_{\{q>0\}}$ as indicator function. Hence, letting K increase, differentiability of order $q+1$ of the random function $P(x)\theta$ is ensured by Assumption 3. This assumption also corresponds to the observed behavior of $\hat{\sigma}_u^2$ and K, so that it is required in order for $\hat{\sigma}_u^2$ to be a valid estimator of σ_u^2. Considering model (2), we get for the corresponding Maximum Likelihood estimate $\hat{\sigma}_u^2 = O_p(1)K^{-1}$. In Figure 1 in plot (c), we show the behavior of $1/\hat{\sigma}_u^2$, which clearly mirrors a linear behavior once K is large enough.

We can now investigate the behavior of (6) for K increasing (and n being kept fixed). We start by looking at $\hat{\sigma}_\epsilon^2$ and first argue that $\hat{\sigma}_\epsilon$ has a stable behavior if K increases. Note that unlike σ_u^2, the variance σ_ε^2 in (2) concerns the residuals

which are not affected by K. Hence, if the mean function is estimated consistently we also obtain consistent estimation of σ_ε^2 for sufficiently large n. However, the residual variance σ_ϵ^2 is estimated with (positive) bias if the mean function is estimated with bias. Hence, the misbehavior of $\hat{\sigma}_\varepsilon^2$ occurs when K is small. This behavior can be seen in Figure 1 plot (d). Let $\hat{\epsilon} = Y - X\hat{\beta} - Z\hat{u}$ and let $V_{\hat{\lambda}}$ be the variance matrix with estimated $\hat{\lambda} = \hat{\sigma}_\epsilon^2/\hat{\sigma}_u^2$. Note that $V_{\hat{\lambda}}^{-1}(Y - X\hat{\beta}) = \hat{\epsilon}$ and $Z^T\hat{\epsilon} = \lambda\hat{u}$. Then with (4) and simple calculus we find

$$\hat{\sigma}_\epsilon^2 = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n - tr(S_{\hat{\lambda}})}. \tag{9}$$

If K is too small, then the fitted residuals $\hat{\epsilon}$ are biased and the resulting estimate $\hat{\sigma}_\epsilon^2$ overestimates the true variance. This in return will result in low values for the likelihood-based criterion $l_K(., \hat{\sigma}_\epsilon^2, \hat{\lambda})$ so that such choice of K is not selected following our criterion. Once K is large enough, the bias vanishes and the consistency of the estimator ensures that $\hat{\sigma}_\epsilon^2$ remains stable for a wide range of values of K. Hence, we conclude that for K increasing,

$$\hat{\sigma}_\epsilon^2 = \sigma_\epsilon^2\big\{1 + O_p\big(n^{-\frac{1}{2}}\big)\big\}. \tag{10}$$

It is also interesting to note that estimator $\hat{\sigma}_\epsilon^2$ in (9) has the common structure of variance estimation in smoothing models, see e.g. Hastie and Tibshirani (1990). It remains to look at the second component in (6) and its behavior as K increases. Following the assumptions, we have $\lambda = \sigma_\varepsilon^2/\sigma_u^2 = \tilde{c}K$ for some constant $\tilde{c}$ and hence

$$\hat{\lambda} = \hat{\tilde{c}}K\big\{1 + O_p\big(n^{-\frac{1}{2}}\big)\big\} \tag{11}$$

for K sufficiently large. We can now use (11) to establish the behavior of the determinant of $V_{\hat{\lambda}} = I + ZZ^T/\hat{\lambda}$. With Assumptions 1 and 2, we have

$$Z^TZ = nR_K + \big\{1 + O_p\big(n^{-\frac{1}{2}}\big)\big\}, \tag{12}$$

where $R_K \in \mathbb{R}^{K\times K}$ has elements

$$R_{K,ij} = \int_0^{\frac{K-i}{K}} \left(t + \frac{i-j}{K}\right)^q t^q \,\mathrm{d}t. \tag{13}$$

Let $\rho_k,\ k = 1, \ldots, K$, denote the eigenvalues of R_K. We can numerically calculate the eigenvalues for matrix R_K and observe that

$$\rho_k \approx K k^{\log(a_q)} \tag{14}$$

for some constant a_q dependent on q. In Figure 2 we plot $\log(\rho_k)$ against $\log(k)$ for $q = 0, 1, 2, 3$, which clearly exhibits a linear shape supporting the formula (14). Note that it is easily derived that $\mathrm{tr}(R) = \sum_{k=1}^{K} \rho_k = O(K)$, which implies with (14) that $\sum_{k=1}^{K} k \log(a_q) = O(1)$ so that $\log(a_q) < -1$. Considering now the second component in (6) yields with (11) and (14)

$$\begin{aligned}
\log|V_\lambda| &\approx \sum_{k=1}^{K} \log\left(1 + \frac{n}{\hat{\tilde{c}}}\ k^{\log(a_q)}\right) \\
&= -\sum_{j=1}^{\infty} (-1)^j \frac{n^j}{\hat{\tilde{c}}^j} \sum_{k=1}^{K} k^{j\log(a_q)} \\
&= -\sum_{j=1}^{\infty} (-1)^j \frac{n^j}{\hat{\tilde{c}}^j}\ O\big(1 + K^{\log(a_q)+1}\big) \\
&= \log\left(1 + \frac{n}{\hat{\tilde{c}}}\right)\ O\big(1 + K^{\log(a_q)+1}\big).
\end{aligned} \tag{15}$$

Since $\log(a_q) + 1 < 0$ we see that the second term does not change in K once K is large enough. We also see that the explicit order at which the influence of K diminishes depends on q, the order of the truncated polynomials. In Figure 2 we give the value of $\log(a_q)$ in the bottom of each plot which ranges from $\log(a_0) \approx -1.9$ for truncated constants to $\log(a_3) \approx -8.2$ for truncated cubic polynomials. The values are approximate and calculated by fitting a linear line through the points. This fitted line is also shown in the four plots as solid line. The larger the polynomial degree, the less influence has K as $\log(a_q)$ decreases so that $K^{\log(a_q)+1}$ gets smaller for larger values of q. Overall, for a given sample size n, the effect of K on the determinant of $V_{\hat{\lambda}}$ vanishes. In summary, because of the behavior noted in (10) and (15) for the likelihood (6), we have

$$l_K(\hat{\beta}, \hat{\sigma}_\epsilon, \hat{\lambda}) = O_p(n) + O\big(\log(n)\big)\ O_p(1 + K^{\log(a_q)+1}).$$

We therefore suggest the following strategy for selecting K. Maximizing $l_K(\hat{\beta}, \hat{\sigma}_\epsilon^2, \hat{\lambda})$ with respect to K ensures that we avoid too low values of K for which the spline basis is too inflexible to fit the data well. Once K passes the point K_0, then further increases in K do not increase the likelihood and do not help in achieving a better fit. This automatic balancing of the spline complexity and amount of penalization is a major advantage of the mixed-model approach to spline fitting. Some additional remarks seem worthwhile:

1. Our result assumes n to be given (determined by the sample size) but K to increase. Our experience matches that of Ruppert (2002) in the fixed-penalty framework and shows that K_0 maximizing the likelihood (6) is relatively small. We do not make any statements how K_0 depends on n, i.e. if and how K_0 should increase with increasing sample size. The asymptotic relation between the optimal K, based on the Mean Squared Error and n is investigated in Li and Ruppert (2008), Kauermann, Krivobokova, and Fahrmeir (2009) and Claeskens, Krivobokova, and Opsomer (2009).

2. Looking at Figure 1, it appears that the likelihood reaches its plateau when the estimator of σ_ϵ^2 becomes stable, that is without bias. This can be also seen in the formula. Note that classical Generalized Cross Validation (GCV) (see e.g. Hastie and Tibshirani, 1990) is defined through minimizing

$$\frac{\hat{\epsilon}^T \hat{\epsilon}}{n - \mathrm{tr}(S_\lambda)} = \hat{\sigma}_\epsilon^2.$$

   Hence, using GCV for selecting the spline dimension K would correspond to maximizing the first component in (6) only. With our results above it appears that GCV is approximately equivalent to maximizing the marginal log likelihood because the determinant $|V_{\hat{\lambda}}|$ depends in an ignorable way on K. In this way, we can relate our results to Ruppert (2002).

3. Following the arguments above and using (14), we can calculate the degree

of the model explicitly as

$$\begin{aligned}\mathrm{tr}(S_{\hat{\lambda}}) &\approx \sum_{k=1}^{K}\left(\frac{n\rho_k}{n\rho_k + K\hat{\tilde{c}}}\right) \\ &= \sum_{k=1}^{K}\frac{k^{\log(a_q)}}{k^{\log(a_q)} + \hat{\tilde{c}}/n} = O\big(1 + K^{\log(a_q)+1}\big) \qquad (16)\end{aligned}$$

3 Simulations and Example

3.1 Simulations

To demonstrate the practical performance of our selection criterion, we follow the simulation settings given in Ruppert (2002). Covariates x are equidistant in $[0, 1]$. As spline basis we make use of B-splines, which are numerically more stable. These are easily implemented as shown in Wand and Ormerod (2008), i.e. the spline can be written as $X\beta + Zu$ with $u \sim N(o, \sigma_u^2 I)$ and penalty matrix. We make use of five mean functions in what follows. First, we simulate from the "Logit" function

$$\mu(x) = \frac{1}{1 + \exp\{-20(x - 0.5)\}},$$

where we set sample size $n = 100$ and simulate with error variance $\sigma_\epsilon^2 = 0.2^2$. The results are shown in Figure 3, top row. Plot (a) displays a typical realization, and plot (b) shows boxplots of value of the maximized likelihood (6) for different values of K based on 100 replicates of the simulation. The vertical line is the median of the selected K_0. Finally, in plot (c) we give a histogram of the optimally chosen values of K_0. Apparently, the selection is quite stable, preferring small values of K. As second simulation, we employ the "Bump" function

$$\mu(x) = x + 2\ \exp\big[-\{16(x - 0.5)^2\}\big].$$

The results for $\sigma_\epsilon^2 = 0.3^2$ are shown in Figure 3 (d) to (f) with the same organization as in the plots (a) to (c). As before, the value of the likelihood does not increase with K and the selection of a small value K is preferred.

The next two functions we simulate from are sine curves

$$\mu(x) = \sin(2\ \pi\ \theta\ x),$$

where $\theta = 3$ and $\theta = 6$, respectively. In these cases, we set $\sigma_\epsilon^2 = 1$ and provide the results for $n = 200$ in Figure 4, plots (a)-(c) for $\theta = 3$ and (d)-(f) for $\theta = 6$. The interpretation is the same, and for $\theta = 3$ it appears that the O'Sullivan splines with a small number of knots are performing best. For $\theta = 6$, one needs a larger number of knots as the proposed selector criterion shows.
Finally, we use a function with spatial heterogeneity, defined as

$$\mu(x) = \sqrt{x(1-x)}\ \sin\left\{\frac{2\pi(1+2^{-3/5})}{x+2^{-3/5}}\right\}$$

(see Wand, 2000). The results are shown for $n = 200$ in plots (a)-(c) in Figure 5 and for $n = 2000$ in plots (d)-(f), respectively. Again, the O'Sullivan splines do not need a large number of knots and the maximized likelihood is stable for increasing K.

3.2 Data Example

We consider data on heating demand in the district heating network in the city of Wuppertal in North-Western Germany. Heating demand here refers to both heating of houses as well as providing of hot water. Heating is induced by steam taken from the district's steam network, which is fed by two power plants in the city of Wuppertal. Steam (i.e. heating) has to be provided throughout the whole year and not only in winter months. Let y be the total heating demand on day i. The heating demand is modelled to depend on the continuous covariates day of the year ($\texttt{yday}_i$), the maximum and minimum temperature of the day ($\texttt{max.temp}_i, \texttt{min.temp}_i$) and the mean global radiation ($\texttt{mean.rad}_i$). Moreover, the heating demand depends on the day of the week ($\texttt{wday}_i$), where we group the days into the four categories Sunday, Monday, Tuesday to Friday and Saturday,

respectively. Public holidays are generally classified as Sundays and in leap years we omit the last day of the year for simplicity. We model the heating demand as

$$y_i = \texttt{wday}_i + \mu_1(\texttt{yday}_i) + \mu_2(\texttt{max.temp}_i) + \mu_3(\texttt{min.temp}_i) + \mu_4(\texttt{mean.rad}_i) + \epsilon_i \quad (17)$$

for $i = 1, \ldots, n$.

Our data base contains data from January 1st 2006 to December 31st 2008. Since the data are observed sequentially, it is plausible to assume that the residuals ϵ_i are serially correlated. We therefore assume an AR(1) structure. Note that in case of correlated residuals, smoothing parameter selection becomes unstable, see Opsomer, Wang & Yang (2001) . The use of a Mixed Model for smoothing parameter selection however exhibits some robustness with respect to misspecification of the correlation structure, as shown in Krivobokova and Kauermann (2007). Based on this result we feel confident that even if the assumed AR(1) structure is too simplistic, the selection of the smoothing parameter based on the Maximum Likelihood estimate in the Mixed Model will work properly. To do so, we replace $\mu_j(.)$ by a K_j-dimensional B-spline basis and impose an a priori normality on the spline coefficients. Hence, we need maximize the marginal likelihood from the resulting Mixed Model (17) with respect to the 4-dimensional vector $(K_1, K2, K_3, K_4)$. Instead of running a 4-dimensional optimization, we start with a small K and increase the elements of K sequentially over the 4 functions until the marginal likelihood no longer increases. We thereby increase K_j by a step of size 2, $j = 1, \ldots, 4$. If the marginal likelihood does not increase for one function, we sequentially select the next function to increase K. If the likelihood does not increase for all functions when increasing K, the algorithm terminates.

Table 1 shows the outcome of the selection routine applied to the data. The resulting optimal fit based on penalized B-splines is shown in Figure 6. We see that even a small number of splines provide a satisfactory fit. The interpretation of the functions is straightforward, showing an increased heating demand in winter

Table 1: Performance of the selection algorithm of K in the heating demand example

step	K_1	K_2	K_3	K_4	log likelihood
0	4	4	4	4	-6970.949
1	6	4	4	4	-6953.221
2	6	6	4	4	-6940.536
3	6	6	4	6	-6940.433
4	6	8	4	6	-6936.722
5	6	10	4	6	-6936.516

and for cold temperatures. The effect of mean radiation is overall weak. The reason for this is that global radiation is strongly correlated with temperature.

4 Generalized Spline Smoothing

4.1 Extention to Generalized Smoother Models

The likelihood-based approach we just described can be extended to the generalized smoother model $E(Y_i|x) = h\{\eta(x_i)\},\ i = 1, \ldots, n$ with $h(.)$ as known link function and $\eta(.)$ as unknown smooth function. Response Y_i for given x follows an exponential family distribution. For simplicity, we will assume that $h(.)$ is the natural or canonical link and the dispension parameter is set equal to 1. Replacing the smooth function $\eta(.)$ by a high dimensional basis and imposing a penalization prior on the spline coefficients extends model (2) to a Generalized Linear Mixed Model (GLMM) of the form

$$E(Y|u) = h\{X\beta + Zu\},\ u \sim N(0, \sigma_u^2 I_{K-1}). \tag{18}$$

Integrating out the random spline components yields the marginal likelihood

$$l_K(\beta, \sigma_u^2) = -\frac{K}{2}\log(\sigma_u^2) + \log \int \exp\left\{ l_{K,c}(\beta, u) - \frac{1}{2}\frac{u^T u}{\sigma_u^2} \right\} \mathrm{d}u, \tag{19}$$

where $l_{K,c}(\beta, u)$ is the (conditional) likelihood treating u as fixed coefficient as above. Unless the response is normal, the above integral is not analytical so that numerical approximation techniques need to be applied to proceed. We here pursue Laplace approximation resulting in the approximate likelihood

$$\tilde{l}_K(\beta, \sigma_u^2) = -\frac{K}{2}\log(\sigma_u^2) + l_{K,c}(\beta, \hat{u}) - \frac{1}{2}\frac{\hat{u}^T \hat{u}}{\sigma_u^2} - \frac{1}{2}\log|Z^T W Z + I/\sigma_u^2|, \tag{20}$$

where $\hat{u}$ is the maximizer of $l_{K,c}(\beta, u) - \frac{1}{2}u^T u/\sigma_u^2$ and diagonal weight matrix $W = W(\beta, u)$ contains the variance function of the underlying exponential family. To assess the accuracy of (20), one can investigate the error term $\delta_K = l_K(\beta, \sigma_u^2) - \tilde{l}_K(\beta, \sigma_u^2)$. We show in Appendix A.1 that under Assumption 3, δ_K is of ignorable size as K increases (with n fixed). This implies that we may consider $\tilde{l}_K(.)$ instead of $l_K(.)$.

Maximizing (20) yields $\hat{\sigma}_u^2 = \hat{u}^T \hat{u}/\mathrm{tr}(S_{\hat{\lambda}})$ with $S_\lambda = Z(Z^T \hat{W} Z + \lambda D)^{-1} Z^T \hat{W}$ where $\hat{W} = W(\hat{\beta}, \hat{u})$ and $\hat{\lambda} = \hat{\sigma}_u^{-2}$. Substituting the estimators back into expression (20), we obtain the Laplace approximated maximized log likelihood

$$\tilde{l}_K(\hat{\beta}, \hat{\sigma}_u^2) = l_{K,c}(\hat{\beta}, \hat{u}) - \frac{1}{2}\mathrm{tr}(S_{\hat{\lambda}}) + \frac{1}{2}\log|I - S_{\hat{\lambda}}|, \tag{21}$$

which is now solely a function in K. In Figure 7 plot (b), we show the course of $\tilde{l}_K(\hat{\beta}, \hat{\sigma}_u^2)$ as a function of K in a simulated example with binary data shown in plot (b) with true (solid) and fitted (dotted) mean function $h\{\eta(x)\}$. The behavior of the log likelihood is comparable to the normal case treated in Section 2. In particular, we see from plot (c) that the relationship between σ_u^2 and K follows Assumption 3. Plot (d) shows the behavior of the first two components in (21). Comparing plot (d) with (b) indicates that the last component in (21) does not depend on K. We explore this property in more depth in Appendix A.2 and A.3. It is shown there that $E\big(l_{K,c}(\hat{\beta}, \hat{u})\big) = E\big(l_{K,c}(\beta, u = 0)\big) + O(1)$. For $u = 0$

the conditional likelihood $l_K(\beta, u)$ does naturally not depend on K. Moreover, following arguments which lead to (16), we find that the second component in (21) is of approximate order $O(1 + K^{\log(a_q)+1})$ with $\log(a_q) < -1$. Finally, since $I - S_\lambda = \tilde{V}_\lambda$, with $\tilde{V}_\lambda = W + \sigma_u^2 Z Z^T$ as generalized variance matrix, we can apply arguments as in (15) to show that the third component in (21) is of order $O(\log(n))\ O(1 + K^{\log(a_q)+1})$.

Similarly to remark 2 in Section 2, the first two components in (21) define the classical Akaike Information Criterion (AIC) in generalized (additive) models, see e.g. Hastie and Tibshirani (1990). Seeing the behavior of the component and reflecting that the third component in (21) depends in an ignorable way on K it becomes obvious that an optimal K can also be chosen via the AIC approach leading to substantially the same optimal K_0. In other words, formula (21) mirrors the relation between the Generalized Linear Mixed Model approach and the classical Akaike Information Criterion.

4.2 Simulations and Example

To demonstrate the performance of our selection strategy of K, we run a small simulation study. We make use of the "Bump" function and the sine curve already from Section 3.2 and assume Poisson $n = 200$ distributed response variables $Y_i \sim \text{Poisson}\big(\exp\{\eta(x)\}\big)$ with $\eta(x) = -1 + 2\exp\big[-\{16(x - 0.5)^2\}\big]$ ("Bump") and $\eta(x) = \sin(6\pi x)$ ("sine"). The results are shown in Figure 8 in plot (a) to (c) for the "Bump" curve and (d) to (f) for the sine curve, respectively. In (a) and (d) we show a typical simulation and its true and fitted mean function $\exp\{\eta(x)\}$. Plots (b) and (c) give the shape of the log likelihoods based on 200 simulations. Finally, plots (c) and (f) give the histogram of the optimally chosen K. Overall, the behavior of the marginal log likelihood is comparable to the normal response case discussed in Section 2 and 3.

As a real example, we consider data showing the seasonal variation of toxoplasmosis in pregnancy. Let $\mathtt{y_i}$ denote the observed number of pregnant women with acute toxoplasmosis cases at time $\mathtt{t_i}$ found in regular screenings of more than 50,000 women in federal state of Upper Austria (see Sagel, Mikolajczyk, and Kraemer (2009)). The data are provided on a weekly basis for the years 2000 to 2005. We assume that $\mathtt{y_i}$ is Poisson distributed depending on seasonal variation and trend. Let $\mathtt{yweek_i} \in \{1, \ldots, 52\}$ be the week in the year, so that our model becomes

$$\mathtt{Y_i} \sim \text{Poisson}\big(\exp\{m(\mathtt{yweek_i}) + \mathtt{t_i}\beta_t\}\big).$$

We fit the model using a cyclic basis for m which guarantees that $\hat{m}(52)$ and $\hat{m}(1)$ merge. We start by fitting the model with small K and increase K sequentially until the likelihood reaches its maximum. This is shown in Figure 9 plot (a). The resulting optimal fit of the seasonal fit is shown in plot (b). Finally, in plot (c) we show the data and the fitted mean curve. The downward slope, that is the trend in the data, is clearly visible and significant (p-value=0.002).

5 Conclusion

In this article, we have proposed a likelihood-based method for selecting the number of knots in penalized spline regression. While such a method already existed under the smoothing framework (Ruppert, 2002), our approach is the first that can do so under the mixed model framework. An important advantage of the latter framework is that it makes it possible to combine model fitting and smoothing parameter (spline basis dimension and penalty) settings using a unified objective function. Not only does this simplify the computational aspects, it also further removes some of the "subjective" decisions often associated with applying nonparametric methods.

An interesting finding is that the knot selection rules of thumbs currently in use in penalized spline regression generally agree with our proposed approach. Similarly,

we showed the connections between the proposed likelihood-based approach and other approaches such as GCV and, in the case of generalized regression, AIC. A possible extension of this work is to evaluate the likelihood-based criterion for knot placement and for determining the degree of the spline. While this could certainly be considered, both factors tend to have only a minor impact of the spline fits once the penalty and the number of knots are allowed to vary. We therefore do not further pursue this here.

A Appendix

A.1 *Laplace Approximation*

Instead of truncated polynomials, we use the equivalent representation in the form of B-splines. Let P be the matrix with rows $P(x_i)$ and set $\theta = (\beta^T, u^T)^T$. With B we denote the normed B-spline basis $B = K^q P L_q$, where L_q is the $(q+1+K) \times (q+1+K)$ dimensional invertible matrix constructed from a $(q+1)$th order difference matrix, see for instance Fahrmeir et al. (2004), Kauermann et al. (2009) or Claeskens et al. (2009). For instance, L_1 is obtained as

$$L_1 = \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ -1 & 1 & 0 & 0 & \vdots & 0 \\ 1 & -2 & 1 & 0 & & \\ 0 & 1 & -2 & 1 & \vdots & \vdots \\ \vdots & & & & & \\ 0 & \dots & \dots & 1 & -2 & 1 \end{pmatrix}.$$

Defining $\omega = K^{-q} L_q^{-1} \theta$, we get $P\theta = B\omega$ and we assume that

$$\omega \sim N\left(0, \frac{\sigma_u^2}{K^{2q}} \left(L_q^T D L_q\right)^-\right) \tag{22}$$

with superscript $^-$ referring to the generalized inverse. Since $\tilde{D} = L_q^T D L_q$ is not of full rank, we decompose ω into $\omega = (\omega_1, \omega_2)$ with $\omega_1 \in \mathbb{R}^{(q+1)}$ as parameter

(corresponding to β in the truncated lines representation) and ω_2 as normally distributed spline coefficients (corresponding to u in the truncated lines version). In fact, with L_q decomposed into

$$L_q = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \in \begin{pmatrix} \mathbb{R}^{(q+1)\times(q+1)} & \mathbb{R}^{(q+1)\times K} \\ \mathbb{R}^{K\times(q+1)} & \mathbb{R}^{K\times K} \end{pmatrix}$$

and since $L_{12} = 0$, we obtain $\omega_1 = K^{-q} L_{11}^{-1} \beta$ and

$$\omega_2 \sim N\left(-L_{22}^{-1} L_{21} \omega_1, \frac{\sigma_u^2}{K^{2q}} L_{22}^{-1} L_{22}^{-T} \right).$$

We write $X\beta + Zu = K^q B_1 \omega_1 + K^q B_2 \omega_2$ with obvious definition for B_1, and B_2 and the marginal likelihood (19) becomes

$$l_K(\omega_1, \sigma_u^2) = -\frac{K}{2} \log\left(\sigma_u^2 / K^{2q}\right) + \log|L_{22}| + \int \exp\left\{ l_{K,p}(\omega_1, \omega_2, \sigma_u^2) \right\} \, \mathrm{d}\omega_2$$

where $l_{K,p}(\omega_1, \omega_2, \sigma_u^2)$ equals $l_{K,p}(\beta_1, u, \sigma_u^2)$ with a slight but obvious misuse of notation.

Following Shun and McCullagh (1995), we can now express the error of the Laplace approximation through

$$l_K(\omega_1, \sigma_u^2) = \tilde{l}_K(\omega_1, \sigma_u^2) + \delta_K, \tag{23}$$

with δ_K as approximation error and $\tilde{l}_K(\omega_1, \sigma_u^2) = \tilde{l}_K(\beta, \sigma_u^2)$, again with a slight misuse of notation. The dominating terms in δ can be written following Einstein's summation convention (see Barndorff-Nielsen and Cox, 1989). Let $l_{jr} = \partial l_{K,p}(\omega_1, \omega_2, \sigma_u^2)/(\partial\omega_{2j}\ \partial\omega_{2r})$ denote the second order derivative of the penalized likelihood with respect to the j-th and r-th component of ω_2. In the same way we denote l_{jrs} the component of the third order derivative, and so on. With l^{jr} we denote the (j,r)-th element of the inverse of $\partial l_{K,p}(\omega_1, \omega_2, \sigma_u^2)/(\partial\omega_2)(\partial\omega_2^T)$. Summing now over equal super- and subscripts, we obtain (see Shun and McCullagh, 1995)

$$\delta = -l_{jlrs}\ l^{jl}\ l^{rs}[3]/24 + l_{jlr}\ l_{stv}\left(l^{jl}\ l^{rs}\ l^{tv}[9] + l^{js}\ l^{lt}\ l^{rv}[6]\right)/72 + \ldots. \tag{24}$$

Bracketed terms above indicate the number of possible permutations which are not explicitly listed in the formula for brevity. Note that

$$\frac{\partial l_{K,p}(\omega_1, \omega_2, \sigma_u^2)}{\partial \omega_2 \, \partial \omega_2^T} = B_2^T W(\omega) B_2 + \frac{K^{2q}}{\sigma_u^2}(L_{22}^T \, L_{22}) \tag{25}$$

and with the structure of B-splines and the difference matrix L we easily see that (25) gives a band diagonal matrix with bandwidth of order $q+1$ and diagonal elements of order $O(n/K) + O(K^{2q}/\sigma_u^2)$. Accordingly, the third and fourth order derivatives are of (three and four dimensional) band diagonal structure with, e.g. $l_{jlr} = 0$ if $|j-l|, |j-r|$ or $|l-r| > q+1$ and $l_{jlr} = O(n/K)$ otherwise. This implies that the second order derivative written as a matrix is strictly diagonal dominant, which allows us to make use of results derived by Demko (1977). We can find an upper bound $0 < \xi < 1$ such that $l^{jl} \leq \xi^{|j-l|} \, O\big\{(n/K + K^{2q}/\sigma_u^2)^{-1}\big\}$. Bearing in mind that any summation is over K elements in (24), we calculate the order of δ_K in (23) to be

$$\delta_K = O\left\{ n \left(\frac{n}{K} + \frac{K^{2q}}{\sigma_u^2} \right)^{-2} \right\} + O\left\{ \frac{n^2}{K} \left(\frac{n}{K} + \frac{K^{2q}}{\sigma_u^2} \right)^{-3} \right\}.$$

For K increasing and assuming n to be fixed, we obtain with Assumption 3 that $\delta_K = O(nK^{-4q-2})$. Hence, the error is of ignorable size, so that the Laplace approximation is justifiable.

A.2 *Approximate Expectation of $l_{K,c}(\hat{\beta}, \hat{u})$ in (21)*

First, let

$$l_{K,p}(\theta, \sigma_u^2) := l_{K,p}(\beta, u, \sigma_u^2) = l_{K,c}(\beta, u) - \frac{1}{2}\frac{u^T u}{\sigma_u^2}$$

denote the penalized likelihood and let $\hat{\theta} = (\hat{\beta}, \hat{u})$ be the maximizer of $l_{K,p}(\theta, \sigma_u^2)$. We define $s_{K,p}(\theta, \sigma_u^2) = \partial l_{K,p}(\theta, \sigma_u^2)/\partial\theta = \partial l_{K,c}(\theta)/\partial\theta - D\theta/\sigma_u^2$ as the penalized score with $s_{K,c}(\theta) = \partial l_{K,c}(\theta)/\partial\theta$ as unpenalized version and D as defined in Section 2, i.e. as diagonal matrix with 0_{q+1} and 1_K on the diagonal. Conditioning on u, we have $E\big(s_{K,c}(\theta)|u\big) = 0$ which implicitly defines β. Moreover,

$E(s_{K,c}(\theta)s_{K,c}^T(\theta)|u) = PW(\theta)P =: F_{K,c}(\theta)$ defines the Fisher information matrix conditional on u. Using standard expansions, we obtain

$$\begin{aligned} 0 = s_{K,p}(\hat{\theta}, \sigma_u^2) &= s_{K,p}(\theta, \sigma_u^2) + \frac{\partial s_{K,p}(\theta, \sigma_u^2)}{\partial \theta}(\hat{\theta} - \theta) + \dots \\ \Leftrightarrow \hat{\theta} - \theta &= F_{K,p}^{-1}(\theta, \sigma_u^2)\ s_{K,p}(\theta, \sigma_u^2) + \dots \\ &= \big\{F_{K,c}(\theta) + \sigma_u^{-2}\ D\big\}^{-1}\big\{s_{K,c}(\theta) - \sigma_u^{-2}D\theta\big\} + \dots \end{aligned} \tag{26}$$

From the definition of θ, we obtain $E(\hat{\theta} - \theta|u) \approx -\big\{F_{K,c}(\theta) + \sigma_u^{-2}D\big\}^{-1}\sigma_u^{-2}D\theta$ as approximate conditional smoothing or penalization bias, and following the Generalized Linear Mixed Model (18) we obtain approximate unbiasedness $E(\hat{\theta} - \theta) \approx 0$. Expanding the maximum value of the likelihood in (21) now yields

$$\begin{aligned} &l_{K,c}(\hat{\beta}, \hat{u}) = l_{K,c}(\hat{\theta}) \\ &\approx l_{K,c}(\theta) + s_{K,c}(\theta)\big\{F_{K,c}(\theta) + \sigma_u^{-2}D\big\}^{-1}\big\{s_{K,c}(\theta) - \sigma_u^{-2}D\theta\big\} \\ &- \frac{1}{2}\big\{s_{K,c}(\theta) - \sigma_u^{-2}D\theta\big\}\big\{F_{K,c}(\theta) + \sigma_u^{-2}D\big\}F_{K,c}(\theta)\big\{F_{K,c}(\theta) + \sigma_u^{-2}D\big\}^{-1}\big\{s_{K,c}(\theta) + \sigma_u^{-2}D\theta\big\} + \dots \end{aligned}$$

Taking expectation with respect to Y and conditioning on u gives

$$\begin{aligned} &E_Y\big\{l_{K,c}(\hat{\beta}, \hat{u})|u\big\} \\ &\quad\approx E_Y\big\{l_{K,c}(\beta, u)|u\big\} + \mathrm{tr}\big\{H_K(\theta, \sigma_u^2)\big\} - \frac{1}{2}\ \mathrm{tr}\big\{H_K(\theta, \sigma_u^2)\ H_K(\theta, \sigma_u^2)\big\} \qquad (27) \\ &\quad- \frac{1}{2}\ \mathrm{tr}\big[\big\{\sigma_u^{-2}D\theta\ \theta^T\sigma_u^{-2}D\big\}H_K(\theta, \sigma_u^2)\big\{F_{K,c}(\theta) + \sigma_u^{-2}D\big\}^{-1}\big] \qquad (28) \end{aligned}$$

with $H_K(\theta, \sigma_u^2) = \big\{F_{K,c}(\theta) + \sigma_u^{-2}D\big\}^{-1}F_{K,c}(\theta)$. Taking expectation with respect to u, observing that $E_u(\theta\theta^T) = \sigma_u^2 D^-$ and applying Laplace approximation where necessary, yields

$$E\big\{l_{K,c}(\hat{\beta}, \hat{u})\big\} = E\big\{l_{K,c}(\beta, u)\big\} + \frac{1}{2}\ \mathrm{tr}\big\{H_K(\tilde{\theta}, \sigma_u^2)\big\} \tag{29}$$

where $\tilde{\theta} = (\beta, \tilde{u})$ with $\tilde{u}$ as maximizer of $l_{K,p}(\beta, u, \sigma_u^2)$ for given β. Defining $S_{K,\lambda} = P(P^TWP + \lambda D)^{-1}P^TW$ as a generalized smoothing matrix, we can express the trace component in (29) as $\mathrm{tr}(S_{K,\lambda})$, mirroring the degrees of the model (see e.g. Hastie and Tibshirani, 1990). We may now apply similar arguments as in (16) for

the normal case to show that $\mathrm{tr}(S_{K,\lambda}) = O\big(1 + K^{\log(a_q)+1}\big)$. Hence, we conclude that

$$E\big\{l_{K,c}(\hat{\beta},\hat{u})\big\} = \big\{l_{K,c}(\beta,u)\big\} + O\big(1 + K^{\log(a_q)+1}\big).$$

A.3 *Behavior of* $E\big\{l_{K,c}(\beta,u)\big\}$

We write the log likelihood as $l_{K,c}(\beta,u) = \sum_{i=1}^{n}\{Y_iP_i\theta - b(P_i\theta)\}$ where $P_i = P(x_i)$ as basis and $b(.)$ is the cumulant generating function of the underlying exponential family distribution. We define the conditional mean value of the likelihood $M(\beta,u) = E_Y\{l_{K,c}(\beta,u)|u\} = \sum_{i=1}^{n}\{b'(P_i\theta)P_i\theta - b(P_i\theta)\}$ where $b'(P_i\theta) = h(P_i\theta)$ with $h(.)$ as canonical link. Let m_{rs} denote the derivative $\partial^2 M(\beta,u)/(\partial u_r)(\partial u_s)\big|_{u=0}$, and corresponding notation for higher order derivatives. We assume that all elements of the derivative are of order $O(1)$, which is implied if $\eta_i = X_i\beta$ is strictly inside the natural parameter space for all $i = 1,\ldots,n$. Using again Einstein's summation convention (see McCullagh, 1987) and considering the independence of coefficients u as stated in model (18), we obtain

$$E\big\{l_{K,c}(\beta,u)\big\} = M(\beta,0) + \sigma_u^2\ 1^{jj}\ M_{jj} + 3\ \sigma_u^4\ 1^{jj}\ 1^{ll}\ M_{jjll} + \ldots \qquad (30)$$

with $1^{lj} = 1$ for all l,j. Considering the structure of M_{ij} as matrix gives Z^TWZ with W as diagonal weight matrix with elements $b''(X_i\beta)$. With similar arguments used in Section 2 to derive the order of (12), we find that $\mathrm{tr}(Z^TWZ) = 1^{jj}m_{jj} = O(nK)$. Utilizing Assumption 3, we conclude that the second term in (30) does not depend on K for K sufficiently large. The same holds for the third and not explicitly listed higher order components in (30). Hence, we obtain $E\big\{l_{K,c}(\beta,u)\big\} = E\big\{l_{K,c}(\beta,0)\big\} + O(1)$.

References

Barndorff-Nielsen, O. and D. Cox (1989). *Asymptotic Techniques for use in Statistics*. London: Chapman and Hall.

Claeskens, G., T. Krivobokova, and J. Opsomer (2009). Asymptotic properties of penalized spline estimators. to appear.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

Demko, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM Journal on Numerical Analysis 14*, 616–619.

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

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized sturctured additive regression for space-time data: A bayesian perspective. *Statistica Sinica 14*, 715–745.

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

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. London: Chapman and Hall.

Kauermann, G., T. Krivobokova, and L. Fahrmeir (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B 71*, 487–503.

Krivobokova, T. and G. Kauermann (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*. to appear.

Li, Y. and D. Ruppert (2008). On the asymptotics of penalized splines. *Biometrika 95*, 415–436.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman & Hall.

Ngo, L. and M. P. Wand (2004). Smoothing with mixed model software. *Journal of Statistical Software 9*, 1–54.

Opsomer, J., G. Claeskens, G. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society 70*, 265–286.

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

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

Ruppert, D., M. Wand, and J. Carroll (2009). Semiparametric regression during 2003-2007. Technical report.

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

Sagel, U., R. Mikolajczyk, and A. Kraemer (2009). Seasonal trends in acute toxoplasmosis in pregnancy in the federal state of upper austria. *Clinical Microbiology and Infection*. in press.

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

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

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

Wand, M. and J. Ormerod (2008). On semiparametric regression with o'sullivan penalized splines. *Australian and New Zealand Journal of Statistics 50*, 179–198.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics 15*, 443–462.

Wood, S. (2006). *Generalized Additive Models*. London: Chapman & Hall.

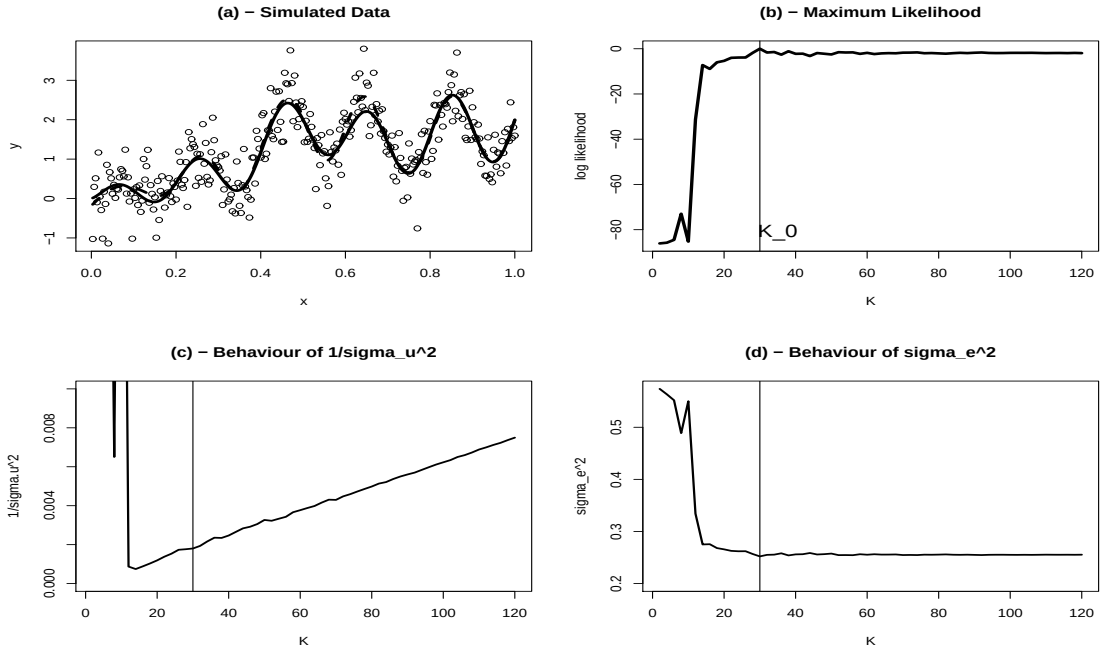

Figure 1: (a) Simulated data and true function (bold) and best fitted function (dashed). (b) Value of marginal log likelihood as a function of K. (c) Behavior of $1/\hat{\sigma}_u^2$ and (d) behavior of $\hat{\sigma}_\epsilon^2$.

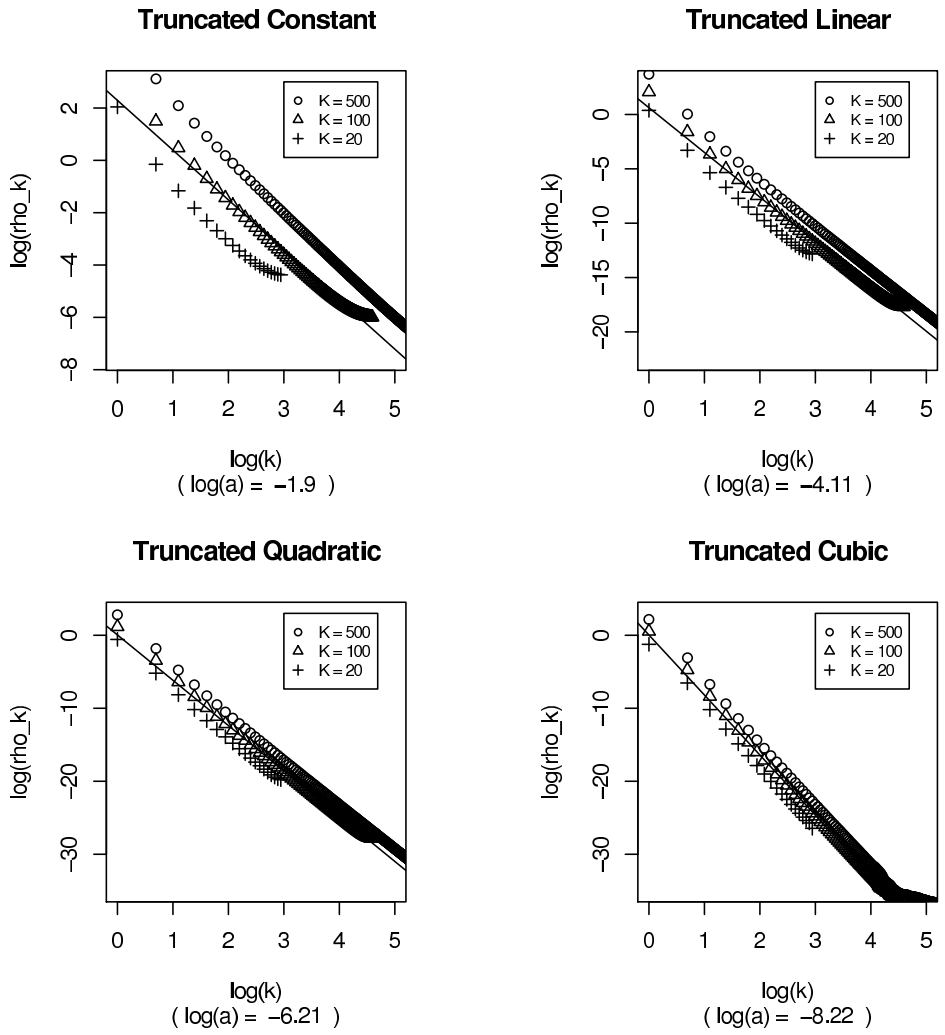

Figure 2: Eigenvalues ρ_k of matrix R_K defined in (13) plotted against $\log(k), k = 1, \ldots, K$, for different values of q ranging from $q = 0$ to $q = 3$. Approximate linear slope ($log(a_q)$) is included as linear diagonal line, see also bottom of each plot.

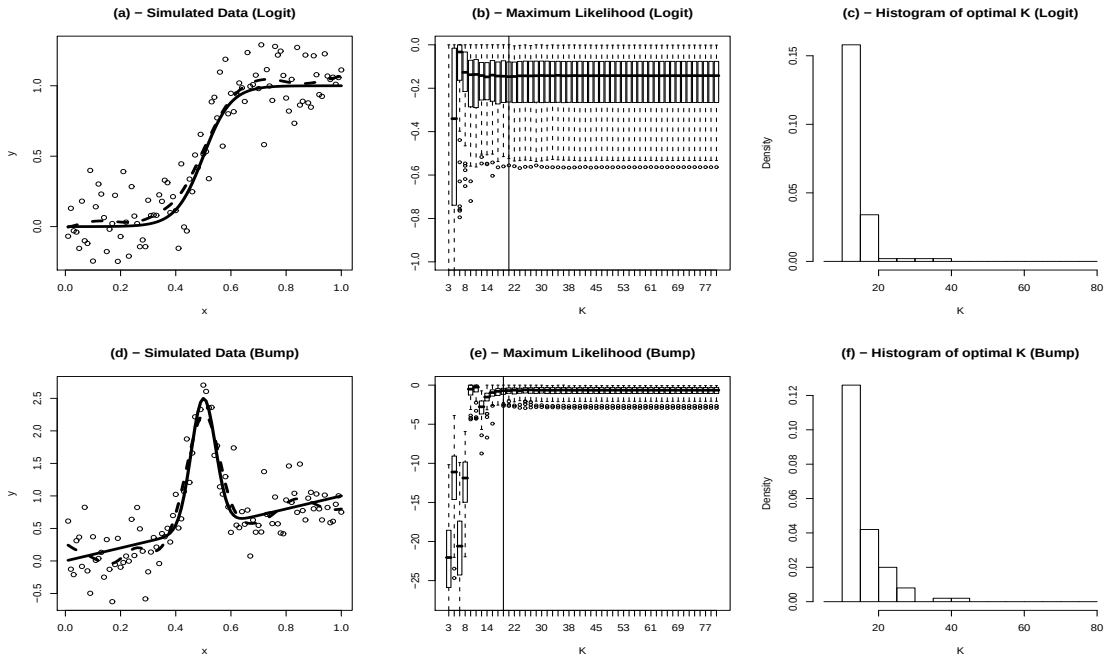

Figure 3: Simulation results for the "Logit" and "Bump" mean functions. Logit function: (a) typical realization with mean function (solid line) and spline fit (dashed line), (b) boxplots of the value of the maximized likelihood for different values of K, vertical line gives the median of the optimally selected K, (c) histogram of the values of K maximizing the sample likelihoods; corresponding plots for Bump function are in (d)-(f).

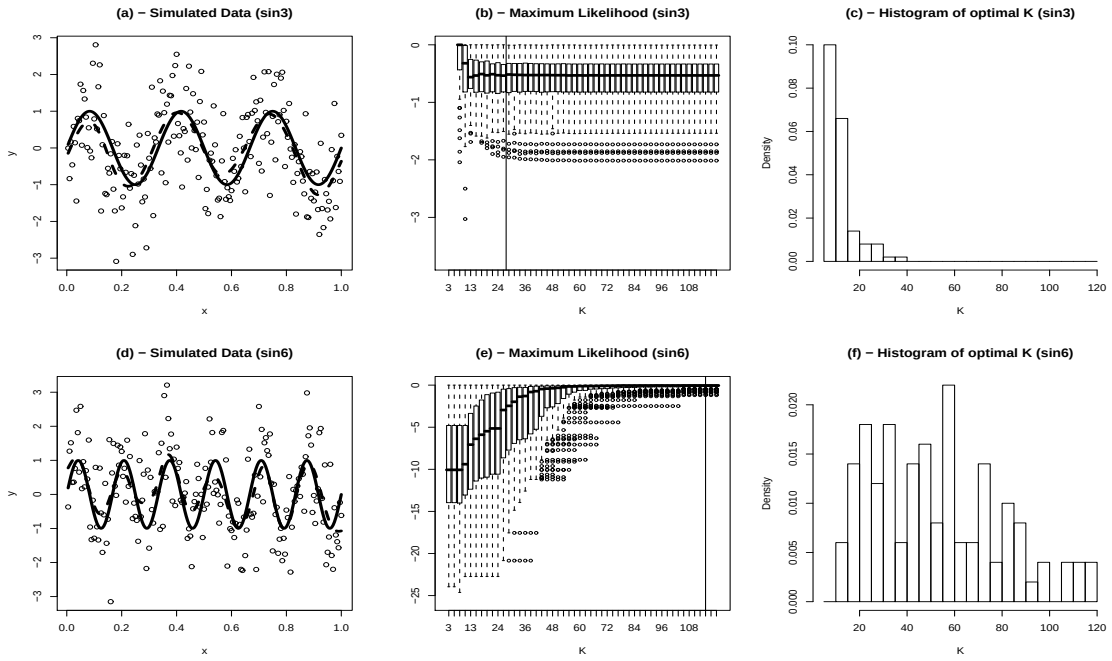

Figure 4: Simulation results for two sine mean functions. Plots (a)-(c) show results for sine function with $\theta = 3$ and plots (d)-(f) show results for $\theta = 6$. See Figure 3 for description of individual plots.

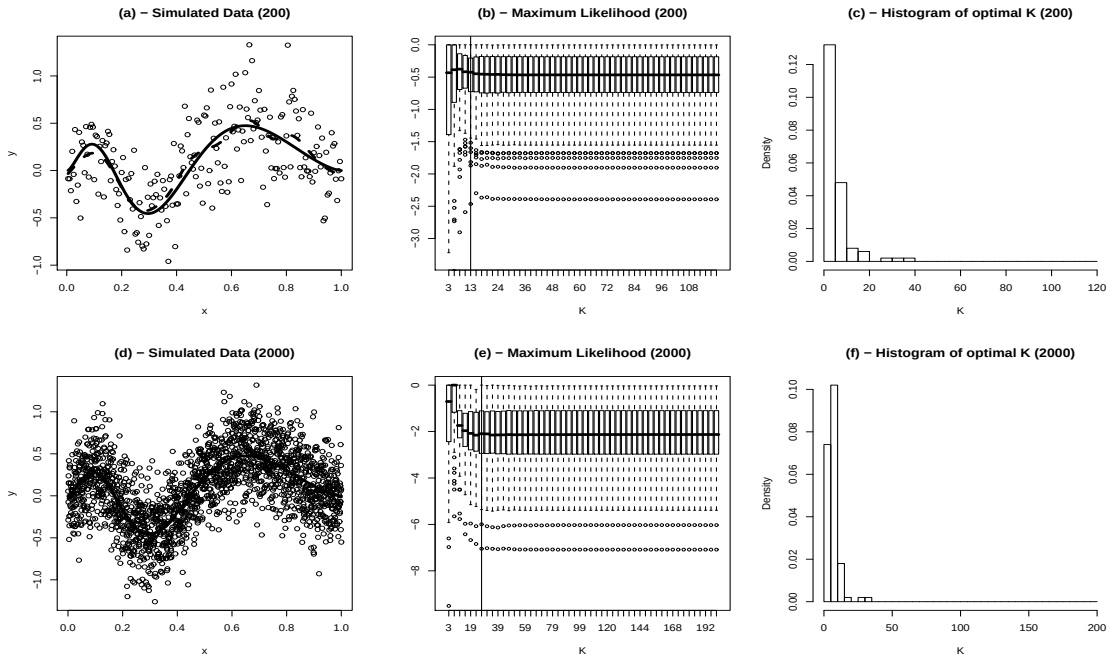

Figure 5: Simulation results for spatial heterogeneity function. Plots (a)-(c) show results for $n = 200$ and plots (d)-(f) show results for $n = 2000$. See Figure 3 for description of individual plots.

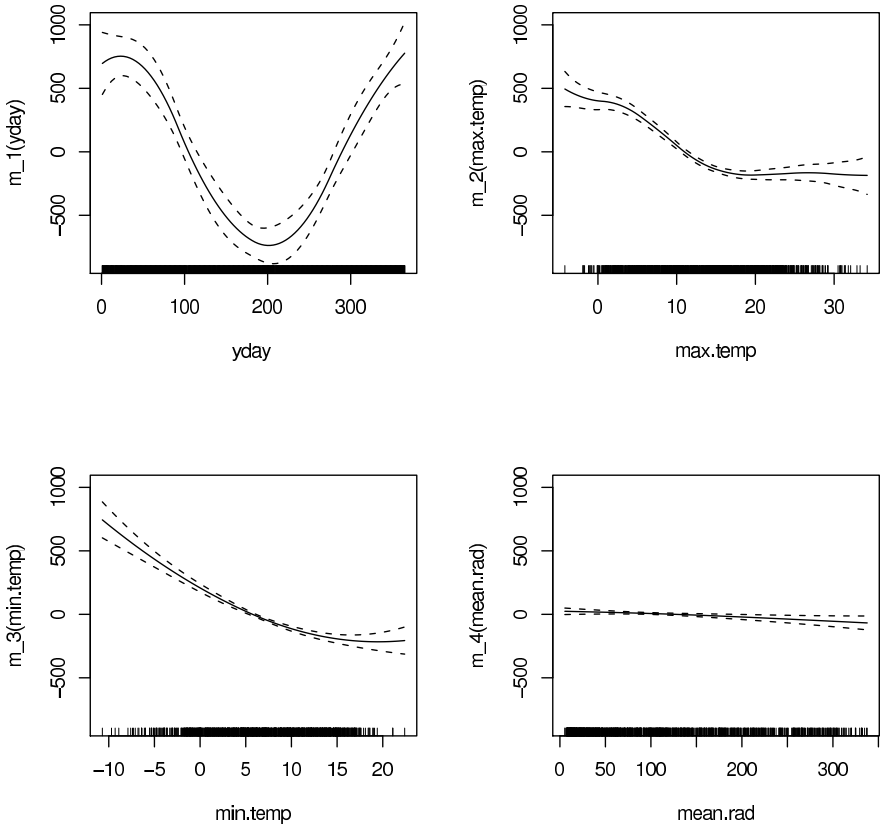

Figure 6: Additive model fit with optimized spline dimension showing the influence of `yday`, `max.temp`, `min.temp` and `mean.rad`.

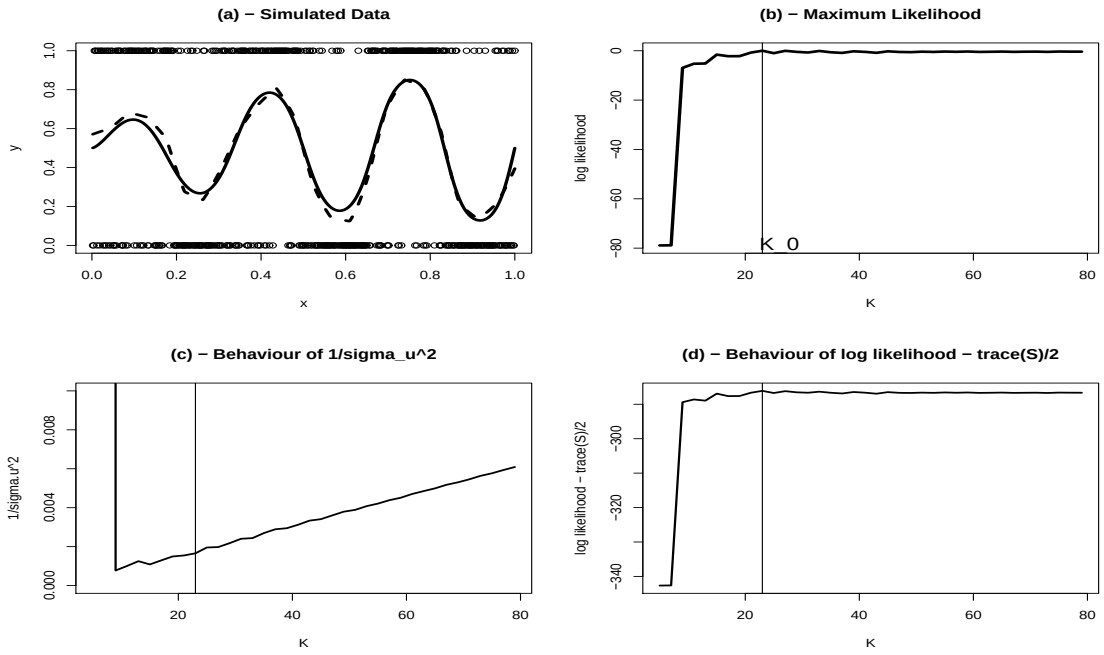

Figure 7: (a) Simulated binary data with true mean function (bold) and best fitted function (dashed). (b) Marginal approximate log likelihood as a function of K. (c) Behavior of $1/\hat{\sigma}_u^2$ and (d) behavior of $\tilde{l}_K(\hat{\beta}, \hat{\sigma}_u^2) - \frac{1}{2}\,\mathrm{tr}(S_{\tilde{\lambda}})$.

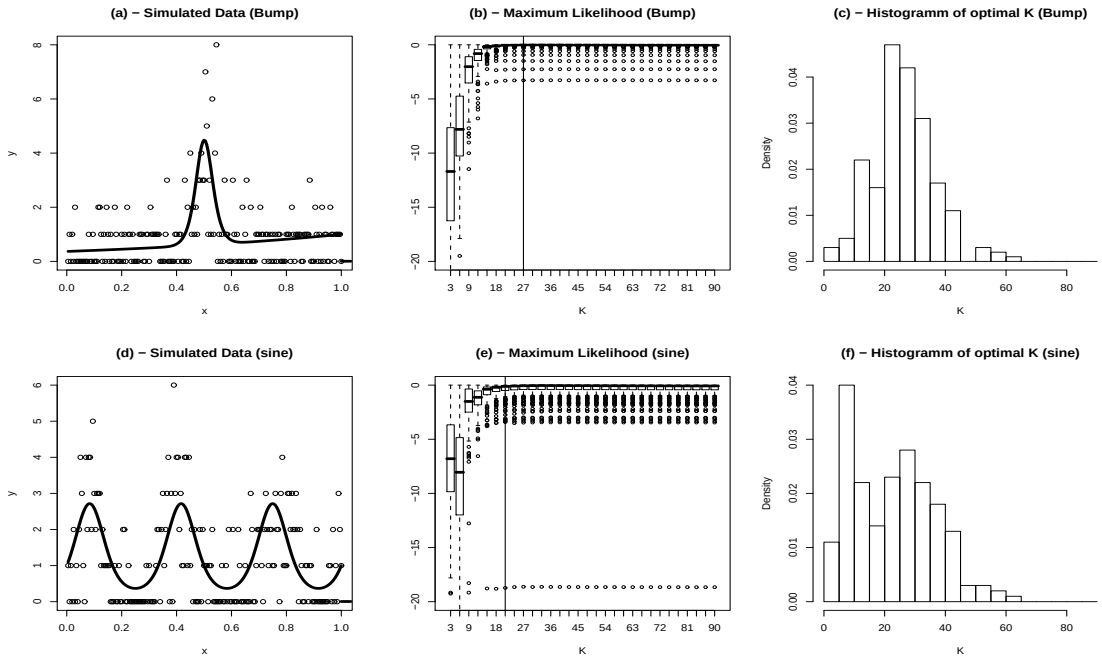

Figure 8: Simulation results for Poisson data with "Bump" and "Sine" mean functions. Bump function: (a) typical realization with mean function (solid line) and spline fit (dashed line), (b) boxplots of the value of the maximized likelihood for different values of K, vertical line gives the median of the optimally selected K, (c) histogram of the values of K maximizing the marginal sample likelihoods; corresponding plots for Sine function are in (d)-(f).

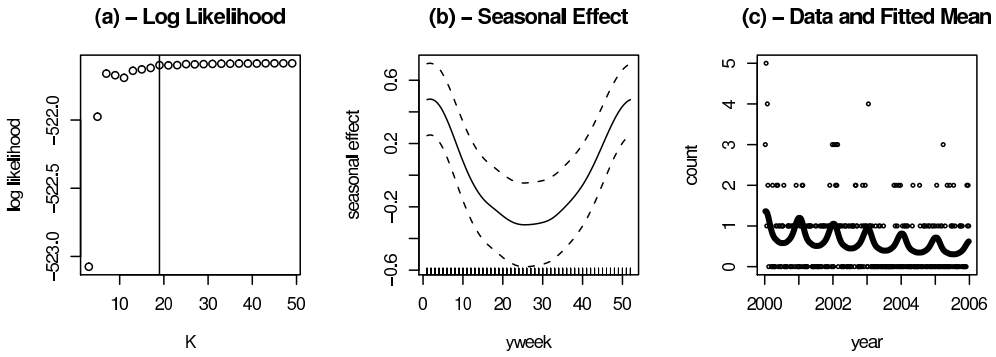

Figure 9: Marginal log likelihood for toxoplasmosis data (a), fitted seasonal effect $\hat{m}(\texttt{yweek})$ (b) and fitted mean function and data (c).