Some Problems in Model Specification and Inference for Generalized Additive Models

Marra, Giampiero

Award date:
2010

Awarding institution:
University of Bath

Some Problems in Model Specification and Inference for Generalized Additive Models

submitted by

Giampiero Marra

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

December 2010

Summary

Regression models describing the dependence between a univariate response and a set of covariates play a fundamental role in statistics. In the last two decades, a tremendous effort has been made in developing flexible regression techniques such as generalized additive models (GAMs) with the aim of modelling the expected value of a response variable as a sum of smooth unspecified functions of predictors. Many nonparametric regression methodologies exist including local-weighted regression and smoothing splines. Here the focus is on penalized regression spline methods which can be viewed as a generalization of smoothing splines with a more flexible choice of bases and penalties.

This thesis addresses three issues. First, the problem of model misspecification is treated by extending the instrumental variable approach to the GAM context. Second, we study the theoretical and empirical properties of the confidence intervals for the smooth component functions of a GAM. Third, we consider the problem of variable selection within this flexible class of models. All results are supported by theoretical arguments and extensive simulation experiments which shed light on the practical performance of the methods discussed in this thesis.

Acknowledgements

I am very grateful to my advisor Professor Simon N. Wood for his assistance, guidance and support and for sharing his ideas and knowledge with me over the past three years. Simon introduced me to a fascinating area of statistics and gave me the opportunity to work on my ideas.

I am indebted to my wife Rosalba Radice for supporting me in every aspect of my scientific and personal life and for our ongoing productive and stimulating collaboration. Without her, I genuinely believe I would not have made it this far.

I also would like to thank Professor Julian Faraway for many detailed and helpful suggestions on the earlier draft of this thesis, and David L. Miller for many interesting discussions and for providing valuable advice on linguistic matters. Finally, I thank my family for believing in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the main objectives of regression modelling is to model the expected value of a response variable Y as a flexible function of regressors $x_1, \ldots, x_p$. In other words, the aim is to specify a function f such that

$$\mathbb{E}(Y|x_1, \ldots, x_p) = h\{f(x_1, \ldots, x_p)\},$$

where $h(\cdot)$ is the inverse of a link function, and Y follows an exponential family distribution. Replacing $f(\cdot)$ with a linear combination of some known functions of covariates, e.g. $f(x_1, \ldots, x_p) = \theta_0 + \sum_{j=1}^{p} \theta_j x_j$, leads to a generalized linear model (GLM; McCullagh and Nelder, 1989) which is easy to estimate and to interpret, and for which well-developed statistical frameworks are available. However, since the functional shape of any relationship is rarely known *a priori* and the response of interest may depend on the predictors in a complicated manner, it is more convenient to model $f(\cdot)$ as the sum of some unspecified smooth function of covariates, e.g. $f(x_1, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j)$, hence giving rise to a generalized additive model (GAM; Hastie and Tibshirani, 1990). Such a model allows for rather flexible specification of the dependence of the response on the covariates, but this flexibility and convenience comes at the cost of new methodological problems, some of which will be the objective of this thesis.

1.1 Objectives of Thesis and Outline

This thesis deals with some aspects of penalized regression spline smoothing. We shall begin by discussing some background material and then concentrate on three issues.

First, we consider the problem of model misspecification in the GAM context. Specifically, when unobservables are associated with included regressors and have an impact on the response, standard estimation methods will not be valid. This means, for example, that estimation results from observational studies, whose aim is to evaluate the impact of a treatment of interest a response variable, will be biased and inconsistent in the presence of unmeasured confounders if these are not accounted for. One method for obtaining consistent estimates of treatment effects when dealing with linear models and GLMs is the instrumental variable (IV) approach. Fitting procedures to carry out IV analysis within the GAM context have not been developed. Following the idea first introduced by Hausman (1978, 1983), we propose a two-stage approach for IV estimation when dealing with GAMs, and a correction procedure for confidence intervals. We explain under which conditions the proposed method works and illustrate its empirical validity through an extensive simulation experiment and a health study where unmeasured confounding is suspected to be present.

Second, we study the coverage properties of the Bayesian 'confidence' intervals for the smooth component functions of GAMs. The intervals are the usual generalization of Wahba (1983) or Silverman (1985) intervals to the GAM component context. We present simulation evidence showing these intervals have close to nominal across-the-function frequentist coverage probabilities, except when the truth is close to a straight line/plane function. We extend Nychka's (1988) argument for univariate smoothing splines to explain these results. The theoretical results allow us to derive alternative intervals from a purely frequentist point of view, and to explain the impact that the neglect of smoothing parameter variability has on confidence interval performance. They also suggest switching the target of inference for component-wise intervals away from smooth components in the space of the GAM identifiability constraints.

Third, we face the problem of GAM component selection. We propose two effective methods and extend the nonnegative garrote estimator, originally introduced by Breiman (1995), to achieve smooth term selection. The proposals avoid having to use nonparametric testing methods for which there is not a general reliable distributional theory. Moreover, variable selection is carried out in one single step as opposed to many selection procedures which involve an exhaustive search of all possible models. The empirical performance of the proposed methods is compared to that of some available techniques via an extensive simulation study. Our results show under which conditions one

method can be preferred over another, hence providing applied researchers with some practical guidelines. The procedures are also illustrated analysing data on plasma beta-carotene levels from a cross-sectional study conducted in the United States.

This thesis is based on the following papers:

- MARRA, G., RADICE, R., 2010. Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, 19, pp. 107–125.
- MARRA, G., RADICE, R., 2012. A Flexible Instrumental Variable Approach. *Statistical Modelling*, in press.
- MARRA, G., WOOD, S. N. Coverage Properties of Confidence Intervals for Generalized Additive Model Components. Submitted.
- MARRA, G., WOOD, S. N. Practical Variable Selection for Generalized Additive Models. Submitted.

Chapter 2

Generalized Additive Models: An Overview

GAMs allow for flexible functional dependence of a response variable on covariates. The aim of this chapter is to provide a brief overview of this flexible class of models, based on the penalized likelihood framework with regression splines, by discussing some aspects that are relevant to this thesis.

2.1 Introduction

GAMs are becoming among the most useful and used of statistical methods. An ISI Web of Knowledge search on the keyword "generalized additive models" reveals over 800 articles published during the last decade in the fields of biology, ecology, economics, environmental science, epidemiology, genetics and medicine (e.g. Marra and Radice, 2011; Marra and Radice, 2010; Zanin and Marra, 2011). This approach extends traditional GLMs by allowing the determination of possible nonlinear effects of covariates on a response variable of interest. In other words, GLMs model the effects of predictor variables x_j in terms of a linear predictor of the form $\theta_0 + \sum_j \theta_j x_j$, where the θ_j are regression parameters, whereas GAMs replace $\theta_0 + \sum_j \theta_j x_j$ with, for instance, $\sum_j f_j(x_j)$, where the f_j are unknown smooth functions of regressors. The use of smooth terms is crucial since the functional shape of any relationship is rarely known *a priori* and the response of interest may depend on the predictors in a complicated manner.

A number of procedures can be employed for fitting GAMs, some of them documented in two recent monographs (Ruppert *et al.*, 2003; Wood, 2006), and there is ongoing research on new ones such as the likelihood-based boosting

approach (Tutz and Binder, 2006). Our investigation is not meant to be extensive. Rather, our goal is to present some background material on the penalized likelihood based approach with regression splines since this is the framework that will be adopted throughout this thesis.

2.2 Model Structure

A GAM can be seen as a GLM with a linear predictor involving smooth functions of covariates

$$g\{\mathbb{E}(Y_i)\} = \eta_i = \mathbf{X}_i^*\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots, \tag{2.1}$$

where $g(\cdot)$ is a smooth monotonic twice differentiable link function, Y_i is a univariate response variable, $\mathbf{X}_i^*$ is the i^{th} row of $\mathbf{X}^*$, which is the model matrix for any strictly parametric model components, with corresponding parameter vector $\boldsymbol{\theta}$, and the f_j are smooth functions of the covariates x_j. The f_j are subject to identifiability constraints such as $\sum_i f_j(x_{ji}) = 0 \ \forall j$. The right hand side of (2.1) is called the linear predictor and is denoted as η_i, and the response Y_i follows an exponential family distribution whose probability density functions are of the form

$$m_\vartheta(y) = \exp\left[\frac{y\vartheta - h(\vartheta)}{\phi} + c(y, \phi)\right], \tag{2.2}$$

where $h(\cdot)$ and $c(\cdot)$ are arbitrary functions, ϑ is the natural and ϕ the dispersion parameter. The mean and variance of such a distribution are $\mathbb{E}(Y) = \partial h(\vartheta)/\partial\vartheta = \mu$ and $\text{var}(Y) = \phi\partial\mu/\partial\vartheta = \phi V(\mu)$, respectively, where $V(\mu)$ denotes the variance function. Several distributions are possible within this family, such as the binomial, gamma, Gaussian and Poisson. In fact, a whole variety of outcome measures (e.g. counts, binary and skewed data) can be modelled within this model structure. In some cases, the nature of the response distribution is not known, and it is only possible to specify what the relationship between the variance of the response and its mean should be. It turns out that it is possible to develop theory for fitting and inference based on the notion of quasi-likelihood. Here, maximum quasi-likelihood parameter estimates can be found by the usual method used to fit a GLM, described in the next section, and the classic large sample distribution of GLM parameter estimators also hold for maximum quasi-likelihood.

Model (2.1) can flexibly determine the functional shape of the relationship between a response and some explanatory variables, hence avoiding the draw-

backs of modelling data using parametric relationships. As an example, let us consider a group of patients from a single hospital who underwent Coronary Artery Bypass Graft surgery. One may wish to identify the risk factors of in-hospital mortality following surgery, where the outcome of interest is *Status* (0=alive, 1=died) and the explanatory variables associated with surgical mortality could be *Age*, *BSA* (Body Surface Area), and *Ejection Fraction* (a measure of heart function summarized in the categories 'Good', 'Fair' and 'Poor'). In order to explain the in-hospital mortality following surgery from these explanatory variables, several model specifications can be adopted. A possibility would be to fit a GLM with linear predictor given by

$$\eta_i = \theta_0 + \theta_1 EF_{fair,i} + \theta_2 EF_{poor,i} + \theta_3 Age_i + \theta_4 BSA_i, \tag{2.3}$$

where θ_0 represents the baseline group *Ejection Fraction* = 'Good'. But we do not know whether the variables *Age* and *BSA* enter the model linearly, and (2.3) makes the assumption of linear relationships between the two continuous variables and response. Instead, one could employ the following GAM

$$\eta_i = \theta_0 + \theta_1 EF_{fair,i} + \theta_2 EF_{poor,i} + f_1(Age_i) + f_2(BSA_i).$$

In this way the relationship between the in-hospital mortality and the continuous variables in the model can be determined flexibly. One of main advantages of GAMs is that residual confounding may be avoided. This is supported by the simulation study of Benedetti and Abrahamowicz (2004) which shows that the use of spline models reduces residual confounding as compared to fully parametric modelling which typically leads to biased and spurious estimated impacts of the exposure of interest, in the presence of unmodelled nonlinearities. However, when unmeasured covariates are correlated with included regressors and have an impact on the response, any GAM estimation method will not be valid, no matter how reliable and computationally robust the method is. Chapter 3 addresses this issue by showing how model misspecification can be dealt with in the GAM context.

The smooth terms can be represented using regression splines. In particular, the regression spline of a predictor is made up of a linear combination of known basis functions, $b_{jk}(x_j)$, and unknown regression parameters, β_{jk},

$$f_j(x_j) = \sum_{k=1}^{q_j} \beta_{jk} b_{jk}(x_j), \tag{2.4}$$

where j indicates the smooth term for the j^{th} explanatory variable, q_j is the number of basis functions, hence regression parameters, used to represent the j^{th} smooth term, and the subscript i is dropped for simplicity. Similarly, the regression spline of two covariates can be written as $f_{jp}(x_j, x_p) = \sum_{k=1}^{q_j} \beta_{jp,k} b_{jp,k}(x_j, x_p)$. As mentioned earlier on, in order to identify (2.1), each smooth component is subject to some identifiability constraint. Basis functions have to be chosen in order to come up with smooth component estimates. For instance, suppose that $f_1(Age)$ is believed to be a 3^{th} order polynomial. A basis for this space is $b_{11}(Age) = 1$, $b_{12}(Age) = Age$, $b_{13}(Age) = Age^2$ and $b_{14}(Age) = Age^3$. Here, expression (2.4) becomes

$$f_1(Age) = \sum_{k=1}^{4} \beta_{1k} b_{1k}(Age) = \beta_{11} + \beta_{12} Age + \beta_{13} Age^2 + \beta_{14} Age^3,$$

which can be easily estimated using standard regression techniques. The number of basis functions, q_j, determines the maximum possible flexibility allowed for a smooth term. For example, a q_j equal to 20 will yield a "wigglier" nonlinear estimate as compared to the estimate that can be obtained when this parameter is set to 10. It is worth observing that, although quite illustrative, polynomial bases are not very useful in practice. As the number of basis functions increases, polynomial bases become increasingly collinear. This yields highly correlated parameter estimators, hence leading to high estimator variance and numerical problems (e.g. Royston, 2005). For these reasons, such basis functions should not generally be employed to model nonlinear relationships. As a practical solution, in some applied work, continuous variables are categorized into groups based on intervals or frequencies. However, categorization has several disadvantages since it introduces the problem of defining cut-points and implies that the relationship between a response variable and a set of covariates is flat within intervals (Royston and Altman, 1994; Johansen *et al.*, 2005). To overcome all these issues, spline bases are typically used to determine flexibly the relationship between the continuous predictors and the outcome of interest. In fact, they avoid the disadvantages of categorization, are not as correlated as polynomial basis functions, have convenient mathematical properties and good numerical stability. Common choices for representing smooth functions include smoothing splines (e.g. Hastie and Tibshirani, 1990; Wahba, 1990). These place knots at every data point, and are indeed sometimes referred to as full rank smoothers because the size of the spline basis is equal to the number of observations. However, such smoothers have as many

unknown parameters as there are data which results in expensive computations. The thin plate regression spline basis proposed by Wood (2003) is a valid alternative. This basis is a low rank eigen-approximation version of the full rank thin plate spline introduced by Duchon (1977). It represents a general solution to the problem of estimating efficiently, and without having to choose knot locations, a smooth function of multiple predictor variables from noisy observations of the function, at particular values of those predictors. Figures 2-1 and 2-2 illustrate thin plate regression spline bases in one dimension and two dimensions, respectively. Full mathematical details can be found in Wood (2003, 2006). This spline basis will be used throughout this thesis.

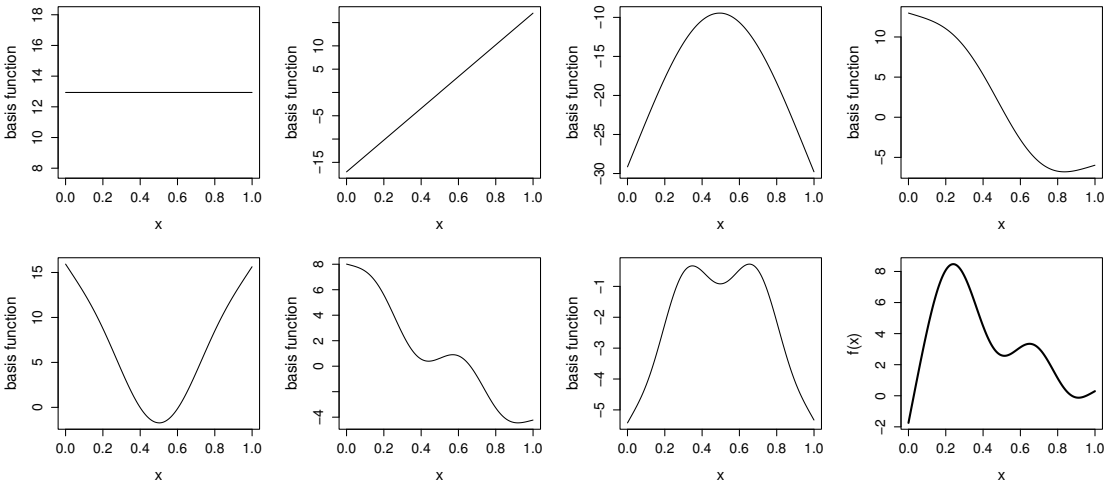

Figure 2-1: This plot (source: Wood (2006)) illustrates a rank 7 thin plate regression spline basis for representing a smooth function of one variable. The first 7 panels (starting at top left) show the basis functions multiplied by some coefficients. These are then summed to give the smooth curve in the lower right panel. The first two bases span the space of functions that are completely smooth, according to the roughness measure defined in Section 2.3. The remaining basis functions represent the wiggly component of the smooth curve.

2.3 Some Model Fitting Details

Given a vector of n independent observations, where $Y_i \sim m_{\vartheta_i}(y_i)$, the substitution of the terms $f_j(x_j)$ with their regression spline expression into a model equation like (2.1) yields a GLM, which can be estimated by maximum likelihood. Specifically, η_i can be rewritten as $\mathbf{X}_i\boldsymbol{\beta}$, where $\mathbf{X}_i$ includes $\mathbf{X}_i^*$ and the terms representing the spline bases for the f_j, while $\boldsymbol{\beta}$ contains $\boldsymbol{\theta}$ and all the smooth coefficient vectors, $\boldsymbol{\beta}_j$. $m_{\vartheta_i}(y_i)$ denotes an exponential family distribution with probability density function (2.2) for which h and c are fixed and depend on the chosen distribution. The natural parameter ϑ_i is determined by

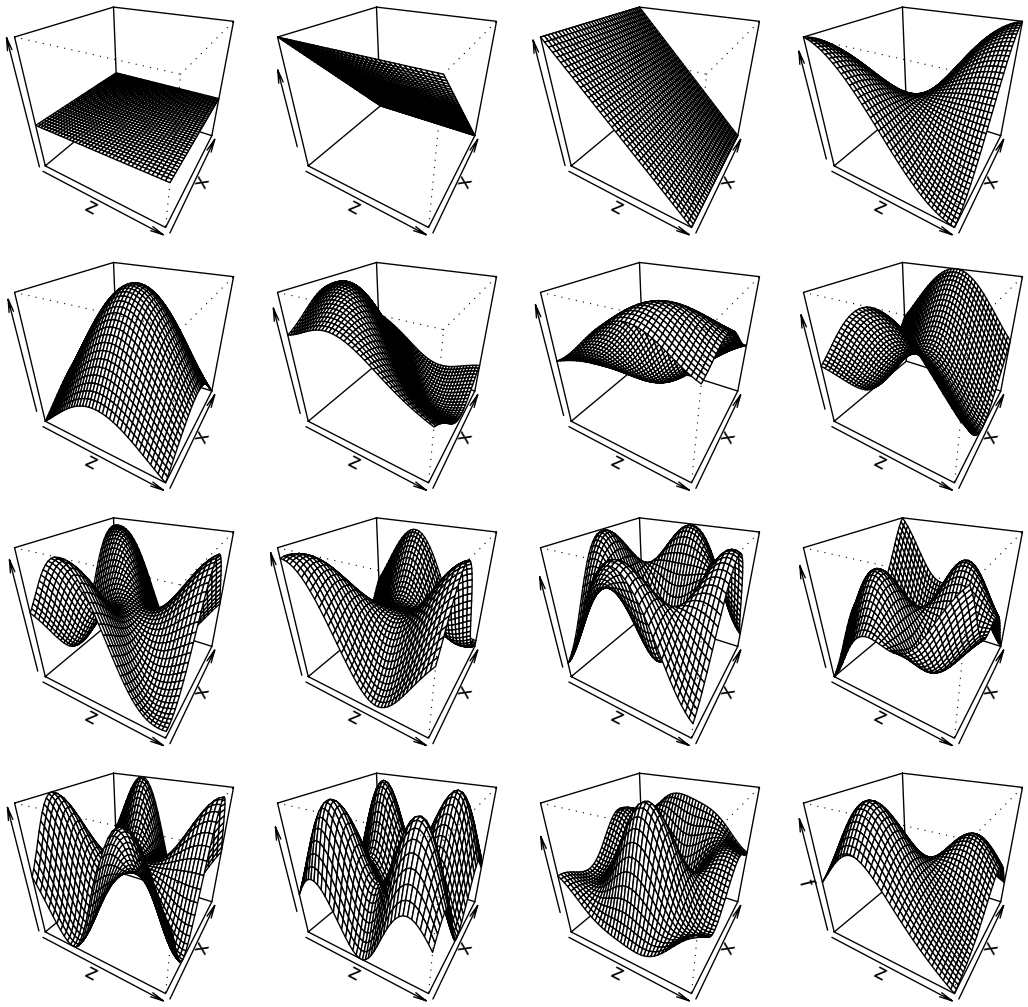

Figure 2-2: This plot (source: Wood (2006)) illustrates a rank 15 thin plate regression spline basis for representing a smooth function of two variables. The first 15 panels (starting at top left) show the basis functions multiplied by some coefficients. These are then summed to give the smooth surface in the lower right panel. The first three bases span the space of functions that are completely smooth, according to the roughness measure defined in Section 2.3. The remaining basis functions represent the wiggly component of the smooth curve.

μ_i via $\mathbb{E}(Y_i)$ and hence ultimately by $\boldsymbol{\beta}$. The dispersion parameter ϕ can either be fixed or estimated, depending on the chosen distribution. For example, for the binomial and Poisson cases, ϕ is known and equal to 1.

Since the Y_i are assumed to be independent, the likelihood of $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} m_{\vartheta_i}(y_i)$$

and its log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{y_i \vartheta_i - h(\vartheta_i)}{\phi} + c(y_i, \phi) \right\},$$

where $\boldsymbol{\beta}$ enters the right-hand side through the ϑ_i. Log-likelihood maximization is achieved by partially differentiating l with respect to each element of $\boldsymbol{\beta}$ setting the resulting equations to zero, and solving for $\boldsymbol{\beta}$. In formulae, the maximum likelihood estimate of $\boldsymbol{\beta}$ satisfies the score equations

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \{ y_i \vartheta_i - h(\vartheta_i) \} = \frac{1}{\phi} \sum_{i=1}^{n} \left\{ \frac{(y_i - \mu_i)}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) x_{ji} \right\} = 0 \ \ \forall j,$$

whose solution does not depend on ϕ. These equations can not be solved algebraically, hence a numerical iterative procedure has to be employed. In practice, the likelihood can be maximized by Iteratively Re-Weighted Least Squares (IRLS), where the GLM is fitted by iterative minimization of the problem

$$\| \sqrt{\mathbf{W}^{[k]}} (\mathbf{z}^{[k]} - \mathbf{X}\boldsymbol{\beta}) \|^2 \ \text{ w.r.t. } \boldsymbol{\beta}.$$

k is the iteration index, $\mathbf{z}^{[k]} = \mathbf{X}\boldsymbol{\beta}^{[k]} + \mathbf{G}^{[k]}(\mathbf{y} - \boldsymbol{\mu}^{[k]})$, $\mu_i^{[k]}$ is the current model estimate of $\mathbb{E}(Y_i)$, $\mathbf{G}^{[k]}$ is a diagonal matrix such that $G_{ii}^{[k]} = g'(\mu_i^{[k]})$, and $\mathbf{W}^{[k]}$ is a diagonal matrix given by $W_{ii}^{[k]} = [G_{ii}^{[k]2} V(\mu_i^{[k]})]^{-1}$. To avoid overfitting it is necessary to fit the model by penalized maximum likelihood estimation in which roughness measures are used to control overfit. For the case of smooth functions of one variable, the penalized likelihood is maximized by penalized IRLS (P-IRLS), so that the GAM is fitted by iteratively minimizing the problem

$$\| \sqrt{\mathbf{W}^{[k]}} (\mathbf{z}^{[k]} - \mathbf{X}\boldsymbol{\beta}) \|^2 + \sum_j \lambda_j \int \left\{ f_j^{d_j}(x_j) \right\}^2 dx_j \ \text{ w.r.t. } \boldsymbol{\beta}.$$

The terms in the summation measure the roughness of the smooth functions, d_j (usually set to 2) indicates the order of the derivatives for the j^{th} smooth

term to be used in the fitting process, and the λ_j are smoothing parameters that control the trade-off between fit and smoothness. Since regression splines are linear in their model parameters, the penalty $\sum_j \lambda_j \int \left\{ f_j^{d_j}(x_j) \right\}^2 dx_j$ can be written as a quadratic form in $\boldsymbol{\beta}$ with known coefficient matrices $\mathbf{S}_j$. As an example, by setting $d_j = 2$ and for a regression spline basis in one dimension, we have that

$$\begin{aligned}
\int \left\{ f_j^2(x_j) \right\}^2 dx_j &= \int \left\{ \frac{\partial^2 f_j(x_j)}{\partial x_j^2} \right\}^2 dx_j = \int \left\{ \frac{\partial^2 \sum_{k=1}^{q_j} \beta_{jk} b_{jk}(x_j)}{\partial x_j^2} \right\}^2 dx_j \\
&= \int \left\{ \boldsymbol{\beta}^\mathsf{T} \mathbf{b}_j''(x_j) \right\}^2 dx_j = \int \boldsymbol{\beta}^\mathsf{T} \mathbf{b}_j''(x_j) \mathbf{b}_j''(x_j)^\mathsf{T} \boldsymbol{\beta} dx_j \\
&= \boldsymbol{\beta}^\mathsf{T} \left\{ \int \mathbf{b}_j''(x_j) \mathbf{b}_j''(x_j)^\mathsf{T} dx_j \right\} \boldsymbol{\beta} = \boldsymbol{\beta}^\mathsf{T} \mathbf{S}_j \boldsymbol{\beta},
\end{aligned}$$

where $\mathbf{b}_j''(x_j)$ is a vector containing the second derivatives of the basis functions for the j^{th} smooth term with respect to x_j. It follows that

$$\sum_j \lambda_j \int \left\{ f_j^{d_j}(x_j) \right\}^2 dx_j = \sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T} \mathbf{S}_j \boldsymbol{\beta}.$$

The precise mathematical expression of a thin regression spline basis and its penalty depends on the value of d_j and the dimension of x_j; see Wood (2003, 2006) for full mathematical details. The smoothing parameters play a crucial role in penalized regression spline estimation: very large values for λ_j lead to very smooth estimates and vice versa. Given smoothing parameters, the penalized nonlinear least squares problem can be solved by using the IRLS algorithm. It turns out that the form of the parameter estimators of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T} \mathbf{W} \mathbf{X} + \mathbf{S})^{-1} \mathbf{X}^\mathsf{T} \mathbf{W} \mathbf{z},$$

where $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. It follows that the estimator for $\boldsymbol{\beta}$ is biased because of penalty-induced bias.

Smoothing parameter estimation has to be addressed as well. This can be achieved by minimization of a prediction error estimate, such as the generalized cross validation (GCV) score, if a dispersion parameter has to be estimated, or the generalized Akaike's information criterion (AIC). Following Wood (2008), smoothing parameter selection via the GCV score consists of minimizing

$$V_g(\boldsymbol{\lambda}) = \frac{n D(\hat{\boldsymbol{\beta}})}{\{n - \mathrm{tr}(\mathbf{A})\}^2},$$

where $D(\hat{\boldsymbol{\beta}})$, the model deviance, is defined as $2\phi(\hat{l}_{\text{sat}} - \hat{l}(\hat{\boldsymbol{\beta}}))$, $\hat{l}(\hat{\boldsymbol{\beta}})$ is the log-likelihood of the fitted model and $\hat{l}_{\text{sat}}$ the maximum value for the log-likelihood of the model with one parameter per datum. The matrix $\mathbf{A}$ is given by $\mathbf{W}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}$, and the λ_j enter the GCV score through $\mathbf{A}$. In case ϕ is known, the following generalized AIC is minimized instead

$$V_a(\boldsymbol{\lambda}) = D(\hat{\boldsymbol{\beta}}) + 2\text{tr}(\mathbf{A})\phi.$$

As an alternative, REML can be employed. Within this framework, the penalized likelihood estimates, $\hat{\boldsymbol{\beta}}$, can be seen as the posterior modes of the distribution of $\boldsymbol{\beta}|\mathbf{y}$ if $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}^{-}\phi)$, where $\mathbf{S}^{-}$ is an appropriate generalized inverse. Viewing the spline parameters as random effects allows for the possibility to estimate the λ_i via REML (Wahba, 1985). Wahba (1985) showed that asymptotically prediction error criteria are better in a mean square error sense, even though Härdle *et al.* (1988) pointed out that these criteria give slow convergence to the optimal smoothing parameters. The recent work by Reiss and Ogden (2009) shows that at finite sample sizes GCV or AIC is prone to undersmoothing and is more likely to develop multiple minima than REML (e.g. Wood, 2010). So, it would appear that REML should be preferred over GCV/AIC especially when the primary purpose of the analysis is to carry out smooth component selection. The computational methods for automatic smoothing parameter estimation of Wood (2006, 2008, 2010) are based on the criteria mentioned above, and will be used throughout this thesis.

2.4 Confidence Intervals

The well known Bayesian 'confidence' intervals originally proposed by Wahba (1983) or Silverman (1985) in the univariate spline model context, and then generalized to the component-wise case when dealing with GAMs (e.g. Gu, 1992; Gu, 2002; Gu and Wahba, 1993; Wood, 2006), are typically used to reliably represent the uncertainty of smooth terms. This is because such intervals include both a bias and variance component (Nychka, 1988), a fact that makes these intervals have good observed *frequentist* coverage probabilities across the function.

The large sample posterior used for interval calculations is given by

$$\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda}, \phi, \sim N(\hat{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}}), \tag{2.5}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum penalized likelihood estimate of $\boldsymbol{\beta}$, $\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\phi$, and $\mathbf{W}$ and $\mathbf{z}$ are the diagonal weight matrix and the pseudodata vector at convergence of the P-IRLS algorithm. Notice that $\mathbf{W}$ and ϕ are equal to the identity matrix and σ^2, respectively, when a normal response is assumed and an identity link function selected; furthermore, in the normal errors case, the result above holds independently of asymptotic arguments.

In Chapter 4, we study the coverage properties of these intervals. Specifically, we present simulation evidence showing these intervals have close to nominal across-the-function frequentist coverage probabilities, and extend Nychka's (1988) argument for univariate smoothing splines to the GAM component case to explain these results.

2.5 Testing for No Effect

In order to achieve component selection, a number of hypothesis testing approaches have been proposed in the literature, each of them with advantages and disadvantages. Here, we follow the approach by Wood (2006).

Asymptotic arguments for maximum likelihood estimators suggest that if a model is correctly specified, then in the large sample limit

$$\hat{\boldsymbol{\beta}} \backsim N\left(\mathbb{E}(\hat{\boldsymbol{\beta}}), \mathbf{V}_{\hat{\boldsymbol{\beta}}}\right),$$

where $\mathbb{E}(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$ because of penalty-induced bias, and the frequentist covariance matrix is given by

$$\begin{aligned}
\mathbf{V}(\hat{\boldsymbol{\beta}}) &= \mathbf{V}\left((\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{z}\right) \\
&= (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{V}(\mathbf{z})\mathbf{W}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1} \\
&= \mathbf{B}\mathbf{V}\left(\boldsymbol{\eta} + \mathbf{G}(\mathbf{y} - \boldsymbol{\mu})\right)\mathbf{B}^{\mathsf{T}} = \mathbf{B}\mathbf{V}(\mathbf{G}\mathbf{y})\mathbf{B}^{\mathsf{T}} = \mathbf{B}\mathbf{G}\mathbf{V}(\mathbf{y})\mathbf{G}^{\mathsf{T}}\mathbf{B}^{\mathsf{T}},
\end{aligned}$$

where $\mathbf{B} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}$, and $\mathbf{V}(\mathbf{y})$ is a diagonal matrix with elements $V_{ii} = V(\mu_i)\phi$. Recalling that $\mathbf{G}$ is a diagonal matrix such that $G_{ii} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)$, it follows that $\mathbf{G}\mathbf{V}(\mathbf{y})\mathbf{G}^{\mathsf{T}} = \mathbf{W}^{-1}\phi$ and therefore that

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\phi.$$

The dispersion parameter ϕ can be estimated by the Pearson estimator $\hat{\phi} = \|\sqrt{\mathbf{W}}(\mathbf{y} - \hat{\boldsymbol{\mu}})\|^2 / \{n - \mathrm{tr}(\mathbf{A})\}$, where $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$. The trace of $\mathbf{A}$ represents the estimated degrees of freedom (edf) of the fitted model. It is also known as

degree of complexity or number of effective parameters in the model. The edf of the model is given by the sum of the edf of the smooth functions. For the case of a smooth function of one covariate with $d_j = 2$, if $\text{edf}_j = 1$ then it means that the explanatory variable x_j can enter the model linearly.

The parameter estimators of $\boldsymbol{\theta}$ are unpenalized. This means that classic distributional results for GLMs can be used for parametric terms. In particular, hypothesis testing and confidence intervals can be based on the Gaussian or an appropriate t distribution, depending on whether ϕ is known. As for the penalized regression spline parameter estimators, given that $\mathbb{E}(\hat{\boldsymbol{\beta}}_j) \neq \boldsymbol{\beta}_j$, the usual distributional results for GLMs can not be employed for hypothesis testing. However, when the goal of the analysis is testing that a smooth term of a GAM is equal to zero, we have that if $\boldsymbol{\beta}_j = \mathbf{0}$ then $\mathbb{E}(\hat{\boldsymbol{\beta}}_j) \approx \mathbf{0}$ (Wood, 2006). It follows that, under the null hypothesis that the coefficients of a smooth component are zero,

$$\hat{\boldsymbol{\beta}}_j^{\mathsf{T}} \mathbf{V}_{\hat{\boldsymbol{\beta}}_j}^{r-} \hat{\boldsymbol{\beta}}_j \dot{\sim} \chi_r^2,$$

where r denotes the rank of the covariance matrix of $\hat{\boldsymbol{\beta}}_j$, and $\mathbf{V}_{\hat{\boldsymbol{\beta}}_j}^{r-}$ is the rank r generalized pseudoinverse of $\mathbf{V}_{\hat{\boldsymbol{\beta}}_j}$ that has to be employed to overcome possible matrix rank deficiencies deriving from the fact that the smoothing penalty may suppress some dimensions of the parameter space. r is determined heuristically as follows. It is the minimum value between the maximum edf value allowed for the j^{th} smooth term (which is also the number of basis functions used for the term) and the smallest integer not less than the quantity calculated as $2 * \text{edf}_j$ (Wood, 2006). If ϕ is unknown, then the null hypothesis can be tested using the following result

$$\frac{\hat{\boldsymbol{\beta}}_j^{\mathsf{T}} \mathbf{V}_{\hat{\boldsymbol{\beta}}_j}^{r-} \hat{\boldsymbol{\beta}}_j / r}{\hat{\phi}/\phi} = \hat{\boldsymbol{\beta}}_j^{\mathsf{T}} \hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_j}^{r-} \hat{\boldsymbol{\beta}}_j / r \dot{\sim} F_{r,n-\text{edf}},$$

since $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_j}$ is based on $\hat{\phi}$. As pointed out in Wood (2006), these two p-value definitions are only approximate and one has to be careful when using these results for variable selection purposes.

Despite the fact that some testing methods have been introduced in the GAM context, such as the one discussed in this section, a general reliable distributional theory for the smooth terms of a GAM has not been developed to date. In Chapter 5, we tackle this problem and propose three practical methods to achieve GAM component selection.

2.6 Model Comparison

In any model building process, the researcher might be interested in comparing two nested models. Nesting implies that the simpler model (H_0) is a special case of the more complex model (H_1). For example, the explanatory variables present in H_0 are a subset of those present in H_1. In such cases, the generalized likelihood ratio test is often applied. Specifically, consider testing

$$H_0 : \mathbf{X}_0\boldsymbol{\beta}_0 \text{ against } H_1 : \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}_0 \subset \mathbf{X}$. If H_0 is true then in the large sample limit

$$D_0 - D_1 \backsim \chi^2_{\text{edf}_1 - \text{edf}_0},$$

where D_0 and D_1 are the deviances under H_0 and H_1, respectively. When a dispersion parameter has to be estimated the F-ratio test may be used

$$F = \frac{(D_0 - D_1)/(\text{edf}_1 - \text{edf}_0)}{D_1/(n - \text{edf}_1)} \dot{\backsim} F_{\text{edf}_1 - \text{edf}_0, n - \text{edf}_1},$$

which does not require knowledge of ϕ (Wood, 2006).

2.7 Model Checking

Before interpreting model results and for model comparison purposes, the adequacy of the fitted models has to be checked. This is perhaps the most important part of any statistical analysis since unexplained systematic structure in the residuals of a fitted model typically leads to misleading inferences.

Model checking for GAMs, which is similar to what is done for linear models and GLMs, can be performed using Pearson or deviance residuals (McCullagh and Nelder, 1989). The Pearson residual has the form

$$r_i^{(P)} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{var}}(Y_i)}},$$

where $\hat{\mu}_i$ and $\hat{\text{var}}(Y_i)$ are the fitted mean and variance for the i^{th} observation in the dataset. Under the assumption that the fitted model is correct, the Pearson residuals have approximately zero mean and standard deviation close to 1. However, as discussed in Cameron and Trivedi (1998), these residuals are generally asymmetrically distributed. As an alternative, one can use deviance

residuals since they can suggest which observations cause lack of fit, and are expected to behave something like standard normal random variables. The deviance residual is defined as

$$r_i^{(d)} = \text{sign}(Y_i - \hat{\mu}_i)\sqrt{c_i},$$

where c_i represents the contribution of the i^{th} observation to the overall goodness of fit of the model as D can be seen as $\sum_{i=1}^{n} c_i$.

Chapter 3

A Flexible Instrumental Variable Approach

Classical regression model literature has sometimes assumed that measured and unmeasured or unobservable covariates are statistically independent. For many applications this assumption is clearly tenuous. When unobservables are associated with included regressors and have an impact on the response, standard estimation methods will not be valid. This means, for example, that estimation results from observational studies, whose aim is to evaluate the impact of a treatment of interest on a response variable, will be biased and inconsistent in the presence of unmeasured confounders. One method for obtaining consistent estimates of treatment effects when dealing with linear models is the instrumental variable (IV) approach. Linear models have been extended to GLMs and GAMs, and although IV methods have been proposed to deal with GLMs, fitting methods to carry out IV analysis within the GAM context have not been developed. We propose a two-stage procedure for IV estimation when dealing with GAMs represented using any penalized regression spline approach, and a correction procedure for confidence intervals. We explain under which conditions the proposed method works and illustrate its empirical validity through an extensive simulation experiment and a health study where unmeasured confounding is suspected to be present.

3.1 Introduction

Observational data are often used in statistical analysis to infer the effects of one or more predictors of interest (which can be also referred to as treatments) on a response variable. The main characteristic of observational studies is a

lack of treatment randomization which usually leads to selection bias. In a regression context, the most common solution to this problem is to account for confounding variables that are associated with both treatments and response (e.g. Becher, 1992). However, the researcher might fail to adjust for pertinent confounders as they might be either unknown or not readily quantifiable. This constitutes a serious limitation to covariate adjustment since the use of standard estimators typically yields biased and inconsistent estimates. Hence, a major concern when estimating treatment effects is how to account for unmeasured confounders.

This problem is known in econometrics as *endogeneity* of the predictors of interest. The most commonly used econometric method to model data that are affected by the unobservable confounding issue is the instrumental variable (IV) approach (Wooldridge, 2002). This technique only recently has received some attention in the applied statistical literature. This method can yield consistent parameter estimates and can be used in any kind of analysis in which unmeasured confounding is suspected to be present (e.g. Beck *et al.*, 2003; Leigh and Schembri, 2004; Linden and Adams, 2006; Wooldridge, 2002). The IV approach can be thought of as a means to achieve pseudo randomization in observational studies (Frosini, 2006). It relies on the existence of one or more IVs that induce substantial variation in the endogenous/treatment variables, are independent of unobservables, and are independent of the response conditional on all measured and unmeasured confounders. Provided that such variables are available, IV regression analysis can split the variation in the endogenous predictors into two parts, one of which is associated with the unmeasured confounders (Wooldridge, 2002). This fact can then be used to obtain consistent estimates of the effects of the variables of interest.

The applied and theoretical literature on the use of IVs in parametric and nonparametric regression models with Gaussian response is large and well understood (Ai and Chen, 2003; Das, 2005; Hall and Horowitz, 2005; Newey and Powell, 2003). In many applications, however, Gaussian regression models have been replaced by GLMs and GAMs (McCullagh and Nelder, 1989; Hastie and Tibshirani, 1990), as they allow researchers to model data using the response variable distribution which best fits the features of the outcome of interest, and to make use of nonparametric smoothers since the functional shape of any relationship is rarely known *a priori*. Simultaneous maximum-likelihood estimation methods for GLMs in which selection bias is suspected to be present have been proposed. Here consistent and efficient estimates can be obtained by jointly modelling the distribution of the response and the endoge-

nous variables (Heckman, 1978; Maddala, 1983; Wooldridge, 2002). However, the main drawbacks are typically computational cost and the derivation of the joint distribution, issues that may become more severe in the GAM context. Amemiya (1974) proposed an IV generalized method of moments (GMM) approach to consistently estimate the parameters of a GLM. An epidemiological example is provided by Johnston *et al.* (2008). Here it is not clear how such an approach can be implemented for GAMs so that reliable smooth component estimates can be obtained in practice. This is because when fitting a GAM the amount of smoothing for the smooth components in the model has to be selected with a certain degree of precision. In this respect, it might be difficult to develop a reliable computational multiple smoothing parameter method by taking an IV GMM approach, and, to the best of our knowledge, such a procedure has not been developed to date.

The IV extension to the GAM context is a topic under construction. This generalization is important because even if we use an IV approach to account for unmeasured confounders, we can still obtain biased estimates if the functional relationship between predictors and outcome is not modelled flexibly. The aim of this chapter is to extend the IV approach to GAMs by exploiting the two-stage procedure idea first proposed by Hausman (1978, 1983) and employing one of the reliable smoothing approaches available in the GAM literature. To simplify matters, we first discuss a two-step estimator for GLMs which can be then easily extended to GAMs. The proposed approach can be efficiently implemented using some standard existing software. Our proposal is illustrated through an extensive simulation study and in the context of a health study.

The rest of the chapter is structured as follows. Section 3.2 discusses the IV properties, the classical two-stage least squares (2SLS) method, and the Hausman's endogeneity testing approach. For simplicity of exposition, Section 3.3 illustrates the main ideas using GLMs, which are then extended to the GAM context in Section 3.4. Section 3.5 proposes a confidence interval correction procedure for the two-stage approach of Section 3.4. Section 3.6 evaluates the empirical properties of the two-step GAM estimator through an extensive simulation experiment, whereas Section 3.7 illustrates the method via a health observational study of medical care utilization where unmeasured confounding is suspected to be present.

3.2 Preliminaries and motivation

In empirical studies, endogeneity typically arises in three ways: omitted variables, measurement error, and simultaneity (see Wooldridge (2002, p. 50) for more details on these forms of endogeneity). Here, we approach the problem of endogenous explanatory variables from an omitted variables perspective.

To fix ideas, let us consider the model

$$Y = \beta_0 + \beta_e X_e + \beta_o X_o + \beta_u X_u + \epsilon_Y, \;\; \mathbb{E}(\epsilon_Y | X_e, X_o, X_u) = 0, \tag{3.1}$$

where ϵ_Y is an error term normally distributed with mean 0 and constant variance, β_0 represents the intercept of the model, and X_e, X_o and X_u are the endogenous variable, observable confounder, and unmeasured confounder, with parameters β_e, β_o and β_u, respectively. We assume that X_u influences the response variable Y and is associated with X_e.

Since X_u can not be observed, (3.1) can be written as

$$Y = \beta_0 + \beta_e X_e + \beta_o X_o + \zeta, \tag{3.2}$$

where $\zeta = \beta_u X_u + \epsilon_Y$. OLS estimation of equation (3.2) results in inconsistent estimators of all the parameters, with β_e generally the most affected. In order to obtain consistent parameter estimates, an IV approach can be employed. Specifically, to *clear up* the endogeneity of X_e, we need an observable variable X_{IV}, called instrument or IV, that satisfies three conditions (e.g. Greenland, 2000):

1. The first requirement can be better understood by making use of the following model

   $$X_e = \alpha_0 + \alpha_o X_o + \alpha_{IV} X_{IV} + \alpha_u X_u + \epsilon_{X_e}, \tag{3.3}$$

   where ϵ_{X_e} has the same features as ϵ_Y. (3.3) can also be written as

   $$X_e = \alpha_0 + \alpha_o X_o + \alpha_{IV} X_{IV} + \xi_u, \;\; \mathbb{E}(\xi_u | X_o, X_{IV}) = 0,$$

   where ξ_u, defined as $\alpha_u X_u + \epsilon_{X_e}$, is assumed to be uncorrelated with X_o and X_{IV}, and α_{IV} must be significantly different from 0. In other words, X_{IV} must be associated with X_e conditional on the remaining covariates in the model.

2. The second requirement is that X_{IV} is independent of Y conditional on

the other regressors in the model and X_u.

3. The third condition requires X_{IV} to be independent of X_u.

As an example, let us consider the study by Leigh and Schembri (2004). The aim of their analysis was to estimate the effect of smoking on physical functional status. Smoking was considered as an endogenous variable since it was assumed to be associated with health risk factors which could not be observed. The IV was cigarette price as it was believed to be logically and statistically associated with smoking, and not to be directly related to any individual's health. Also, it was logically assumed to be unrelated to those unmeasured health risk confounders which could affect physical functional status. Cigarette price therefore appeared to satisfy the conditions for a valid and strong instrument. In many situations identification of a valid instrument is less clear than in the case above, and is usually heavily dependent on the specific problem at hand. This is because some of the necessary assumptions can not be verified empirically, hence the selection of an instrument has to be based on subject-matter knowledge, not statistical testing. Assuming that an appropriate instrument can be found, several methods can be employed to correctly quantify the impact that a predictor of interest has on the response variable, 2SLS being the most common.

In 2SLS estimation, least squares regression is applied twice. Specifically, the first stage involves fitting a linear regression of X_e on X_o and X_{IV} to obtain $\hat{\mathbb{E}}(X_e|X_o, X_{IV})$ or $\hat{X}_e$. In the second stage, a regression of Y on $\hat{X}_e$ and X_o is performed. We see why this procedure yields consistent estimates of the parameters by taking the conditional expectation of (3.2) given X_o and X_{IV}. That is,

$$\mathbb{E}(Y|X_o, X_{IV}) = \beta_0 + \beta_e\hat{X}_e + \beta_o X_o.$$

Thus, the 2SLS estimator can produce an estimate of the original parameter of interest. However, this approach does not yield consistent estimates of the coefficients when dealing with generalized models (Amemiya, 1974). This is because the unobservable is not additively separable from the systematic part of the model. The following argument better explains this point. 2SLS implies the replacement of $\beta_e X_e$ with $\beta_e(\hat{X}_e + \hat{\xi}_u)$. Thus, the error of model (3.2) is allowed to become $(\beta_e\hat{\xi}_u + \beta_u X_u + \epsilon_Y)$, which can be readily shown to be uncorrelated with $\hat{X}_e$ and X_o. This result does not hold for GLMs because $\beta_e\hat{\xi}_u$ and $\beta_u X_u$ can not become part of the error term given the presence of a link function that has to be employed when dealing with GLMs.

The developments of the next two sections are based on the two-stage approach introduced by Hausman (1978, 1983) as a means of directly testing the endogeneity hypothesis for the class of linear models. His procedure has the same first stage as 2SLS, but in the second stage X_e is not replaced by $\hat{X}_e$. Instead, the first-stage residual is included as an additional predictor in the second-stage regression, and its parameter significance tested. 2SLS and the Hausman's procedure are equivalent for Gaussian models in terms of estimated parameters. However, they do not yield the same results when dealing with generalized models since 2SLS would produce biased and inconsistent estimates (for the reasons given in the previous paragraph) whereas a Hausman-like approach would consistently estimate the parameters of interest, as it will be discussed in the next section.

3.3 IV estimation for GLMs

The purpose of this section is to discuss a two-step IV estimator for GLMs which can be then easily extended to GAMs. As explained in Section 3.1, several valid methods have already been proposed to deal with GLMs in which selection bias is suspected to be present. In fact, our aim is not to discuss an alternative IV approach for GLMs, but to illustrate the main ideas using this simpler class of models. The generalization to the GAM context will then easily follow.

A GLM has the model structure

$$\mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{3.4}$$

where $\mathbf{g}(\cdot)$ is a smooth monotonic link function, $\boldsymbol{\mu} \equiv \mathbb{E}(\mathbf{y}|\mathbf{X})$, $\mathbf{y}$ is a vector of independent response variables $(Y_1, ..., Y_n)^\mathsf{T}$, $\boldsymbol{\eta}$ is called the linear predictor, $\mathbf{X}$ is an $n \times k$ matrix of k covariates, and $\boldsymbol{\beta}$ represents the $k \times 1$ vector of unknown regression coefficients. The generic response variable Y follows an exponential family distribution.

Model (3.4) can also be written as

$$\mathbf{y} = \mathbf{g}^{-1}(\boldsymbol{\eta}) + \boldsymbol{\epsilon}, \;\; \mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0}, \tag{3.5}$$

where $\mathbf{g}^{-1}(\boldsymbol{\eta}) = \mathbb{E}(\mathbf{y}|\mathbf{X})$, and $\boldsymbol{\epsilon}$ is an additive, unobservable error trivially defined as $\boldsymbol{\epsilon} \equiv \mathbf{y} - \mathbf{g}^{-1}(\boldsymbol{\eta})$. Recall that equation (3.5) only implies that $\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0}$. Certainly, depending on the nature of Y, the error term may have some un-

desired properties. As explained in Section 3.2, we assume three types of covariates. That is, $\mathbf{X} = (\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_u)$, where $\mathbf{X}_e$ is an $n \times h$ matrix of endogenous variables, $\mathbf{X}_o$ an $n \times j$ matrix of observable confounders, and $\mathbf{X}_u$ an $n \times h$ matrix of unmeasured confounders that influence the response variable and are associated with the endogenous predictors. Correspondingly, $\boldsymbol{\beta}^\mathsf{T}$ can be written as $(\boldsymbol{\beta}_e^\mathsf{T}, \boldsymbol{\beta}_o^\mathsf{T}, \boldsymbol{\beta}_u^\mathsf{T})$. Notice that, as, e.g., in Terza *et al.* (2008), we assume we have as many endogenous variables as there are unobservables. To simplify notation we do not write the intercept vector in $\mathbf{X}$ even though we assume it is included. If $\mathbf{X}_u$ is available, then $\hat{\boldsymbol{\beta}}$ can yield consistent estimates of $\boldsymbol{\beta}$.

The problem with equation (3.5) is that we can not observe $\mathbf{X}_u$, hence it can not be included in the model. This violates the assumption that $\mathbb{E}(\mathbf{X}^\mathsf{T}\boldsymbol{\epsilon}) = \mathbf{0}$, therefore leading to biased and inconsistent estimates. To this end, it is useful to model the variables in $\mathbf{X}_e$ through the following set of auxiliary (or reduced form) equations (e.g. Terza *et al.*, 2008)

$$\mathbf{x}_{ep} = \mathbf{g}_p^{-1}(\mathbf{Z}_p\boldsymbol{\alpha}_p) + \boldsymbol{\xi}_{up}, \quad p = 1, \ldots, h, \tag{3.6}$$

where $\mathbf{x}_{ep}$ represents the p^{th} column vector from $\mathbf{X}_e$, $\mathbf{g}_p^{-1}$ is the inverse of the link function chosen for the p^{th} endogenous/treatment variable, $\mathbf{Z}_p = (\mathbf{X}_o, \mathbf{X}_{IVp})$, $\mathbf{X}_{IVp}$ is the p^{th} matrix of dimension $n \times n.iv_p$ where $n.iv_p$ indicates the number of identifying instrumental variables available for $\mathbf{x}_{ep}$, $\boldsymbol{\alpha}_p$ denotes the $(j+n.iv_p)\times 1$ vector of unknown parameters, and $\boldsymbol{\xi}_{up}$ is a term containing information about both structured and unstructured terms. It is well known in the IV literature that, in order to *identify* the set of reduced form equations, there must be at least as many instruments as there are endogenous regressors. This means that each $n.iv_p$ must be equal or greater than 1. This will be assumed to be the case throughout the chapter.

The reason why the equations in (3.6) can be used to "correct" the parameter estimates of the equation of interest is as follows. Once the measured confounders have been accounted for and provided the instruments meet the conditions discussed in Section 3.2, the $\boldsymbol{\xi}_{up}$ contain information about the unmeasured confounders that can be used to obtain corrected parameter estimates of the endogenous variables. To shed light on this last point, using an argument similar to that of Johnston *et al.* (2008), let us assume that the true model underlying the p^{th} reduced form equation is

$$\mathbf{x}_{ep} = \mathbb{E}(\mathbf{x}_{ep}|\mathbf{Z}_p, \mathbf{x}_u) + \boldsymbol{\upsilon}_p, \tag{3.7}$$

where $\mathbb{E}(\mathbf{x}_{ep}|\mathbf{Z}_p, \mathbf{x}_u) = \mathbf{h}_p(\mathbf{Z}_p\boldsymbol{\alpha}_p + \mathbf{x}_u)$, $\mathbf{h}_p = \mathbf{g}_p^{-1}$, and $\boldsymbol{\upsilon}_p$ is an error term. Now, $\mathbf{h}_p(\cdot)$ can be replaced by the Taylor approximation of order 1

$$\mathbf{h}_p(\mathbf{Z}_p\boldsymbol{\alpha}_p + \mathbf{x}_u) \approx \mathbf{h}_p(\mathbf{Z}_p\boldsymbol{\alpha}_p) + \mathbf{x}_u\mathbf{h}_p'(\mathbf{Z}_p\boldsymbol{\alpha}_p), \tag{3.8}$$

hence (3.7) can be written as

$$\mathbf{x}_{ep} = \mathbf{h}_p(\mathbf{Z}_p\boldsymbol{\alpha}_p) + \mathbf{x}_u\mathbf{h}_p'(\mathbf{Z}_p\boldsymbol{\alpha}_p) + \boldsymbol{\upsilon}_p,$$

which in turn leads to model (3.6) where

$$\boldsymbol{\xi}_{up} = \mathbf{x}_u\mathbf{h}_p'(\mathbf{Z}_p\boldsymbol{\alpha}_p) + \boldsymbol{\upsilon}_p.$$

The next section shows how the fact that the $\boldsymbol{\xi}_{up}$ contain information about the unobservables can be used to clear up the endogeneity of the treatment variables in the model. Notice that in the Gaussian case, approximation (3.8) is not needed since $\mathbf{x}_u$ would enter the error term linearly.

3.3.1 The two-step GLM estimator

In order to obtain consistent estimates for model (3.5) in the context defined earlier, we employ a Hausman-like approach. Specifically, the following two-step generalized linear model (2SGLM) procedure can estimate the parameters of interest consistently:

1. For each endogenous predictor in the model, obtain consistent estimates of $\boldsymbol{\alpha}_p$ by fitting the corresponding auxiliary equation through a GLM method. Then, calculate the following set of quantities

$$\hat{\boldsymbol{\xi}}_{up} = \mathbf{x}_{ep} - \mathbf{g}_p^{-1}(\mathbf{Z}_p\hat{\boldsymbol{\alpha}}_p), \;\; p = 1, \ldots, h. \tag{3.9}$$

2. Fit a GLM defined by

$$\mathbf{y} = \mathbf{g}^{-1}(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \hat{\boldsymbol{\Xi}}_u\boldsymbol{\beta}_{\boldsymbol{\Xi}_u}) + \boldsymbol{\varsigma}, \;\; \mathbb{E}(\boldsymbol{\varsigma}|\mathbf{X}) = \mathbf{0}, \tag{3.10}$$

   where $\hat{\boldsymbol{\Xi}}_u$ is an $n \times h$ matrix containing the $\hat{\boldsymbol{\xi}}_{up}$ obtained in the previous step, with parameter vector $\boldsymbol{\beta}_{\boldsymbol{\Xi}_u}$, and $\boldsymbol{\varsigma}$ represents an error term.

The parameter vector $\boldsymbol{\beta}_{\boldsymbol{\Xi}_u}$ can not be used as a means to explain the impact that the unmeasured confounders have on the outcome (see Section 3.4.1 for an explanation). However, this is not problematic since we are not interested

in $\boldsymbol{\beta}_{\boldsymbol{\Xi}_u}$. All that is needed is to account for the presence of unobservables, and this can be achieved by including a set of quantities which contain information about them. As a result, we can replace equation (3.5) by (3.10) since in this context $\boldsymbol{\beta}_e$ is the parameter vector of interest. It is important to stress that better empirical results are expected when the endogenous variables in the first step can be modelled using Gaussian regression models. In this case approximation (3.8) does not come into play which means that we can better control for unobservables.

Following, e.g., Terza *et al.* (2008), 2SGLM works since if the $\boldsymbol{\alpha}_p$ were known then by using (3.6) the column vectors of $\boldsymbol{\Xi}_u$ would be known. Hence information about the unobservables could be incorporated into the model by using $\boldsymbol{\Xi}_u$. In this respect, the endogeneity issue would disappear since the assumption that the error term is uncorrelated with the predictors would be satisfied. However, we do not know the $\boldsymbol{\alpha}_p$. By using (3.9) we can get consistent estimates for the $\boldsymbol{\alpha}_p$ thereby obtaining a good estimate for $\boldsymbol{\Xi}_u$. It can be readily shown that $\hat{\boldsymbol{\beta}}^\mathsf{T}$, now defined as $(\hat{\boldsymbol{\beta}}_e^\mathsf{T}, \hat{\boldsymbol{\beta}}_o^\mathsf{T}, \hat{\boldsymbol{\beta}}_{\boldsymbol{\Xi}_u}^\mathsf{T})$, is consistent for the vector value $\boldsymbol{\gamma}^\mathsf{T} = (\boldsymbol{\gamma}_e^\mathsf{T}, \boldsymbol{\gamma}_o^\mathsf{T}, \boldsymbol{\gamma}_u^\mathsf{T})$ that solves the *population* problem

$$\text{minimize} \quad \mathbb{E}[\|\mathbf{y} - \mathbf{g}^{-1}(\mathbf{X}_e\boldsymbol{\gamma}_e + \mathbf{X}_o\boldsymbol{\gamma}_o + \boldsymbol{\Xi}_u\boldsymbol{\gamma}_u)\|^2] \quad \text{w.r.t.} \quad \boldsymbol{\gamma}. \tag{3.11}$$

In (3.11) we ignore estimation for $\boldsymbol{\Xi}_u$ as the endogeneity issue only concerns the second-step equation, and because consistent estimates for it can be obtained. Provided the IVs meet the assumptions discussed in Section 3.2, we have that

$$\mathbb{E}(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_{IV}) = \mathbf{g}^{-1}(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \boldsymbol{\Xi}_u\boldsymbol{\beta}_{\boldsymbol{\Xi}_u}),$$

from which follows that $\boldsymbol{\beta} = \boldsymbol{\gamma}$. The sample analogue follows similar principles. These arguments are standard and can be found in Wooldridge (2002, pp. $341 - 345, 353 - 354$).

3.4 The GAM extension

The IV extension to the GAM context is important because even if we use an IV approach to account for unmeasured confounders, as shown in the previous section, we can still obtain biased estimates if the functional relationship between predictors and outcome is not modelled flexibly.

GAMs extend GLMs by allowing the determination of possible nonlinear

effects of predictors on the response variable. A GAM has the model structure

$$\mathbf{y} = \mathbf{g}^{-1}(\boldsymbol{\eta}) + \boldsymbol{\epsilon}, \;\; \mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0}. \tag{3.12}$$

Here, $\mathbf{X} = (\mathbf{X}^*, \mathbf{X}^+)$, $\mathbf{X}^* = (\mathbf{X}^*_e, \mathbf{X}^*_o, \mathbf{X}^*_u)$, and $\mathbf{X}^+ = (\mathbf{X}^+_e, \mathbf{X}^+_o, \mathbf{X}^+_u)$. The symbols $*$ and $+$ indicate whether the matrix considered refers to discrete predictors (such as dummy variables) or continuous regressors. Matrix dimensions can be defined following the same criterion adopted in the previous section. The linear predictor of a GAM is typically given by

$$\boldsymbol{\eta} = \mathbf{X}^*\boldsymbol{\beta}^* + \sum_j \mathbf{f}_j(\mathbf{x}^+_j), \tag{3.13}$$

where $\boldsymbol{\beta}^*$ represents the vector of unknown regression coefficients for $\mathbf{X}^*$, and the $\mathbf{f}_j$ are unknown smooth functions of the covariates, $\mathbf{x}^+_j$, represented using regression splines (see Chapter 2).

Since we can not observe $\mathbf{X}^*_u$ and $\mathbf{X}^+_u$, inconsistent estimates are expected. However, provided that IVs are available to correct for endogeneity, consistent estimates can be obtained by modelling the endogenous variables in the model. In the GAM context, this can be achieved through the following set of flexible auxiliary regressions

$$\mathbf{x}_{ep} = \mathbf{g}^{-1}_p\{\mathbf{Z}^*_p\boldsymbol{\alpha}^*_p + \sum_j \mathbf{f}_j(\mathbf{z}^+_{jp})\} + \boldsymbol{\xi}_{up}, \;\; p = 1, \ldots, h, \tag{3.14}$$

where $\mathbf{x}_{ep}$ represents either the p^{th} discrete or continuous endogenous predictor, $\mathbf{Z}^*_p = (\mathbf{X}^*_o, \mathbf{X}^*_{IVp})$ with corresponding vector of unknown parameters $\boldsymbol{\alpha}^*_p$, and $\mathbf{Z}^+_p = (\mathbf{X}^+_o, \mathbf{X}^+_{IVp})$.

3.4.1 The two-step GAM estimator

The 2SGLM estimator can now be extended to the GAM context. In particular, the following two-step generalized additive model (2SGAM) approach can be employed:

1. For each endogenous variable in the model, obtain consistent estimates of $\boldsymbol{\alpha}^*_p$ and the $\mathbf{f}_j$ by fitting the corresponding reduced form equation through a GAM method. Then, calculate the following set of quantities

$$\hat{\boldsymbol{\xi}}_{up} = \mathbf{x}_{ep} - \mathbf{g}^{-1}_p\{\mathbf{Z}^*_p\hat{\boldsymbol{\alpha}}^*_p + \sum_j \hat{\mathbf{f}}_j(\mathbf{z}^+_{jp})\}, \;\; p = 1, \ldots, h.$$

2. Fit a GAM defined by

$$\mathbf{y} = \mathbf{g}^{-1}\{\mathbf{X}^{*}_{eo}\boldsymbol{\beta}^{*}_{eo} + \sum_{j}\mathbf{f}_j(\mathbf{x}^{+}_{jeo}) + \sum_{p}\mathbf{f}_p(\hat{\boldsymbol{\xi}}_{up})\} + \boldsymbol{\varsigma}, \tag{3.15}$$

where $\mathbf{X}^{*}_{eo} = (\mathbf{X}^{*}_{e}, \mathbf{X}^{*}_{o})$ with parameter vector $\boldsymbol{\beta}^{*}_{eo}$, and $\mathbf{X}^{+}_{eo} = (\mathbf{X}^{+}_{e}, \mathbf{X}^{+}_{o})$.

In practice, the 2SGAM estimator can be implemented using GAMs represented via any penalized regression spline approach. For instance, the models in (3.14) and (3.15) can be fitted through penalized likelihood which can be maximized by P-IRLS (see Chapter 2).

As explained throughout the chapter, the presence of a relationship between the outcome and unobservables that are associated with endogenous predictors can lead to bias in the estimated impacts of the latter variables. The use of the $\mathbf{f}_p(\hat{\boldsymbol{\xi}}_{up})$ in (3.15) allows us to properly account for the impacts of unmeasured confounders on the response. This means that the linear/nonlinear effects of the endogenous regressors can be estimated consistently (see Section 3.6).

Let us now consider equation (3.14). The estimated residuals, $\hat{\boldsymbol{\xi}}_{up}$, will contain the linear/nonlinear impacts of the unobservables on the endogenous variables $\mathbf{x}_{ep}$. These effects can be partly or completely different from those that the same unmeasured confounders have on the outcome. However, this is not problematic; the $\mathbf{f}_p$ in (3.15) will automatically yield smooth functions estimates that (i) take into account the nonlinearity already present in the $\hat{\boldsymbol{\xi}}_{up}$, and (ii) recover the residual amount of nonlinearity needed to clear up the endogeneity of the endogenous variables in the model. This also explains why the $\hat{\mathbf{f}}_p(\hat{\boldsymbol{\xi}}_{up})$ can not be used to display the relationship between the unobservables and the response. As mentioned in Section 3.3.1, this is not problematic since all that is needed is to account for information about those unobservables that have a detrimental impact on the estimation of the effects of interest.

Intuitively, the consistency arguments for the 2SGLM estimator could be extended to 2SGAM by recalling that a GAM can be seen as a GLM whose design matrix contains the basis functions of the smooth components in the model, and by adapting the asymptotic results of Kauermann *et al.* (2009) to this context. The discussion of these properties is beyond the scope of this work, hence we do not pursue it further. Implementation of 2SGAM is straightforward. It just involves applying a penalized regression spline approach twice by using one of the reliable packages available to fit GAMs. This is particularly appealing since the amount of smoothing for the smooth com-

ponents in the models of the two-step approach can be selected reliably by taking advantage of the recent computational developments in the GAM literature.

3.5 Confidence interval correction

The large sample posterior for the generic parameter vector containing all regression spline coefficients is given in Section 2.4.

Since the second-stage of 2SGAM can not automatically account for an additional source of variability introduced via the quantities calculated in the first step, the confidence intervals for the component functions of the second-step model will be too narrow, hence leading to poor coverage probabilities. This can be rectified via posterior simulation.

The algorithm we propose is as follows:

1. Fit the first-step models, and let the first-stage parameter estimates be $\hat{\boldsymbol{\alpha}}^{[p]}$ and the estimated parameter covariance Bayesian matrix be $\hat{\mathbf{V}}_{\boldsymbol{\alpha}}^{[p]}$, where $p = 1, \ldots, h$.

2. Fit the second-stage model, and let $\hat{\boldsymbol{\beta}}^{[1]}$ and $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{[1]}$ be the corresponding parameter estimates and covariance Bayesian matrix.

3. Repeat the following steps for $k = 2, \ldots, N_b$.

   (a) For each first-stage model p, simulate a random $N(\hat{\boldsymbol{\alpha}}^{[p]}, \hat{\mathbf{V}}_{\boldsymbol{\alpha}}^{[p]})$, calculate new predicted values $\mathbf{x}_{ep}^{*}$, and then obtain $\hat{\boldsymbol{\xi}}_{up}^{*}$.

   (b) Fit the second-stage model where the $\hat{\boldsymbol{\xi}}_{up}$ are replaced with the $\hat{\boldsymbol{\xi}}_{up}^{*}$. Then store $\hat{\boldsymbol{\beta}}^{[k]}$ and $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{[k]}$.

4. For $k = 1, \ldots, N_b$, simulate N_d random draws from $N(\hat{\boldsymbol{\beta}}^{[k]}, \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{[k]})$, and then find approximate Bayesian intervals for the component functions of the second-stage model.

In words, samples from the posterior distribution of each first-step model are used to obtain samples from the posterior of the quantities of interest $\boldsymbol{\xi}_{up}$. Then, given N_b replicates for each $\boldsymbol{\xi}_{up}$, N_d random draws from the N_b posterior distributions of the second-stage model are used to obtain approximate Bayesian intervals for the smooth functions in the model. In this way, the extra source of variability introduced via the quantities calculated in the first step models can be accounted for. Simulation experience suggests that, depending

on the number of reduced form equations in the first step, small values for N_b and N_d will be tolerable. In practice, as a rule of thumb, $N_b = 25 \times p$ and $N_d = 100$ yield good coverage probabilities.

As explained by Ruppert *et al.* (2003), result (2.5) and, as a consequence, our correction procedure can not be used for variable selection purposes. In the presence of several candidate predictors, it is possible to carry out variable selection using information criteria, test statistics and shrinkage methods. Since in this context it is not straightforward to correct the second-step estimated standard errors analytically, we suggest to use a shrinkage method. This is because shrinkage approaches are based on the estimated components of a model. Hence, we can exploit the fact that 2SGAM yields consistent term estimates which can in turn lead to consistent covariate selection. Some shrinkage smoothers for GAMs are provided in Chapter 5.

3.6 Simulation study

To explore the empirical properties of the 2SGAM estimator, a Monte Carlo simulation study was conducted. The proposed two-stage approach was tested using data generated according to four response variable distributions and two data generating processes (DGP1 and DGP2). For both DGPs the number of endogenous variables in the model was equal to one. All computations were performed using `R` with GAM setup based on the `mgcv` package.

The performance of 2SGAM was compared with naive GAM estimation (i.e. the case in which the model is fitted without accounting for unmeasured confounding), and complete GAM estimation (i.e. the case in which the unobservable is included in the model). No competing methods were employed since, to the best of our knowledge, there are not available IV alternatives that can deal with GAMs in which the amount of smoothing for the smooth components can be selected via a reliable numerical method.

3.6.1 DGP1

The linear predictor was generated as follows

$$\eta = f_1(x_{o1}) + f_2(x_e) + f_3(x_u) + x_{o2}. \tag{3.16}$$

The test functions used for both DGPs are displayed and defined in Figure 3-1 and Table 3.1, respectively. For each set of correlations, sample size and

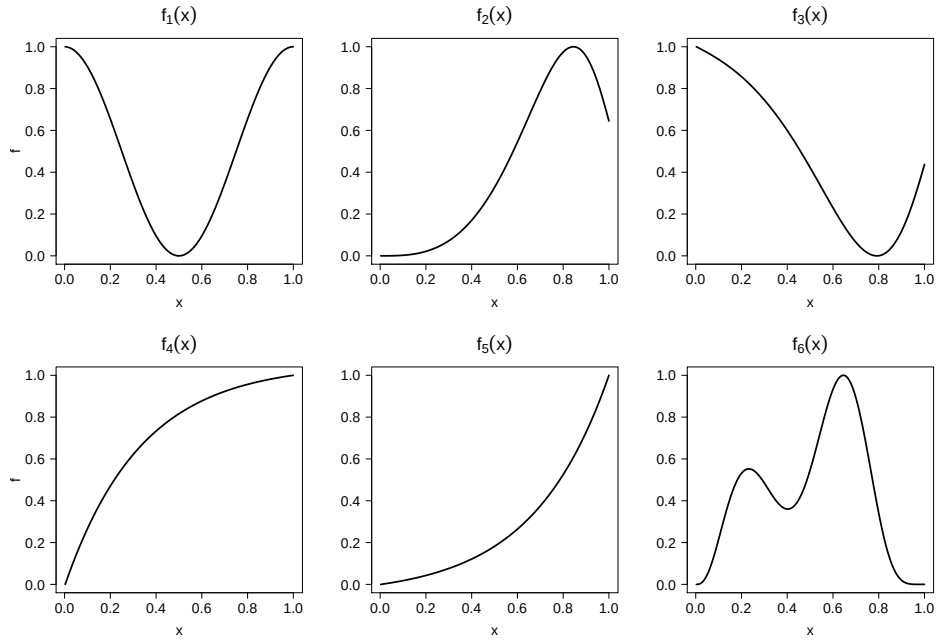

Figure 3-1: The six test functions used in the linear predictors.

response distribution, we carried out the following steps:

1. Simulate x_{o1} and x_{o2} from a multivariate uniform distribution on the unit square. This was achieved using the algorithm from Gentle (2003). Specifically, using R, two uniform variables with correlation approximately equal to 0.5 were obtained as follows

```
library(mvtnorm)
cor <- array(c(1,0.5,0.5,1),dim=c(2,2))
var <- pnorm(rmvnorm(n,sigma=cor))
xo1 <- var[,1]; xo2 <- var[,2]
```

2. Simulate x_u, x_{IV1} and x_{IV2} from independent uniform distributions on (0,1).

3. Simulate the endogenous/treatment variable of interest as follows

$$x_e = \theta_1 f_4(x_u) + \theta_2 f_5(x_{IV1}) + \theta_3 f_6(x_{IV2}) + \zeta,$$

where $\zeta \sim N(0,1)$, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ was chosen to obtain the set of correlations $\rho\{f_2(x_e), f_3(x_u)\} \in \{-0.4, -0.6\}$ and $\rho\{x_e, f_5(x_{IV})\} \in \{0.4, 0.7\}$. The three functions were scaled to have the same importance, and x_e was scaled so that its values were between 0 and 1.

4. Scale the model terms in (3.16) to have the same magnitude, and then generate the linear predictor.

5. Generate data according to the chosen outcome distribution.

$f_1(x) = \cos(2\pi x)$
$f_2(x) = 0.5\{x^3 + \sin(\pi x^3)\}$
$f_3(x) = -0.5\{x + \sin(\pi x^{2.5})\}$
$f_4(x) = -e^{-3x}$
$f_5(x) = e^{3x}$
$f_6(x) = x^{11}\{10(1-x)\}^6 + 10(10x)^3(1-x)^{10}$

Table 3.1: Test function definitions. f_1 - f_6 are plotted in Figure 3-1.

3.6.2 DGP2

Here, the linear predictor was defined as

$$\eta = f_1(x_{o1}) + \beta_e x_e + f_3(x_u) + x_{o2}, \tag{3.17}$$

where x_e was a binary variable with the corresponding parameter β_e. For each set of correlations, sample size and response distribution, we followed the same steps as in DGP1 but steps 3 and 4 were replaced with:

3. Simulate x_e according to the following mechanism

$$\begin{cases} x_e = 1 \quad if \quad x_e^* = \phi_1 + \phi_2 f_4(x_u) + \phi_3 f_5(x_{IV1}) + \phi_4 f_6(x_{IV2}) + \zeta > 0 \\ x_e = 0 \quad if \quad x_e^* \leq 0 \end{cases}$$

   where $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3, \phi_4)$ was chosen to obtain the set of correlations $\rho\{x_e^*, f_3(x_u)\} \in \{-0.4, -0.6\}$ and $\rho\{x_e^*, f_5(x_{IV})\} \in \{0.4, 0.7\}$. The three functions were scaled to have the same magnitude.

4. Generate (3.17) by setting $\beta_e = 2$ and scaling all model terms (except for x_e) to have the same magnitude.

3.6.3 Common parameter settings

One-thousand replicate data sets were generated for each DGP, combination of correlations, sample size and distribution (see Table 3.2). The 2SGAM approach, naive GAM estimation, and complete GAM estimation were employed

using penalized thin plate regression splines (Wood, 2006) based on second-order derivatives and with basis dimensions equal to 10. Step 1 was achieved by fitting an additive model for DGP1 and a GAM with probit link for DGP2. The first step models did not include x_{IV2}. The smoothing parameters were selected by the computational methods for multiple smoothing parameter estimation of Wood (2006, 2008). For each data set and estimation procedure, we obtained the mean squared error (MSE) for the estimated smooth function/dummy parameter of the treatment variable of interest. Then from the resulting 1000 MSEs, an overall mean was taken and its standard deviation calculated. Complete GAM estimation results represented our benchmark.

	$binomial$	$gamma$	$Gaussian$	$Poisson$
n	$250, 500, 1000, 2000, 4000, 8000$			
$g(\mu)$	$logit$	log	$identity$	log
$l \leq \eta \leq u$	$[0.02, 0.98]$	$[0.2, 3]$	$[0, 1]$	$[0.2, pmax]$
s/n	$n_{bin}= 1$	$\phi = 0.6$	$\sigma= 0.4$	$pmax = 3$

Table 3.2: Observations were generated from the appropriate distribution with true response means, laying in the specified range, obtained by transforming the linear predictors by the inverse of the chosen link function. l, u and s/n stand for lower bound, upper bound and signal to noise ratio parameter, respectively. The linear predictor for the binomial case was scaled to produce probabilities in the range $[0.02, 0.98]$; observations were then simulated from binomial distributions with denominator n_{bin}. In the gamma case the linear predictor was scaled to have range $[0.2, 3]$ and one value for ϕ used. For the Gaussian case normal random deviates with mean 0 and standard deviation σ were added to the true expected values, which were then scaled to lay in $[0, 1]$. The linear predictor of the Poisson case was scaled in order to yield true means in the interval $[0.2, 3]$. Notice that the chosen signal to noise ratio parameters yielded low informative responses. See Section 5.3 for further details.

3.6.4 Results

To save space, not all simulation results are shown. Missing plots convey the same information, hence it suffices to use those selected here to draw conclusions.

Figure 3-2 shows the MSE results for $\hat{f}_2(x_e)$ when data are simulated from a Bernoulli distribution using DGP1. Naive GAM yields MSEs that appear to be rather high for all cases, whereas the 2SGAM results indicate that the proposed method performs as well as complete GAM provided that the IV is strong. For the cases in which the IV is not strong, 2SGAM still performs better than the naive method, but worse than complete GAM. This is to be expected since the proposed approach, as well as any other IV method, works satisfactorily provided that the IV induces substantial variation in the endogenous

variable of interest (Wooldridge, 2002). The effect of the endogenous variable will be always better estimated when all confounders can be observed and included in the model as in complete GAM. In fact, all we can hope for is to have a method which is as good as complete GAM when valid and strong instruments are available. 2SGAM yields MSEs that converge to those of complete GAM, that in turn converge to zero as the sample size increases. Naive GAM can not produce better estimates as the sample size increases since unmeasured confounding is not accounted for. This can be clearly seen in Figure 3-3. Excluding x_{IV2} from the first step auxiliary regressions did not significantly affect the 2SGAM performance. In fact, all that is usually required to obtain consistent parameter estimates is that at least one IV is available for each endogenous regressor in the model (Wooldridge, 2002).

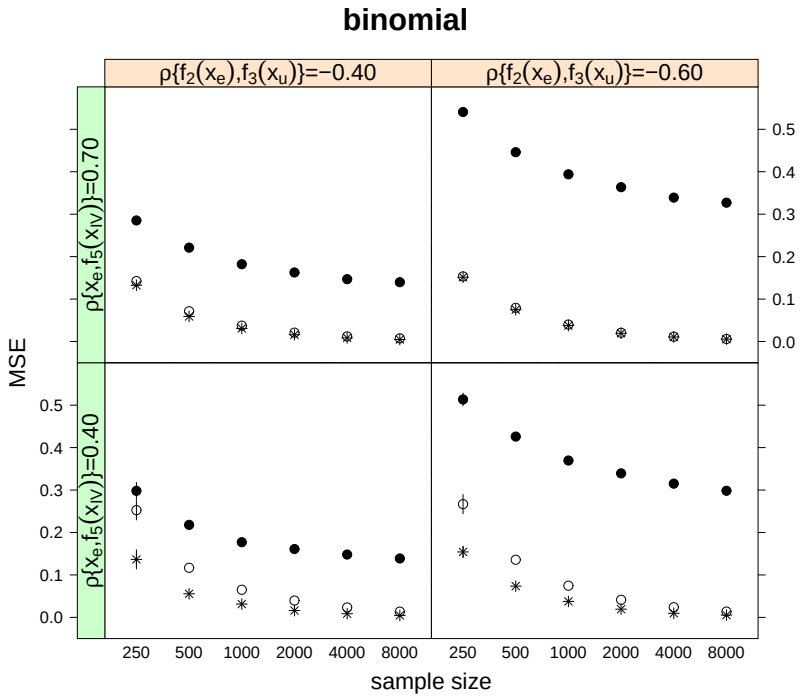

Figure 3-2: MSE results for $\hat{f}_2(x_e)$ when data are simulated from a Bernoulli distribution using DGP1. Details are given in Sections 3.6.1 and 3.6.3. $\circ$ indicates the 2SGAM estimator results, whereas $\bullet$ and $*$ refer to the cases in which estimation is carried out without accounting for unmeasured confounding, and that in which the unobservable is available and included in the model. $*$ represents our benchmark since the right model is fitted. The vertical lines show ± 2 standard error bands, which are only reported for the cases in which they are substantial. Notice the good overall performance of the proposed method for all sets of correlations and sample sizes.

Figure 3-4 shows the MSE results for $\hat{\beta}_e$ when data are simulated from a gamma distribution using DGP2. These findings complement the results discussed above. For sample sizes greater than 2000, all previous considerations apply. For the remaining sample sizes, when $\rho\{x_e^*, f_5(x_{IV})\} = 0.4$, naive GAM

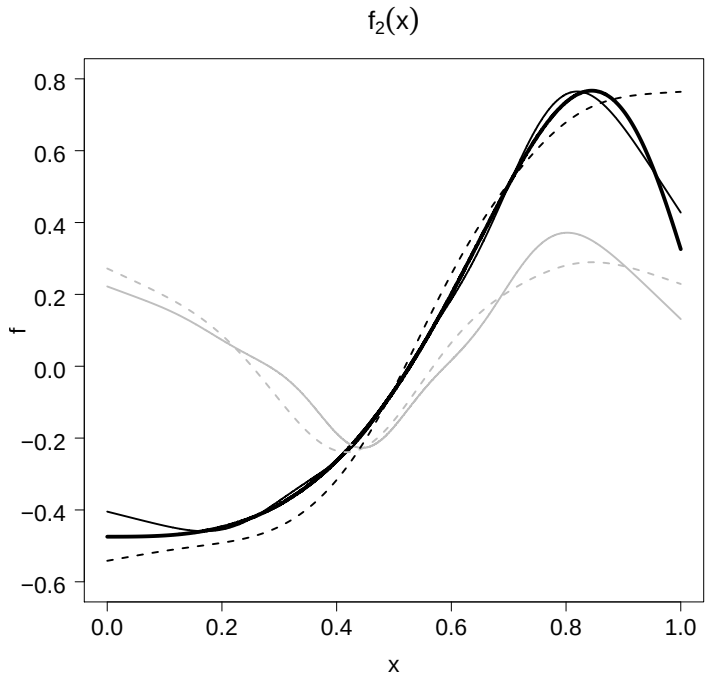

Figure 3-3: Typical estimated smooth functions for $f_2(x_e)$ (ticker solid black line) when employing the 2SGAM approach (black lines) and naive GAM estimation (grey lines). The dotted and solid lines indicate the results for the cases in which $n = 1000$ and $n = 8000$, respectively. Notice the convergence of the proposed method to the true function as opposed to the naive approach.

seems to outperform 2SGAM. This is not surprising since, as discussed in Section 3.3.1, the correction achieved by using the proposed approach when data are generated using DGP2 is approximate, case in which a strong instrument can help to obtain better adjusted estimates.

As pointed out by Staiger and Stock (1997) and Bound *et al.* (1995), IV methods can be ill-behaved if the instruments are not highly correlated with the endogenous variables of interest. This is because seemingly small correlations between instruments and unmeasured confounders can cause severe inconsistency, hence severe finite sample bias if the IVs are weak. Given that there will always be some empirical correlation at finite sample sizes, biased estimates can be avoided if the IVs are strong. In our simulation study, this requirement becomes even more relevant when dealing with DGP2, where more information is usually needed to obtain consistent estimates of the parameter of interest.

The results obtained by using 2SGLM, naive GLM estimation and complete GLM estimation (not reported here) were similar to those reported above, but obviously none of the methods could yield estimates converging to the true values. In fact, for the DGPs considered here, full parametric modelling could

not account for the nonlinear effects of the confounders as well as model the nonlinearities of the treatment variable of interest for DGP1.

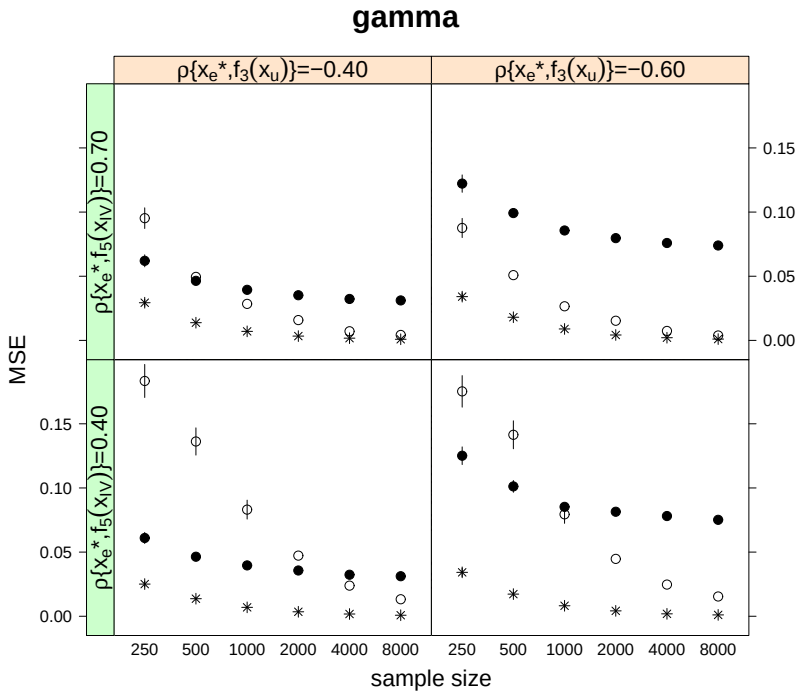

Figure 3-4: MSE results for $\hat{\beta}_e$ when data are simulated from a gamma distribution using DGP1. Details are given in Sections 3.6.2 and 3.6.3, and in the caption of Figure 4-3. For low sample sizes the naive method seems to outperform 2SGAM when the instrument is not strong. See Section 3.6.4 for an explanation of this result.

Table 3.3 shows some of the across-the-function coverage probabilities for $\hat{f}_2(x_e)$ when using the proposed two-step approach without correction for the Bayesian intervals, and the two-stage approach employing the interval correction introduced in Section 3.5, with $N_b = 25$ and $N_d = 100$. The results show that the proposed correction produces Bayesian intervals with coverage probabilities very close to the nominal level. 2SGAM without correction yields intervals which are too narrow. This results in undercoverage. However, as the sample size increases the coverage probabilities improve. This is because, as n increases the first step quantities are estimated more reliably, hence the neglect of the variability of these quantities might not have a substantial detrimental impact on the model term coverages of the second step model. In other words, if the data have high information content, first step quantities will be estimated more reliably and, as a result, uncorrected intervals are likely to yield better coverages. The coverage probability results for $\hat{\beta}_e$ (not reported here) led to the same conclusions, but, as for the estimation results for $\hat{\beta}_e$, at larger sample sizes.

	2SGAM				AD.2SGAM			
	250	500	1000	4000	250	500	1000	4000
binomial	0.92	0.92	0.93	0.93	0.94	0.94	0.95	0.95
gamma	0.90	0.91	0.91	0.92	0.94	0.94	0.94	0.95
Gaussian	0.91	0.92	0.93	0.93	0.94	0.95	0.95	0.95
Poisson	0.90	0.92	0.93	0.93	0.94	0.95	0.95	0.95

Table 3.3: Across-the-function coverage probability results for $\hat{f}_2(x_e)$ at four sample sizes, for the nominal level 95%, when the correlation between instrument and endogenous variable is 0.7 and that between endogeous and unobservable equal to -0.6. 2SGAM, and AD.2SGAM stand for the proposed two-step approach without correction for the Bayesian intervals, and the two-step approach with the correction described in Section 3.5, with $N_b = 25$ and $N_d = 100$. Notice the good coverage probabilities obtained when employing the correction.

It should be pointed out that IV methods are never unbiased when at least one explanatory variable is endogenous in the model. We know that model term estimates are biased when the error term is correlated with some of the regressors. IV approaches solve this problem but only asymptotically, since they are based on the assumption that the instruments are asymptotically uncorrelated with the unobservables (Wooldridge, 2002). Unfortunately, we can not observe the unmeasured confounders, hence we can not know to what extent the issues above affect our empirical analysis. As a rule of thumb, IV methods should be used if the instruments are believed to satisfy the IV assumptions.

3.7 Illustration of 2SGAM

In order to illustrate the 2SGAM approach, we investigated the effect of private health insurance on private medical care utilization using data from an Italian population-based survey. Private health insurance coverage is not randomly assigned as in a controlled trial but rather is the result of supply and demand, including individual preferences and health status. As a consequence, differences in outcomes for insured and uninsured individuals might be due not only to the effect of health insurance, but also to the effect of unobservables which are associated with insurance coverage. If this fact is not accounted for, the estimated impact of private health insurance will not be realistic, leading to biased health policy conclusions. Buchmueller *et al.* (2005) provide an excellent review of these issues.

3.7.1 Data

We used data from the Survey on Health, Aging and Wealth (SHAW, Brugiavini *et al.*, 2002) which was conducted by the leading Italian polling agency DOXA in 2001. The SHAW sample consists of 1068 households whose head is over 50 years old, and mainly provides information about individual health status, utilization of health services, types of insurance coverage, as well as socio-economic features. The outcome was utilization of private health care: an indicator variable that takes value 1 if the subject had private examinations and 0 otherwise. The endogenous/treatment variable was private health insurance: a dummy variable with value 1 if the respondent had private insurance coverage and 0 otherwise. The measured confounders were given by five factors (consumption of strong alcohol, marital status, self-reported health status, sex, smoking status) and three continuous variables (age, body mass index (bmi), income).

As pointed out in Section 3.2, identification of a valid instrument may not be straightforward because this choice has to be based on subject-matter knowledge, not statistical testing. Despite the effort of many researchers in trying to correctly quantify the impact that private health insurance has on utilization of private health care, there is not a general agreement on which instrument should be selected for statistical analysis (see Buchmueller *et al.* (2005), and references therein for a review of the relevant literature). Taking these findings into account, on the basis of the variables already included in the model, and depending on the remaining predictors available in the data set at hand, indemnity insurance (which is a binary variable) was suggested as an instrument possibly meeting the three conditions discussed in Section 3.2. Of course, since some of the necessary assumptions can not be verified in practice, we can not be certain about the empirical validity of this instrument.

3.7.2 Health care modelling

The aim is to quantify the effect that private health insurance has on utilization of private health care by accounting for unmeasured risk factors. Two logistic GAM models were employed to implement the 2SGAM approach. To keep the illustration simple, the response variables of the two models were modelled considering all main effects only. Thin plate regression splines of the continuous regressors with basis dimension 10 and penalties based on second-order derivatives were used. Smoothing parameters were automatically selected as

explained in Section 3.6.3. The factor variables were kept as parametric model components. The estimated smooth functions of bmi and $\hat{\boldsymbol{\xi}}_u$ support the presence of nonlinearities (see Figure 3-5). Recall that $\hat{\mathbf{f}}(\hat{\boldsymbol{\xi}}_u)$ does not have any meaningful interpretation (see Section 3.4.1).

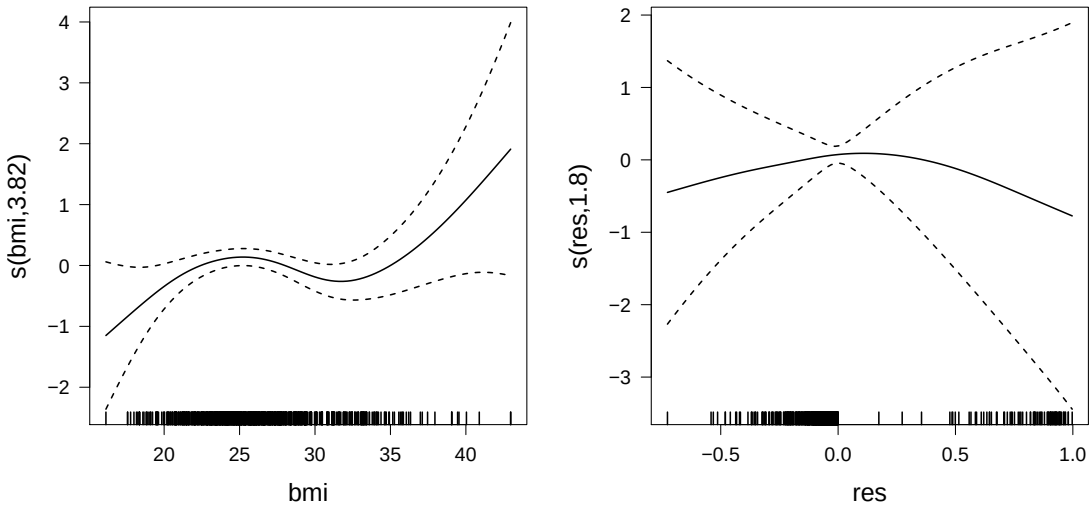

Figure 3-5: Smooth function estimates of body mass index (bmi) and $\hat{\boldsymbol{\xi}}_u$ on the scale of the linear predictor, for the second stage equation. Dashed lines represent 95% Bayesian confidence intervals corrected as discussed in Section 3.5. The numbers in brackets in the y-axis captions are the estimated degrees of freedom or effective number of parameters of the smooth curves. The rug plot, at the bottom of each graph, shows the covariate values

Overall our results are consistent with those reported in the health care utilization literature (Buchmueller *et al.*, 2005; Harmon and Nolan, 2001; Hofter, 2006; Reidpath *et al.*, 2002). As discussed earlier on, the target is to obtain an adjusted estimate of the impact of private health insurance on utilization of private health care. Naive GAM estimation yielded $\hat{\beta}_e = 0.39\ (-0.16, 0.95)$, whereas 2SGAM with corrected intervals produced $\hat{\beta}_e = 0.94\ (0.03, 1.83)$. Although there is not any statistical difference between the two estimates, the former shows no significant effect whereas the latter exhibits a statistically significant estimate. This suggests that the unobservable confounding issue affects the parameter of interest. The differences between the other parametric parameter estimates of naive GAM and 2SGAM were minimal, confirming that the presence of unmeasured confounding has to be accounted for in order to obtain a consistent estimate of β_e. The plot depicting the smooth of bmi for naive GAM has not been reported as it was similar to that in Figure 3-5. These findings were not unexpected since it is well known that the endogenous parameter of interest is generally the most affected (e.g. Wooldridge 2002, ch.

5).

The validity of the 2SGAM results certainly depends on the degree that the IV assumptions are met. However, as pointed out by Johnston *et al.* (2008), "in observational settings where unmeasured confounding is suspected ... analysis using an imperfect instrument can still help in providing a more complete picture than regression alone."

3.8 Discussion

The unobservable confounding issue is likely to affect the majority of observational studies in which the researcher is interested in evaluating the effect of one or more predictors of interest on a response variable. When unmeasured confounding is not controlled for any standard estimation method will yield biased and inconsistent parameter estimates.

The IV approach represents a valid means to account for unmeasured confounding. This technique, first proposed in econometrics, only recently has received some attention in the applied statistical literature. We have proposed a flexible procedure to carry out IV analysis within the GAM context. Our proposal is backed up with an extensive simulation experiment whose results confirmed that 2SGAM represents a flexible theoretically sound means of obtaining consistent curve/parameter estimates in the presence of unmeasured confounding. We have also proposed a Bayesian interval correction procedure for 2SGAM. In simulation, the resulting intervals performed well in terms of coverage probabilities.

The major drawback in all IV methods (including ours) is the difficulty in choosing an appropriate instrument. Given that not all IV assumptions can be tested empirically, logical arguments must be presented to justify the instrument choice. However, statistical analysis using an imperfect instrument can still help in providing insights into the possible effect that unmeasured confounding has on the estimated relationship of interest.

Chapter 4

Coverage Properties of Confidence Intervals

In this chapter, we study the coverage properties of Bayesian confidence intervals for the smooth component functions of GAMs. The intervals are the usual generalization of Wahba (1983) or Silverman (1985) intervals to the GAM component context. We present simulation evidence showing these intervals have close to nominal 'across-the-function' frequentist coverage probabilities, except when the truth is close to a straight line/plane function. We extend Nychka's (1988) argument for univariate smoothing splines to explain these results. The theoretical argument suggests that good coverage probabilities can be achieved, provided that heavy oversmoothing is avoided, so that the bias is not too large a proportion of the sampling variability. Otherwise, because the Bayesian intervals account for bias and variance, the coverage probabilities are surprisingly insensitive to the exact choice of smoothing parameter. The theoretical results allow us to derive alternative intervals from a purely frequentist point of view, and to explain the impact that the neglect of smoothing parameter variability has on confidence interval performance. They also suggest switching the target of inference for component-wise intervals away from smooth components in the space of the GAM identifiability constraints. Instead intervals should be produced for each function as if only the other model terms were subject to identifiability constraints. If this is done then coverage probabilities are improved.

4.1 Introduction

Inference for *univariate* spline models can be effectively achieved using the Bayesian confidence intervals proposed by Wahba (1983) or Silverman (1985). As theoretically shown by Nychka (1988), for the case of univariate models whose Bayesian intervals have close to constant width, a very interesting feature of these intervals is that they work well when evaluated by a frequentist criterion, provided coverage is measured 'across-the-function' rather than pointwise. Specifically, consider the model

$$Y_i = f(x_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2),$$

where the ϵ_i are mutually independent. According to Nychka's results, if the smoothing parameter is sufficiently reliably estimated (e.g. by GCV) that the bias in the estimates is a modest fraction of the mean squared error for $f(x)$, then the average coverage probability (ACP)

$$\mathrm{ACP}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \Pr[f(x_i) \in BI_\alpha(x_i)]$$

is very close to the nominal level $1-\alpha$, where $BI_\alpha(x)$ indicates the $(1-\alpha)100\%$ Bayesian interval for $f(x)$ and α the significance level. This agreement occurs because the Wahba/Silverman type intervals include both a bias and variance component. Note that for convenience we define ACP only over the design points, rather than the whole function (but this restriction makes no practical difference for a smooth well sampled function). These intervals as well as their component-wise extensions have been derived when dealing with Gaussian and non-Gaussian data (e.g. Gu, 1992; Gu, 2002; Gu and Wahba, 1993; Ruppert *et al.*, 2003; Wood, 2006). See Section 2.4 for the expression of the large sample posterior for the generic parameter vector containing all regression spline coefficients.

Several other approaches have also been proposed to produce inferential tools for GAMs. Hastie and Tibshirani (1990) suggested using simple frequentist approximations to produce approximate confidence intervals. Ruppert *et al.* (2003, Section 6.4) showed instead the link between the mixed model approach and the Wahba/Silverman type intervals, but their derivation gives no indication of whether these intervals should have close to nominal coverage probabilities under repeated sampling from a fixed true function. Other frameworks include the use of bootstrap methods as well as the fully Bayesian

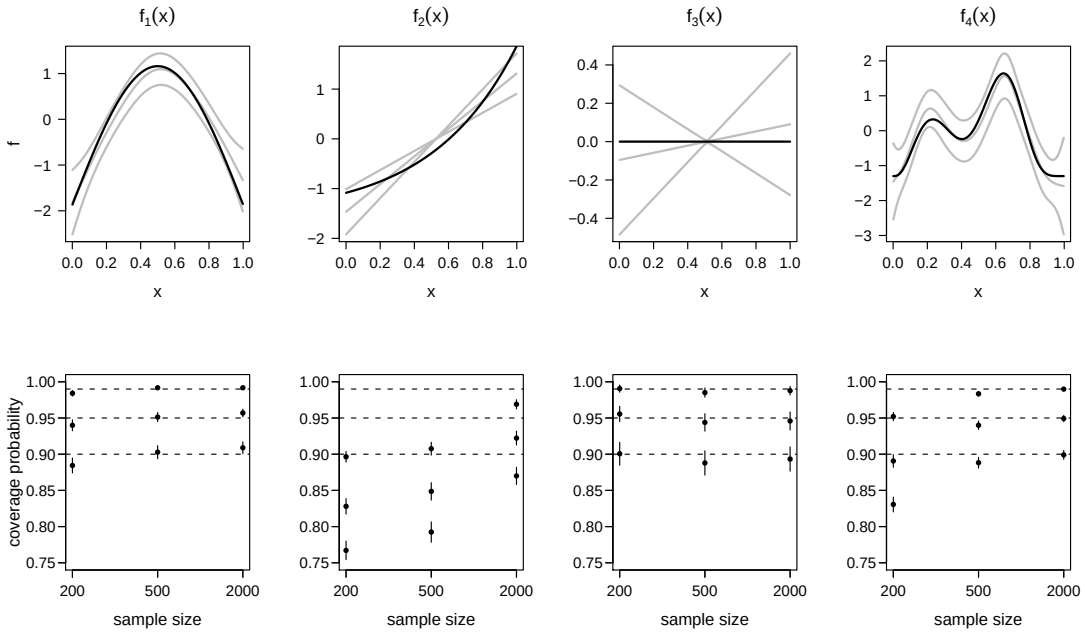

Figure 4-1: Results from component-wise Bayesian intervals for Bernoulli simulated data at three sample sizes. Observations were generated as $logit\{\mathbb{E}(Y_i)\} = \alpha + z_i + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i})$, where Y_i followed a bernoulli distribution and uniform covariates on the unit interval with correlations equal to 0.5 were employed (see section 4.3.1 for details). The function definitions are given in Table 4.1. The functions were scaled to have the same magnitude in the linear predictor and then the sum rescaled to produce probabilities in the range [0.02,0.98]. 1000 replicate datasets were then generated and GAMs fitted using penalized thin plate regression splines (Wood, 2003) with basis dimensions equal to 10, 10, 10 and 20, respectively, and penalties based on second-order derivatives. Multiple smoothing parameter selection was by generalized AIC (Wood, 2008). Displayed in the top row are the true functions, indicated by the black lines, as well as example estimates and 95% Bayesian confidence intervals (gray lines) for the smooths involved. • represents the mean coverage probability from the 1000 across-the-function coverage proportions of the intervals, vertical lines show ± 2 standard error bands for the mean coverage probabilities, and dashed horizontal lines show the nominal coverage probabilities considered.

approach. For example, Härdle and Bowman (1988) and Härdle and Marron (1991) used bootstrap to construct pointwise and simultaneous confidence intervals. An under-smoothing approach has also been taken within the bootstrap framework, as a devise for avoiding smoothing bias (Hall, 1992; Kaurmann and Opsomer, 2003). Direct bootstrapping has been employed as well, as in Härdle *et al.* (2004) who make use of 'Wild' bootstrap methods. As an alternative, Fahrmeir *et al.* (2004) and Fahrmeir and Lang (2001) adopted the fully Bayesian approach, which employs MCMC for practical computations. Wang and Wahba (1995) compared the Wahba/Silverman type confidence intervals with those derived using several variations of the bootstrap approach. They found that the bootstrap framework can yield intervals that are comparable to the Bayesian ones in terms of across-the-function coverage properties. However, they are computationally intensive, a problem which may affect the fully Bayesian approach as well.

Although Nychka's theoretical analysis has not been extended to non-constant width intervals for a smooth function that is a component of a larger model, simulation evidence suggests that result (2.5) might yield intervals for GAM components that perform well. As an example, Figure 4-1 illustrates that coverage probabilities for smooth functions with different degrees of complexity can be rather close to the nominal levels. However, it also suggests that, for smooth components close to a straight line/plane function (such as $f_2(x)$), coverage probabilities can be too low. These results suggest that the use of the $\boldsymbol{\beta}$ posterior distribution (2.5) does not always yield good coverage probabilities, and it is worth investigating when and why this happens. Intuitively it seems that when a 'true' function is 'close' to a straight line (or, more generally, any function in the penalty null space), there is a high chance that the estimate of that function will be a straight line. In this case, because of the identifiability constraints on component functions, the confidence interval width will shrink to zero where the straight line passes through zero, and, it seems likely, as Figure 4-1 suggests, that this leads to poor observed coverage probabilities.

The aim of this chapter is to extend and modify Nychka's analysis to derive non-constant width intervals for GAM components in a way that reveals their coverage properties. Our theoretical arguments show why non-constant Wahba/Silverman type intervals for smooth components work well in a frequentist setting, and explain when and why they fail. As a result, a fix for the case in which the intervals fail is suggested and alternative intervals are derived from a purely frequentist point of view. The impact that the neglect of smoothing parameter uncertainty has on confidence interval performance

is also explained by our results. All findings are supported by an extensive simulation study.

4.2 Confidence Intervals

The aim of this section is to develop a construction of variable width component-wise intervals. The primary purpose of this construction is to reveal the coverage properties of the usual component-wise extension of Wahba/Silverman type intervals as discussed, for instance, by Gu and Wahba (1993), Fahrmeir *et al.* (2004), Ruppert *et al.* (2003) and Wood (2006b), to name a few. The initial part of Section 4.2.1 is similar to the construction that can be found in Ruppert *et al.* (2003, Section 6.4), but thereafter we are forced to follow a line more similar to Nychka (1988) in order to establish coverage properties (which the Ruppert *et al.* (2003) derivation does not reveal). Our theoretical derivations explain why Wahba/Silverman type intervals work well in a frequentist setting, and show why intervals for smooth components that are in the penalty null space are problematic. The good coverage probabilities obtained using result (2.5) are hence explained and a remedy to the near-straight-line/plane case is introduced. The theoretical arguments also allow us to derive alternative intervals when a purely frequentist approach is adopted.

For clarity the normal error, identity link, case is covered first, with the generalization to GAMs discussed subsequently. We need to start by establishing some preliminary results.

4.2.1 Estimation of $\mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2/n$, and σ^2

The subsequent arguments will require that we can estimate the expected mean squared error of linear transformations of the coefficient estimates $\hat{\boldsymbol{\beta}}$, so this needs to be addressed first.

Consider, then, an arbitrary linear transformation defined by a matrix of fixed coefficients $\mathbf{B}$. We seek an estimate for the expected mean squared error, $\mathbb{E}(M_B)$, of $\mathbf{B}\hat{\boldsymbol{\beta}}$. From the Bayesian approach we have $\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\lambda},\sigma^2 \sim N(\hat{\boldsymbol{\beta}},\mathbf{V}_{\boldsymbol{\beta}})$. Taking expectations with respect to this distribution implies

$$\mathbb{E}(M_B) = \frac{1}{n}\mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2 = \frac{1}{n}\mathrm{tr}(\mathbf{B}^{\mathsf{T}}\mathbf{B}\mathbf{V}_{\boldsymbol{\beta}})$$

but this appears to require that we accept the whole Bayesian analysis. Furthermore, using this mean squared error for frequentist purposes involves a

troubling exchange of the the roles of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. In fact we can make do with far fewer of the assumptions present in the Bayesian analysis, and can be much more precise about exactly what assumption is needed where.

First let $\tilde{\boldsymbol{\beta}}$ denote the unpenalized, and hence unbiased, estimate of $\boldsymbol{\beta}$. Let $\mathbf{F} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$ be the matrix such that $\hat{\boldsymbol{\beta}} = \mathbf{F}\tilde{\boldsymbol{\beta}}$ ($\mathbf{X}$ is the full model matrix here). It follows immediately that $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbf{F}\boldsymbol{\beta}$. It is then routine to show that the mean square error can be partitioned into a variance term and a mean squared bias term

$$\mathbb{E}(M_B) = \frac{1}{n}\mathbb{E}\|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = \frac{1}{n}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{V}_{\hat{\boldsymbol{\beta}}}) + \frac{1}{n}\|\mathbf{B}(\mathbf{F} - \mathbf{I})\boldsymbol{\beta}\|^2 \tag{4.1}$$

where $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S})^{-1}\sigma^2$. An obvious option for estimating the final term on the right hand side (the mean squared bias) is to plug in $\hat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}$, but to make the link to Bayesian intervals we also consider an alternative estimate of the term.

Progress is most easily made by re-parameterizing using a 'natural' Demmler-Reinsch type parameterization. Forming the QR decomposition $\mathbf{X} = \mathbf{QR}$ and the eigen decomposition $\mathbf{R}^{-\mathsf{T}}\mathbf{S}\mathbf{R}^{-1} = \mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$, then the reparameterization leads to the new smooth coefficient vector $\mathbf{U}^\mathsf{T}\mathbf{R}\boldsymbol{\beta}$ and matrix of fixed coefficients $\mathbf{B}\mathbf{R}^{-1}\mathbf{U}$. In this parameterization the most important matrices are diagonal, for example $\mathbf{V}_{\boldsymbol{\beta}}/\sigma^2 = \mathbf{F} = (\mathbf{I} + \mathbf{D})^{-1}$ and $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{I} + \mathbf{D})^{-2}\sigma^2$.

We do not know $\boldsymbol{\beta}$, but if we knew the distribution of likely $\boldsymbol{\beta}$ values then we could estimate the bias term by its expectation according to that distribution. The natural assumption is to use the prior employed in the Bayesian analysis, namely that $\mathbb{E}(\boldsymbol{\beta}) = \mathbf{0}$, $\mathrm{var}(\beta_k) = D_{kk}^{-1}\sigma^2$, unless $D_{kk} = 0$ (in which case the variance turns out to be immaterial), and the covariances are zero. It is then routine to show that

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\beta}}\|\mathbf{B}(\mathbf{F} - \mathbf{I})\boldsymbol{\beta}\|^2 = \frac{1}{n}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{H})\sigma^2$$

where $\mathbf{H}$ is a diagonal matrix with elements $H_{kk} = D_{kk}/(1+D_{kk})^2$. Recognizing that $\mathbf{H}\sigma^2 = \mathbf{V}_{\boldsymbol{\beta}} - \mathbf{V}_{\hat{\boldsymbol{\beta}}}$, we have the estimate

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n}\mathbb{E}\|\mathbf{B}\widehat{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}\|^2 = \frac{1}{n}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{V}_{\boldsymbol{\beta}})$$

Reversing the re-parameterization confirms that this is simply

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S})^{-1})\sigma^2 \tag{4.2}$$

in the original parameterization.

If $\mathbf{B} = \mathbf{X}$ the result leads easily to the usual estimator

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - \mathrm{tr}(\mathbf{F})}. \tag{4.3}$$

Alternatively, if we use the plug in estimate of the mean squared bias in (4.1), then we could use

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 - \|\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\|^2}{n - 2\mathrm{tr}(\mathbf{F}) + \mathrm{tr}(\mathbf{F}\mathbf{F}^{\mathsf{T}})}. \tag{4.4}$$

4.2.2 Intervals

Having completed the necessary preliminaries, we now consider the construction of intervals for some component function of a model, such that $[f(x_1), f(x_2), \ldots, f(x_n)]^{\mathsf{T}} \equiv \mathbf{f} = \mathbf{X}\boldsymbol{\beta}$. Here $\mathbf{X}$ may be the model matrix for just one model component: the matrix mapping the vector of all model coefficients to the evaluated values of just one smooth component (so, many of the columns of $\mathbf{X}$ may be zero). The approach modifies Nychka's (1988) construction in order to obtain intervals of variable width, which are also applicable in the case where the function is only a component of a larger model.

Given some convenient constants, C_i, we seek a constant, A, such that

$$\mathrm{ACP} = \frac{1}{n}\mathbb{E}\left\{\sum_i \mathbb{I}(|\hat{f}(x_i) - f(x_i)| \leq z_{\alpha/2}A/\sqrt{C_i})\right\} = 1 - \alpha \tag{4.5}$$

where 'ACP' denotes 'Average Coverage Probability', $\mathbb{I}$ is an indicator function, α is a constant between 0 and 1 and $z_{\alpha/2}$ is the $\alpha/2$ critical point from a standard normal distribution. To this end, define $b(x) = \mathbb{E}\{\hat{f}(x)\} - f(x)$ and $v(x) = \hat{f}(x) - \mathbb{E}\{\hat{f}(x)\}$, so that $\hat{f} - f = b + v$. Defining I to be a random variable uniformly distributed on $\{1, 2, \ldots, n\}$ we have

$$\begin{aligned}
\mathrm{ACP} &= \Pr\left(|b(x_I) + v(x_I)| \leq z_{\alpha/2}A/\sqrt{C_I}\right) \\
&= \Pr\left(|\sqrt{C_I}b(x_I) + \sqrt{C_I}v(x_I)| \leq z_{\alpha/2}A\right) \\
&= \Pr\left(|B + V| \leq z_{\alpha/2}A\right)
\end{aligned}$$

where $B = \sqrt{C_I}b(x_I)$ and $V = \sqrt{C_I}v(x_I)$. We need to be able to approximate the distribution of $B + V$.

Let $[b(x_1), b(x_2), \ldots, b(x_n)]^{\mathsf{T}} \equiv \mathbf{b} = \mathbb{E}(\hat{\mathbf{f}}) - \mathbf{f}$. Hence, defining $\mathbf{c} = (\sqrt{C_1}, \sqrt{C_2}, \ldots, \sqrt{C_n})^{\mathsf{T}}$,

we have

$$\mathbb{E}(B) = \sum_i \frac{1}{n} b(x_i)\sqrt{C_i} = \mathbf{c}^\mathsf{T}(\mathbb{E}(\hat{\mathbf{f}}) - \mathbf{f})/n.$$

In practice this quantity is very small (unless heavy oversmoothing is employed), but in any case it can be estimated as

$$\widehat{\mathbb{E}(B)} = \mathbf{c}^\mathsf{T}(\hat{\hat{\mathbf{f}}} - \hat{\mathbf{f}})/n = \mathbf{c}^\mathsf{T}\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})/n. \tag{4.6}$$

Now consider V. Defining $[v(x_1), v(x_2), \ldots, v(x_n)]^\mathsf{T} \equiv \mathbf{v} = \hat{\mathbf{f}} - \mathbb{E}(\hat{\mathbf{f}})$, we have $\mathbb{E}(\mathbf{v}) = \mathbf{0}$, and hence

$$\mathbb{E}(V) = \sum_i \frac{1}{n} v(x_i)\sqrt{C_i} = 0.$$

The covariance matrix of $\mathbf{v}$ is, $\mathbf{V}_{\hat{\mathbf{f}}} = \mathbf{X}\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\mathsf{T}$, the same as that of $\hat{\mathbf{f}}$. Hence

$$\mathrm{var}(V) = \sum_i \frac{1}{n}\mathbb{E}\{v(x_i)^2 C_i\} = \mathrm{tr}(\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}})/n,$$

where $\mathbf{C}$ is the diagonal matrix with leading diagonal elements C_i. Now since $\mathbf{v} \sim N(\mathbf{0}, \mathbf{V}_{\hat{\mathbf{f}}})$, V is a mixture of normals, which is inconvenient unless $[\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ is independent of i. If this constant variance assumption holds then $V \sim N(0, \mathrm{tr}(\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}})\sigma^2/n)$ (and the lack of dependence on i implies a lack of dependence on $b(x_i)$, implying independence of B and V).

It is the distribution of $B + V$ that is needed. $\mathbb{E}(B + V) = \mathbb{E}(B)$ and by construction $\mathrm{var}(B + V) = \mathbb{E}(M) - \mathbb{E}(B)^2$ where

$$M = \frac{1}{n}\sum_i C_i\{\hat{f}(x_i) - f(x_i)\}^2 = \|\sqrt{\mathbf{C}}(\hat{\mathbf{f}} - \mathbf{f})\|^2/n = \|\sqrt{\mathbf{C}}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/n.$$

Now we can exactly re-use Nychka's (1988) argument: provided B is small relative to V then $B + V$ will be approximately normally distributed, i.e. approximately

$$B + V \sim N(\mathbb{E}(B), \mathbb{E}(M) - \mathbb{E}(B)^2).$$

The $B < V$ assumption is examined in more detail in the Appendix. We can estimate $\mathbb{E}(B)$ and $\mathbb{E}(M)$ (by the results of the previous section). So, defining $\hat{\sigma}^2_{bv} = \widehat{\mathbb{E}(M)} - \widehat{\mathbb{E}(B)^2}$, we have the approximate result

$$B + V \sim N(\widehat{\mathbb{E}(B)}, \hat{\sigma}^2_{bv}).$$

Routine manipulation then results in

$$\hat{f}(x_i) - \widehat{\mathbb{E}(B)}/\sqrt{C_i} \pm z_{\alpha/2}\hat{\sigma}_{bv}/\sqrt{C_i} \tag{4.7}$$

as the definition of intervals achieving close to $1-\alpha$ ACP (i.e. $A = \sigma_{bv}$). So, it is the fact that the convolution of B and V is close to a normal that leads the intervals to have good across-the-function coverage.

So far the choice of C_i has not been discussed, but the constant variance requirement, for $[\mathbf{C}\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ to be independent of i, places strong restrictions on what is possible here. Two choices are interesting:

1. $C_i^{-1} = [\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$ ensures that the constant variance assumption is met exactly. Note that in this case, if we use (4.2) as the expected mean squared error estimate,

$$\widehat{\mathbb{E}(M)} = \frac{\hat{\sigma}^2}{n}\sum_i \frac{[\mathbf{X}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{\mathsf{T}}]_{ii}}{[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}}.$$

   In effect the resulting intervals are using the frequentist covariance matrix $\mathbf{V}_{\hat{\mathbf{f}}}$, but 'scaled up' to the 'size' of the Bayesian covariance matrix $\mathbf{V}_{\mathbf{f}} = \mathbf{X}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{\mathsf{T}}$. Using the plug in estimator of (4.1) we have

$$\widehat{\mathbb{E}(M)} = 1 + \|\sqrt{\mathbf{C}}\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\|^2/n \tag{4.8}$$

2. $C_i^{-1} = [\mathbf{V}_{\mathbf{f}}]_{ii}$. If $[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii} \approx \gamma[\mathbf{V}_{\mathbf{f}}]_{ii}$ for some constant γ, then this choice approximately meets the constant variance assumption. If we use the (typically accurate) approximation $\widehat{\mathbb{E}(B)} \approx 0$, along with the mean squared error estimate (4.2), then the resulting intervals are exactly Bayesian intervals of the Wahba/Silverman kind.

So, we arrive at the key result. Given the derivation of the intervals started from (4.5), we expect component-wise Wahba/Silverman type intervals to have close to nominal coverage properties across the function.

Notice the limited role of smoothing parameter selection in the above:

1. The requirement for B to be smaller than V, requires that we choose smoothing parameters so as not to oversmooth too heavily, but otherwise the choice of smoothing parameter is rather unimportant.

2. Accuracy of $\widehat{\mathbb{E}(B)}$ depends on $\hat{f}$ being close to f, but this is exactly what smoothing parameter selection methods try to achieve. In any case this term is typically small enough to be practically negligible.

3. $\widehat{\mathbb{E}(M)}$ will be a more reliably estimated if $\hat{f}$ is close to f, or the prior assumption for $\boldsymbol{\beta}$ is plausible, in not being too wide nor too narrow. Again, this is what smoothness selection methods try to achieve.

The forgoing immediately suggests the circumstances under which the intervals will behave poorly. If smoothing parameters are substantially overestimated, so that we substantially oversmooth some component, then the requirement for B to be smaller than V will be violated. Two situations are likely to promote oversmoothing. Firstly, highly correlated covariates are likely to mean that it is difficult to identify which corresponding smoothing parameters should be high and which low. For example if one covariate has a very smooth effect, and another a very wiggly effect, but they are highly correlated, it is quite possible that their estimated effects will have the degrees of smoothness reversed. This means that one of the covariate effects is substantially oversmoothed. Secondly, if a true effect is almost in the penalty null space then the estimated smoothing parameter may tend to infinity, forcing the estimate into the penalty null space, and as we will see below, this can be problematic.

In summary: we have shown that Bayesian component-wise variable width intervals, or our proposed alternative, for the smooth components of an additive model should achieve close to nominal across the function coverage probability, provided only that we do not oversmooth so heavily that average bias dominates the sampling variability for a term estimate. Beyond this requirement not to oversmooth too heavily, the results appear to have rather weak dependence on smoothing parameter values, suggesting that the neglect of smoothing parameter variability should not significantly degrade interval performance.

4.2.3 Generalized additive model case

The results of sections 4.2.1 and 4.2.2 can be routinely extended to the *generalized* additive model case. In the case of section 4.2.1 the results follow as large sample approximations with the substitutions

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\phi \tag{4.9}$$

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\phi \tag{4.10}$$

$$\mathbf{F} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} \tag{4.11}$$

and $\sqrt{\mathbf{W}}\mathbf{X} = \mathbf{QR}$. Then (4.2) becomes

$$\widehat{\mathbb{E}(M_B)} = \frac{1}{n}\mathrm{tr}(\mathbf{B}^\mathsf{T}\mathbf{B}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1})\sigma^2.$$

Similarly, (4.3) becomes

$$\hat{\phi} = \frac{\|\sqrt{\mathbf{W}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\|^2}{n - \mathrm{tr}(\mathbf{F})},$$

while the generalization of (4.4) is

$$\hat{\phi} = \frac{\|\sqrt{\mathbf{W}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\|^2 - \|\sqrt{\mathbf{W}}\mathbf{X}(\mathbf{F}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\|^2}{n - 2\mathrm{tr}(\mathbf{F}) + \mathrm{tr}(\mathbf{F}\mathbf{F}^\mathsf{T})}.$$

Section 4.2.2 also follows as before, but again with the substitutions (4.9) – (4.11). The key requirement that $\mathbf{v} \sim N(\mathbf{0}, \mathbf{V}_{\hat{\mathbf{f}}})$ is now a large sample approximation, which follows from the large sample normality of $\hat{\boldsymbol{\beta}}$, which can readily be established (e.g. Wood, 2006b).

4.2.4 What the results explain

Our results explain the success of Bayesian, component-wise, variable width intervals, which is evidenced in the simulation study of the next section. More interestingly, they explain the cases where the Bayesian intervals fail. The major failure, evident from simulations (see Figure 4-1), occurs when a smooth component is close to a function in the null space of the component's penalty (i.e. to a straight line or plane, for the examples in this chapter) and may therefore be estimated as exactly such a function. The component will have been estimated subject to an identifiability constraint, but when intervals are constructed subject to such a constraint, the observed coverage probabilities are poor. The preceding theory explains why. For example, when a term is estimated as a straight line but *subject to an identifiability constraint* then the associated confidence interval necessarily has width 0 where the line passes through the zero line. In this case the sampling variability must be smaller than the bias over some interval surrounding the point, and the assumption that B is less than V will fail, given also that the C_i associated with this interval will be very large.

The theory also suggests a remedy to this problem: compute each term's interval as if it alone were unconstrained, and identifiability was obtained by constraints on the other model terms (see section below).

Notice that the interval failure in the constrained straight line/plane case is not just the result of failing to meet the original Nychka (1988) constant width assumption: variable width intervals are quite acceptable under our extension of Nychka's argument: the problem occurs when the interval width shrinks to zero somewhere.

The poor component-wise coverages reported in Wood (2006b) are also explained by our results. The reported simulations were performed (in 2001/2) with the original smoothness selection method proposed in Wood (2000), and using uncorrelated covariates. This method alternated Newton updates of the smoothing parameters with a computationally cheap global search for an 'overall' smoothing parameter. While superficially appealing, this global search can miss a shallow minimum altogether, and place one or more smoothing parameters in a part of the smoothing parameter space in which the smoothness selection criteria is completely flat. Once smoothing parameters are at such a point, then they can only return by accident, in another global search, and this rarely happens. The upshot is that too many straight line/plane are estimated, and as we have seen this degrades the performance of the associated intervals. The bootstrapping method proposed by Wood (2006b) to improve interval coverage appears to have actually been fixing an artefact of the smoothing parameter selection method, rather than a real deficiency in the confidence interval methods.

4.2.5 Component interval computation

As already mentioned we can improve interval performance by changing the target of inference, a little. For each component smooth term, intervals can be constructed by applying identifiability constraints to all other model components, but allowing the component of interest to 'carry the intercept'. This yields improved coverage probabilities, especially for the near straight line/plane case (see section below), and in the authors' experience also produces intervals that correspond more closely to the way in which users tend to interpret component-wise intervals. Specifically, assume that the linear predictor is

$$\eta_i = \mathbf{X}^*_i \boldsymbol{\beta}^* + \sum_j f_j(x_{ji})$$

where $\sum_i f_j(x_{ji}) = 0 \; \forall j$. Suppose that $\boldsymbol{\beta}_j$ is the coefficient vector for f_j, so that the complete coefficient vector is $\boldsymbol{\beta} = (\boldsymbol{\beta}^{*\mathsf{T}}, \boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}}, \ldots)^{\mathsf{T}}$, with Bayesian covariance matrix $\mathbf{V}_{\boldsymbol{\beta}}$. Let $\alpha = \mathbf{1}^{\mathsf{T}}\mathbf{X}^*\boldsymbol{\beta}^*/n$ define the intercept. We are interested in

inference about $\tilde{f}_j(x_j) = \alpha + f_j(x_j)$.

Now let $\mathbf{X}$ denote the full $n \times p$ model matrix, and assume that its first p^* columns are given by $\mathbf{X}^*$. Further, let $\mathbf{X}^j$ denote the matrix $\mathbf{X}$ with all columns zeroed, except for those corresponding to the coefficients of f_j, and the first p^*: these are modified to $X^j_{ik} = \sum_i X_{ik}/n$ for $1 \leq k \leq p^*$. Then $\mathbf{X}^j$ is the model matrix for $\tilde{f}_j$, i.e. $[\tilde{f}_j(x_{j1}), \tilde{f}(x_{j2}), \ldots, \tilde{f}(x_{jn})]^\mathsf{T} \equiv \tilde{\mathbf{f}}_j = \mathbf{X}^j\boldsymbol{\beta}$. Then the Bayesian covariance matrix for $\tilde{\mathbf{f}}_j$ is just $\mathbf{X}^j\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{j\mathsf{T}}$ and interval computation is straightforward.

This fix is quite important since the near straight-line/plane case is the one that is most important to get right for model selection purposes, both for deciding on which terms should be removed from a model, and which should be treated purely parametrically. In other words, the problem fixed by the preceding theory is of some importance in practice.

Notice that this proposal is not the same as basing intervals on the standard error of overall model predictions, with all but the covariate of interest held constant (e.g. Ruppert *et al.*, 2003, ch. 8). Such intervals are typically much wider than our proposal. Also note that trying to reduce the near straight-line problem by using alternative constraints will unnecessarily widen confidence intervals, since only the given constraint results in $\hat{\mathbf{f}}_j \perp \mathbf{1}$.

4.3 Simulation study

A Monte Carlo simulation study was conducted to compare the practical performance of several component-wise variable width confidence intervals. Specifically, based on equation (4.7), three kinds of intervals were considered:

1. Standard Bayesian intervals: the Wahba/Silverman type Bayesian intervals for smooth functions subject to identifiability constraints, derived employing $C_i^{-1} = [\mathbf{V}_\mathbf{f}]_{ii}$, setting $\widehat{\mathbb{E}(B)} = 0$, and using the mean squared error estimate (4.2) as well as the scale parameter estimate (4.3). That is,

$$\hat{f}(x_i) \pm z_{\alpha/2}\hat{\sigma}\sqrt{[\mathbf{V}_\mathbf{f}]_{ii}}.$$

2. Bayesian intervals with intercept: the same intervals as the previous ones but re-defining the model matrix $\mathbf{X}$ as discussed in section 4.2.5.

3. Alternative intervals with intercept: these intervals are derived using the model matrix of section 4.2.5, $C_i^{-1} = [\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}$, the mean bias estimate (4.6),

and the mean squared error estimate (4.8). That is,

$$\hat{f}(x_i) - \widehat{\mathbb{E}(B)}\sqrt{[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}} \pm z_{\alpha/2}\sqrt{(\widehat{\mathbb{E}(M)} - \widehat{\mathbb{E}(B)^2})[\mathbf{V}_{\hat{\mathbf{f}}}]_{ii}}$$

where the scale parameter contained in $\mathbf{V}_{\hat{\mathbf{f}}}$ is estimated according to (4.4).

For the cases in which non-Gaussian data were considered, the substitutions of section 4.2.3 were carried out. Under a wide variety of settings, and employing the test functions displayed in Figures 4-1 and 4-2, the confidence intervals were compared in terms of coverage probabilities.

4.3.1 Design and model fitting settings

Two different linear predictors have been used for the simulation study. The first one was made up of a parametric component, z, plus four one-dimensional test functions (see Figure 4-1 and Table 4.1)

$$\eta_{1i} = \alpha + z_i + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}).$$

The second one was made up of a parametric component plus three two-dimensional test functions (see Figure 4-2 and Table 4.1)

$$\eta_{2i} = \alpha + z_i + f_5(x_{5i}, x_{6i}) + f_6(x_{7i}, x_{8i}) + f_7(x_{9i}, x_{10i}).$$

Uniform covariates on (0, 1) with equal correlations were obtained using the algorithm from Gentle (2003) (see Section 3.6.1). This procedure was employed to obtain correlation among all covariates involved in a linear predictor. The cases in which ρ was set to 0, 0.5 and 0.9 were considered. The functions were scaled to have the same range and then summed. Data were simulated under four error models at each of three signal to noise ratio levels, at each of three sample sizes, $n = 200, 500, 2000$. The three signal to noise ratio parameters were chosen so that the squared correlation coefficient between μ_i and y_i was about 0.4, 0.55, and 0.7 respectively. See Table 4.2 for further details. One-thousand replicate data sets were then generated at each sample size, distribution and error level combination, and generalized additive models fitted using penalized thin regression splines (Wood, 2003) based on second-order derivatives and with basis dimensions equal to 10, 10, 10 and 20 respectively for the first linear predictor, and with basis dimensions equal to 20, 20 and 50 for the second linear predictor. The smoothing parameters were chosen by GCV in the

normal and gamma cases and by generalized AIC in the Poisson and binomial cases (Wood, 2004, 2008). For each replicate, 90%, 95% and 99% confidence intervals for the linear predictor as well as for each smooth function, evaluated at the simulated covariate values, were obtained. Then for each value of ρ, test function, sample size, signal to noise ratio, error model and $1-\alpha$ level, an overall mean coverage probability from the resulting 1000 across-the-function coverage proportions was taken, and its standard deviation calculated.

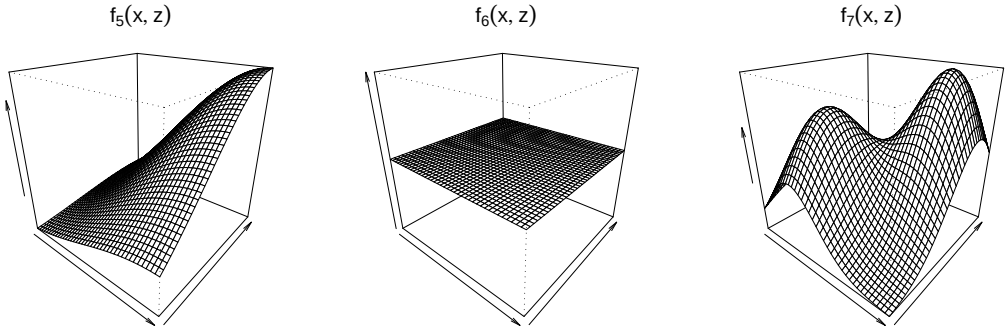

Figure 4-2: The three two-dimensional test functions used in the linear predictor $\eta_{2,i}$.

$$
\begin{aligned}
f_1(x) &= 2\sin(\pi x)\\
f_2(x) &= e^{2x}\\
f_3(x) &= 0\\
f_4(x) &= x^{11}\{10(1-x)\}^6 + 10(10x)^3(1-x)^{10}\\
f_5(x,z) &= 0.7e^{-\{(-3x+3)^2+0.7(3z-3)^2\}/5}\\
f_6(x,z) &= 0\\
f_7(x,z) &= 0.075e^{-\frac{(x-0.3)^2}{0.3^2}-(z-0.3)^2} + 0.094e^{-\frac{(x-0.8)^2}{0.3^2}-\frac{(z-0.8)^2}{0.4^2}}
\end{aligned}
$$

Table 4.1: Test function definitions. f_1 - f_4 are plotted in Figure 4-1, and f_5 - f_7 in Figure 4-2.

	binomial	*gamma*	*Gaussian*	*Poisson*
$g(\mu)$	*logit*	*log*	*identity*	*log*
$l \leq \eta \leq u$	$[0.02, 0.98]$	$[0.2, 3]$	$[0, 1]$	$[0.2, pmax]$
s/n	n_{bin}= $1, 3, 5$	$\phi = 0.6, 0.4, 0.2$	σ= $0.4, 0.2, 0.1$	$pmax = 3, 6, 9$

Table 4.2: Observations were generated as described in Table 3.2. The linear predictor for the binomial case was scaled to produce probabilities in the range $[0.02, 0.98]$; observations were then simulated from binomial distributions with denominator n_{bin}. In the gamma case the linear predictor was scaled to have range $[0.2, 3]$ and three levels of ϕ used. For the Gaussian case normal random deviates with mean 0 and standard deviation σ were added to the true expected values, which were then scaled to lay in $[0, 1]$. The linear predictor of the Poisson case was scaled in order to yield true means in the interval $[0.2, pmax]$.

4.3.2 Coverage probability results

To save space, not all simulation results are shown. Instead the most significant examples chosen. We have been careful to choose plots that are representative of the results in general, so that intuition gained from the plots shown fairly reflects the intuition that would be gained from looking at all the plots from the study.

Figures 4-3, 4-4 and 4-6 show coverage probability results when covariates are moderately correlated ($\rho = 0.5$) and data are generated using linear predictor $\eta_{1,i}$. Figure 4-8 refers to the case when employing $\eta_{2,i}$. Figures 4-5 and 4-7 serve to show the impact that covariate correlation has on confidence interval performance.

Results for standard Bayesian intervals

The standard Bayesian intervals yield realized coverage probabilities that appear to be fairly close to their nominal values in most cases, even at small sample sizes when the signal-to-noise ratio is not too low. The exception is for function $f_2(x)$ where the smoothing parameter methods tend to select more straight lines than they should really be selecting. However, as explained in section 4.2.4, it is because of the identifiability constraint that the confidence intervals are too narrow, and the coverage probabilities are poor as a consequence of the violation of the assumption that B is less than V. It is worth observing that when not much information is present in the data, it is likely to come up with straight line estimates for smooth functions like $f_2(x)$. Yet, it is the identifiability constraint that does have a detrimental effect on confidence interval performance, and this is why confidence intervals should really be constructed including the intercept of the model.

Results for Bayesian intervals 'with intercept'

As suggested in section 4.2.4, interval performance can be improved by allowing the component of interest to 'carry the intercept'. When this is done, the component-wise Bayesian intervals produce overall better coverage probabilities, especially for the near straight line/plane case; see functions $f_2(x)$ and $f_5(x, z)$.

Results for alternative intervals 'with intercept'

The alternative intervals also exhibit good component-wise interval performance but with slower convergence rate: this is probably because of the higher number of component quantities that have to be estimated for these intervals.

Impact of covariate correlation

Covariate correlation has an impact on confidence interval performance. Figures 4-5, 4-6 and 4-7 show the coverage probabilities obtained for Poisson data when correlated covariates were generated using $\rho = 0$, 0.5 and 0.9 respectively. It can be seen that although mild correlation does not spoil the coverage probabilities, a heavier one degrades some of them. As explained in section 4.2.2, the confidence intervals exhibit good practical performance if the smoothing parameters are chosen such that the estimated smooth components are not too heavily oversmoothed, but this is less likely to happen when covariates are heavily correlated. In this case, when $\rho = 0.9$ the covariates are quite confounded and this may lead to heavy oversmoothing for some component function of a GAM (often with some other component correspondingly undersmoothed, but this is less detrimental). Looking at the coverage probability results from the standard Bayesian intervals and comparing the results across the different values of ρ reveal that the coverages for $f_1(x)$ and $f_2(x)$ worsen. This means that for these two functions a smoother estimate is often selected, and this is why interval performance degrades. Specifically, the major failure evident from our results occurs for $f_2(x)$ where a substantial number of straight line estimates are selected. On the other hand, it would be rather remarkable if a smoothing method could select the right curve for a function which is so close to a straight line, when $\rho = 0.9$. The interval performance for $f_3(x)$ and $f_4(x)$ does not change significantly. Concerning $f_3(x)$, straight line estimates and wiggly ones are selected, but always centered on the right level. For this reason, interval performance does not degrade. Regarding $f_4(x)$, the smoothing parameter can still be reasonably estimated as a result of the fact that the degree of complexity of this function is not low and hence heavy oversmoothing is not likely to occur, even when $\rho = 0.9$. Obviously, the Bayesian intervals 'with intercept' exhibit better coverage probabilities. Notice, once again, that as the information contained in the data increases interval performance improves.

The following argument explains the 'strange' downward coverage probability pattern that can be observed for a near-straight-line function like $f_2(x)$

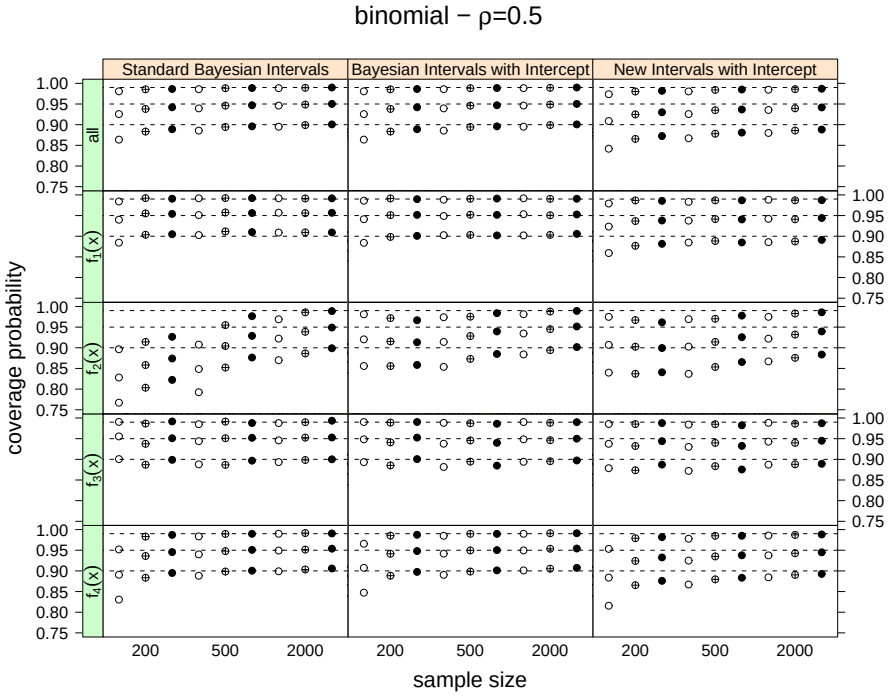

Figure 4-3: Coverage probability results for binomial data generated using $\eta_{1,i}$ as a linear predictor. Covariate correlation was equal to 0.5. Details are given in section 4.3. $\circ$, $\oplus$ and $\bullet$ stand for high, medium and low noise level respectively. Standard error bands are not reported since they are smaller than the plotting symbols. Notice the improvement in the performance of the component-wise intervals for $f_2(x)$, when the intercept is included in the calculations.

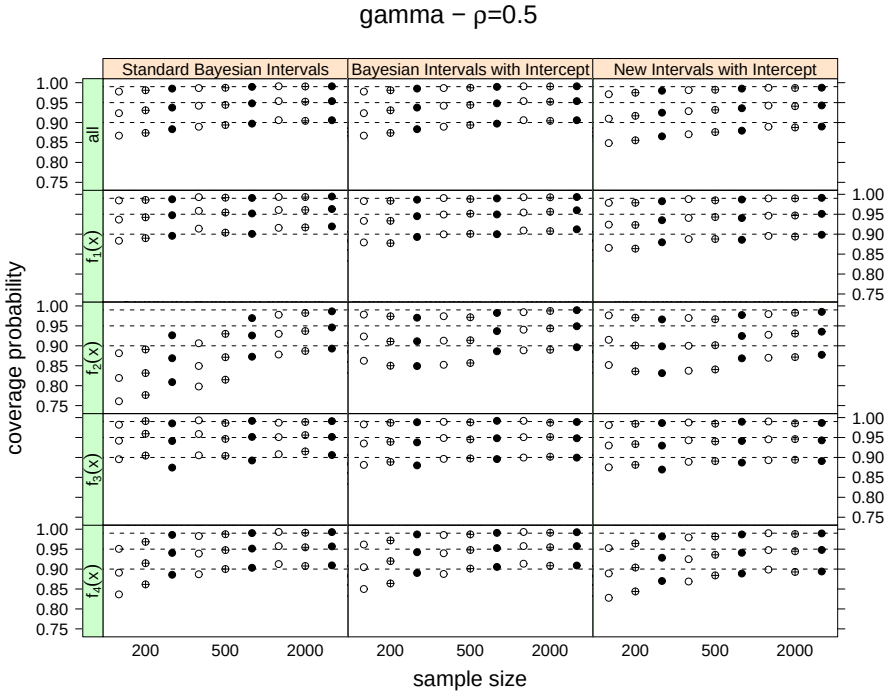

Figure 4-4: Coverage probability results for gamma data. Details are given in the caption of Figure 4-3.

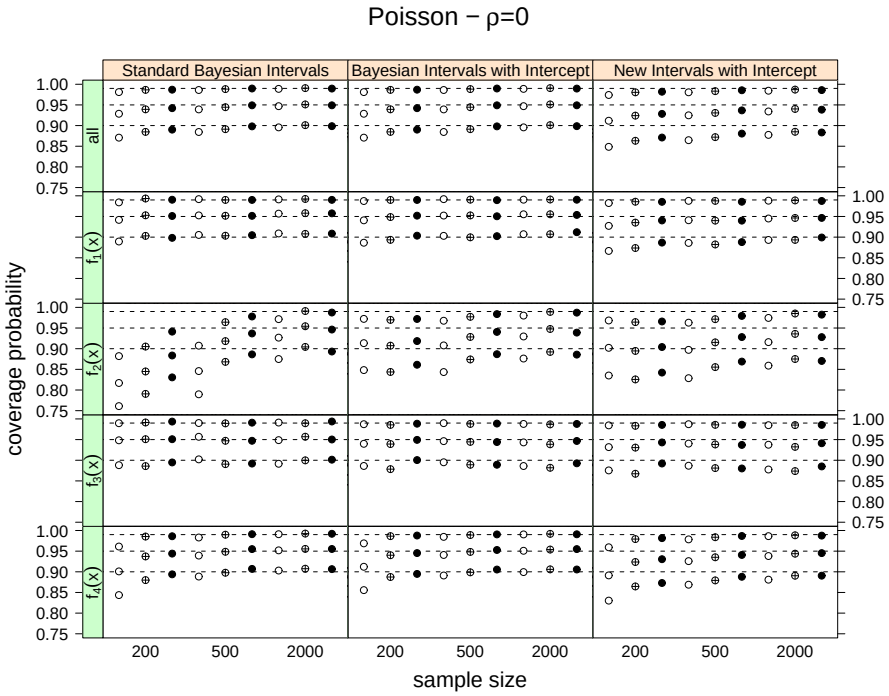

Figure 4-5: Coverage probability results for Poisson data for the case in which correlated uniform covariates were obtained setting $\rho = 0$. Details are given in the caption of Figure 4-3.

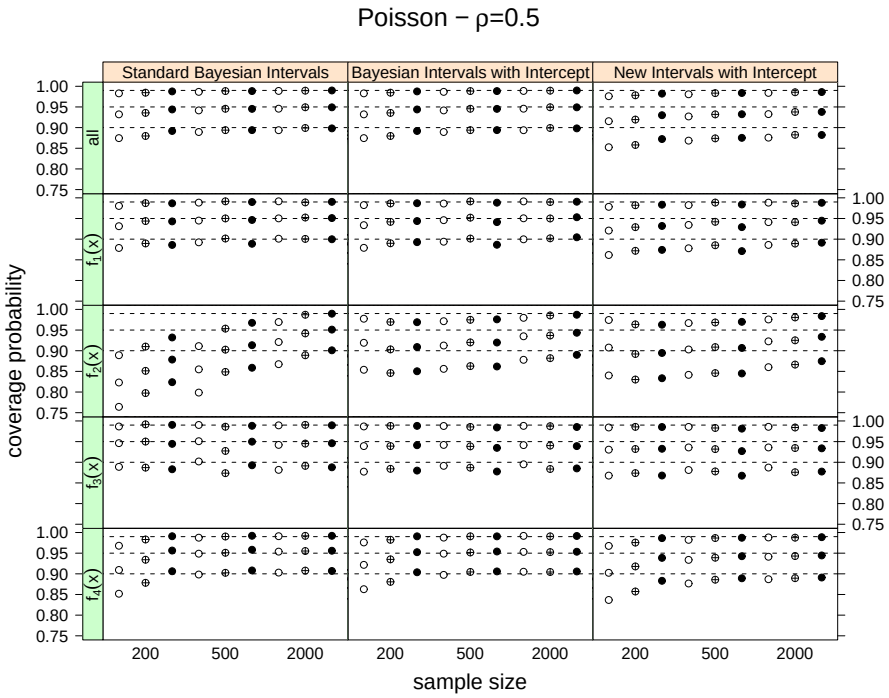

Figure 4-6: Coverage probability results for Poisson data for the case in which ρ was set to 0.5. Details are given in the caption of Figure 4-3.

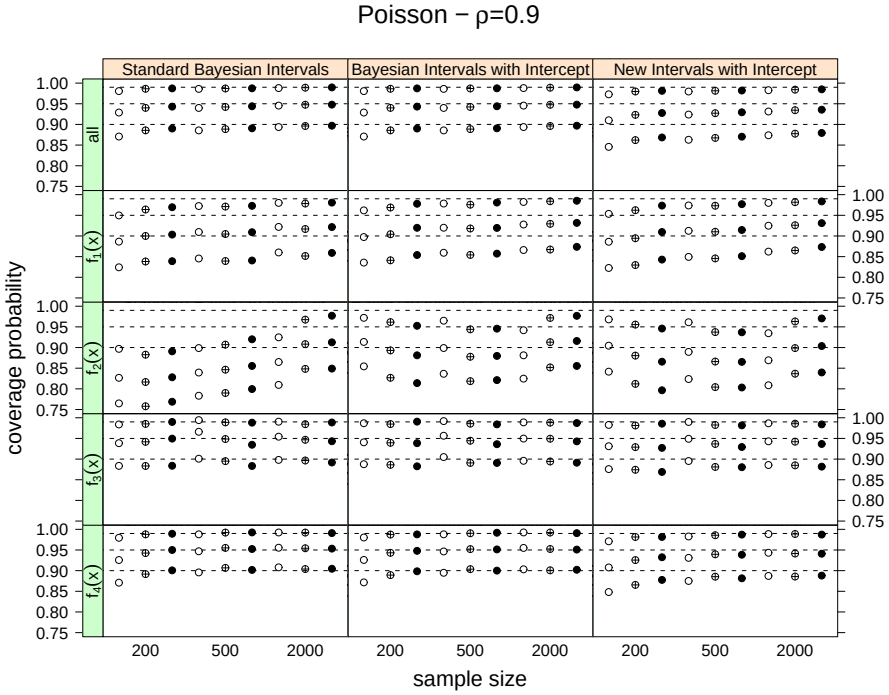

Figure 4-7: Coverage probability results for Poisson data for the case in which ρ was set to 0.9. Details are given in the caption of Figure 4-3. Notice how the confidence interval performance for $f_1(x)$ and $f_2(x)$ degrades when oversmoothing, due to high covariate correlation, occurs.

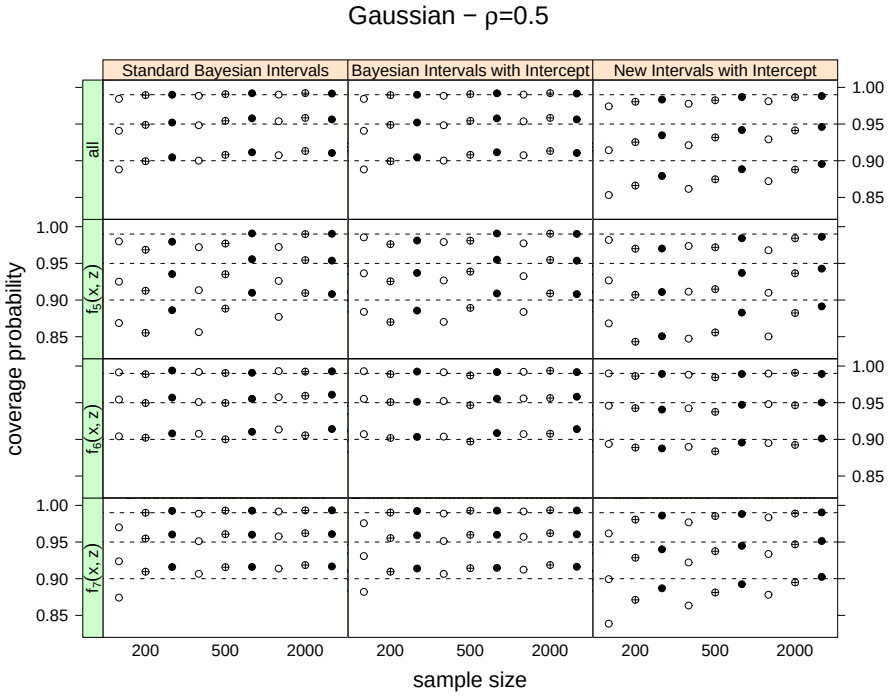

Figure 4-8: Coverage probability results for Gaussian data generated using $\eta_{2,i}$ as a linear predictor. Covariate correlation was equal to 0.5. Details are given in the caption of Figure 4-3. Notice the improvement in the performance of the intervals for $f_5(x, z)$, when the intercept is included in the calculations.

(see Figure 4-7 for example). Let us consider a situation in which a straight line estimate is always selected for $f_2(x)$. The confidence intervals for the case in which $n = 200$ and the signal-to-noise ratio is low will be wider than the ones obtained for the case in which $n = 200$ but the signal-to-noise ratio is high. This means that, for some functions like the one being analyzed here, oversmoothing can become unimportant since the intervals will be so wide that they will do a better job than the ones calculated when more information is contained in the data. In other words, when the complexity of a function is not high and provided data are noisy enough to get very wide confidence intervals, the violation of the assumption that B is less than V can become irrelevant. However, as the signal-to-noise ratio and the sample size increase the intervals become narrower and the conclusions drawn in this results section apply.

4.4 Discussion

We have shown by simulation and extension of Nychka's (1988) analysis, that the Wahba/Silverman type Bayesian intervals for the components of a penalized regression spline based GAM have generally good frequentist properties, across-the-function. As simulation evidence and our theoretical arguments suggest, the exception occurs when components estimated subject to identifiability constraints have interval widths vanishing somewhere as a result of heavy smoothing. Coverage probabilities can be improved if intervals are only obtained for unconstrained quantities, such as a smooth component plus the model intercept. The theoretical results also allow us to define alternative intervals when a frequentist approach is adopted.

The results make a novel contribution in extending Nychka's argument to the GAM component case thereby pinpointing the circumstances in which the intervals will and will not work, and explaining the role of smoothness selection as well as smoothing parameter uncertainty. The findings are backed up with quite extensive simulation testing of the finite sample performance of the three types of confidence intervals considered here. Specifically, our Monte Carlo investigation allowed us to compare the component-wise intervals under a wide variety of setting. In this respect, our simulation results show under which circumstances these intervals can reliably represent smooth term uncertainty, given the smoothing parameter selection methods employed here, and provide in fact applied researchers with some guidance about the practical performance of the intervals.

As evidenced by our theoretical and simulation results, these pointwise confidence intervals reliably represent the sample variability associated with GAM smooth components to the extent that their coverage probabilities, averaged across the observation points, are close to the nominal level. The disadvantage of this approach is that, because the intervals are *only* valid in an average sense, one can not completely rely on them for inferential purposes. Moreover, as pointed out by Hastie and Tibshirani (1990, p. 62), a confidence band is limited in the amount of information it provides. This is because, a confidence band represents just a projection of the confidence sets for $\mathbf{f}$ in n-dimensional space. In other words, as illustrated in Figure 4-9, the functions in the confidence set might exhibit features that are not necessarily enforced by a confidence band.

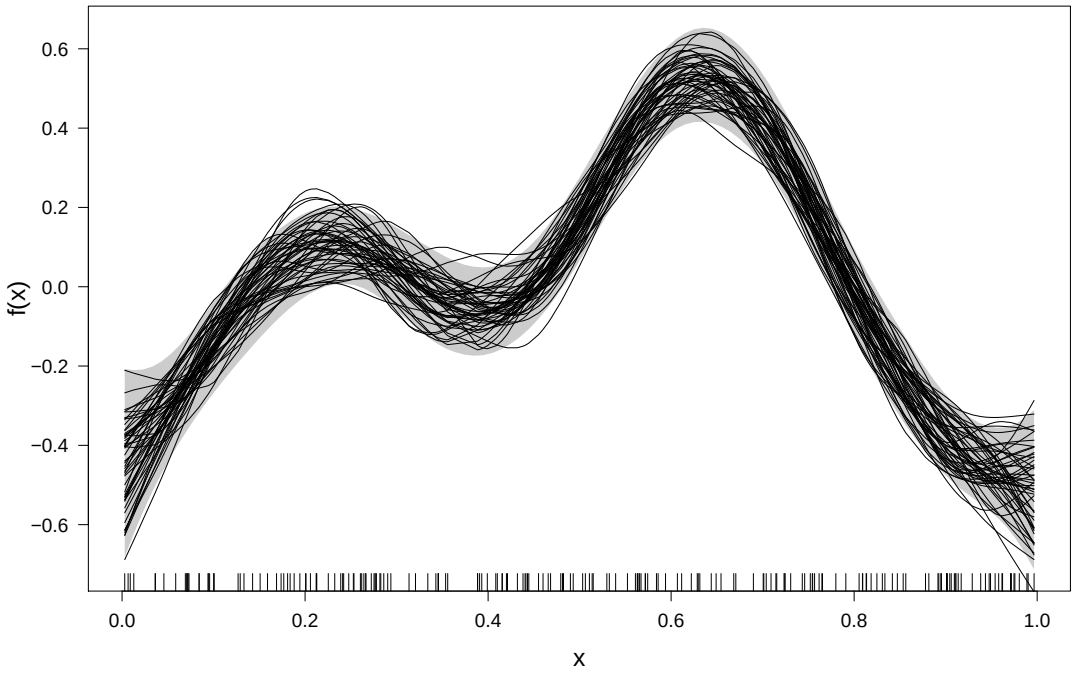

Figure 4-9: Smooths corresponding to 50 draws from (2.5) obtained from fitting an additive model to 200 observations generated as $Y_i = f_4(x_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.3$, and x is a uniform covariate on the unit interval. The function f_4 is displayed and defined in Figure 4-1 and Table 4.1, respectively. The shaded regions represent 95% Bayesian intervals from the fitted model.

Figure 4-9 shows that the intervals provide information about the overall pattern of the curves, although they do not say much about other structure in them. Despite this, we can be confident, for example, that the whole curve has at least 1 peak and that the true relationship is nonlinear since the intervals definitely do not contain a straight line. However, for the case in which a straight

line can be drawn within the variability bands of the estimated smooth, it is not possible to conclude that the true relationship is linear given the arguments in the previous paragraph. So, these intervals are a useful guide to help model selection, but one can not completely rely on them alone to infer certain features in the smooth estimate.

Appendix: The relative magnitude of B and V

Our analysis suggests that the intervals will only work well if B is of substantially smaller magnitude than V. Nychka (1988) provided simulation evidence that this would usually be the case for univariate spline smoothing. This appendix provides some simulation evidence that this also holds in the component wise case, as well as providing a limited theoretical exploration of the issue.

Empirical insight into the relative magnitude of B and V can be gained by examining the percentage mean squared bias and mean variance of the smooth components of a GAM. Table 4.3 reports these two quantities, calculated according to the definitions in Section 4.2.2 with $C_i^{-1} = [\mathbf{V}_{\mathbf{f}_j}]_{ii}$ for each smooth component j, for some of the cases considered in our simulation study. Overall, the mean squared bias is of substantially smaller magnitude than the mean variance, except for f_2 where the opposite happens. As explained in Section 4.2.4, when a true function is close to a term in the null space of the component's penalty, and the corresponding smooth function is estimated as a straight line but subject to an identifiability constraint, the assumption that B is less than V will fail. These results are consistent with our simulations, where poor coverage probabilities were obtained for f_2. Notice that this is not the result of failing to meet the assumption that the mean squared bias of the parameters of a smooth is less than its mean variance. The fact that B is greater than V is due to an 'artifact' induced by the identifiability constraint, an issue which can be explored only if using the extended Nychka argument of this chapter, where some constants C_i have to be used to derive non-constant width intervals for GAM components. Recall that as a remedy, improved coverages are obtained if each term's interval is computed as if it alone were unconstrained, and identifiability was obtained by constraints on the other model terms.

Some limited theoretical insight into the relative magnitude of B and V can be gained by examining the mean squared bias and mean variance of the parameters of a smooth, in the Demmler-Reinsch (or 'natural') parameteriza-

	binomial		*gamma*		*Gaussian*		*Poisson*	
$function$	$\bar{b^2}^*$	$\bar{v^2}^*$	$\bar{b^2}^*$	$\bar{v^2}^*$	$\bar{b^2}^*$	$\bar{v^2}^*$	$\bar{b^2}^*$	$\bar{v^2}^*$
f_1	8.6	91.4	12.8	87.2	9.0	91.0	17.0	83.0
f_2	76.5	23.5	90.4	9.6	62.5	37.5	94.9	5.1
f_3	0.9	99.1	0.1	99.9	0.1	99.9	0.3	99.7
f_4	15.4	84.6	26.5	73.5	13.4	86.6	17.5	82.5

Table 4.3: Percentage mean squared bias ($\bar{b^2}^*$) and mean variance ($\bar{v^2}^*$) results for the smooth components of GAMs fitted to data simulated from four error models at medium noise level. Covariate correlation and sample size were 0.5 and 200 (see Section 4.3 for further details). $\bar{b^2}^* = \bar{b^2}/(\bar{b^2}+\bar{v^2})*100$ and $\bar{v^2}^* = \bar{b^2}/(\bar{b^2}+\bar{v^2})*100$, where $\bar{b^2}$ and $\bar{v^2}$ were calculated following the definitions in Section 4.2, with $C_i^{-1} = [\mathbf{V}_{\mathbf{f}_j}]_{ii}$ for each smooth component j. Notice that the $B < V$ assumption is comfortably met for all terms except for f_2, which is the problematic case in the first columns of Figures 4-3 - 4-7.

tion as in Section 4.2.1. For simplicity, and without loss of generality, let us consider the case of a smooth estimated by a penalized least squares fit to data with variance σ^2. In the new parameterization, provided the model is a reasonable fit, it is easy to find expressions for the mean variance and mean squared bias of the coefficients,

$$\bar{v^2} = \frac{1}{p}\sum_k \frac{\sigma^2}{(1+D_{kk})^2}$$

and

$$\bar{b^2} = \frac{1}{p}\sum_k \frac{\sigma^2 D_{kk}}{(1+D_{kk})^2}.$$

If M is the dimension of the null space of the smooth penalty then M of the D_{kk} will be zero. These unpenalized coefficients contribute $M\sigma^2/p$ to the mean variance, and nothing at all to the mean squared bias. So for $\bar{b^2}$ to exceed $\bar{v^2}$ we require the remaining terms in the $\bar{b^2}$ to tend to exceed the corresponding terms in $\bar{v^2}$ and to be substantial relative to $M\sigma^2/p$. This is difficult to achieve. The largest term in $\bar{b^2}$ is bounded above by $\sigma^2/(4p)$, and is of the same size as the corresponding term in $\bar{v^2}$. Later terms in $\bar{b^2}$ do become larger than the corresponding terms in $\bar{v^2}$, but at the same time they rapidly become very small relative to $M\sigma^2/p$. Given that different smoothers will have different eigen spectra, it is difficult to make this argument more precise, but it does go some way to explaining the simulation results, and also makes the interesting prediction that the B less than V assumption will tend to hold more strongly as the dimension of the penalty null space increases.

Chapter 5

Practical Variable Selection

In this chapter, we work under the assumption of the absence of unmeasured confounders and consider the problem of variable selection within the class of GAMs, when there are many predictors to choose from, but the number of predictors is still somewhat smaller than the number of response observations. Two very simple but effective shrinkage methods and an extension of the nonnegative garrote estimator are introduced. The proposals avoid having to use nonparametric testing methods for which there is not a general reliable distributional theory. Moreover, component selection is carried out in one single step as opposed to many selection procedures which involve an exhaustive search of all possible models. The empirical performance of the proposed methods is compared to that of some available techniques via an extensive simulation study. The results show under which conditions one method can be preferred over another, hence providing applied researchers with some practical guidelines. The procedures are also illustrated analysing data on plasma beta-carotene levels from a cross-sectional study conducted in the United States.

5.1 Introduction

Variable selection is an important area of research. From a pragmatic point of view, it aims at determining which covariates have the strongest effects on the response of interest, whereas from a statistical perspective it represents a means to achieve a balance between goodness of fit and parsimony. In other words, by effectively identifying a subset of important covariates, variable selection can both enhance model interpretability and improve prediction accuracy. Methods such as subset selection, stepwise procedures and shrinkage

methods can be employed (see Guisan *et al.* (2002) for an overview). Subset selection chooses a model containing a subset of predictors according to some criterion, but all possible subset models have to be explored and hence it can become computationally expensive as the number of predictors increases. Stepwise procedures do not make use of all possible models, therefore reducing computational cost, but they might be inconsistent given the dependence on the path chosen through the variable space. The additional drawback of these procedures is that if we perform variable selection and then hypothesis testing using the selected model, the p-values associated with the model terms will not be strictly correct since they neglect variable selection uncertainty. Shrinkage methods are becoming popular in the statistical literature. In fact, they have proved to be a valid alternative to the procedures above in terms of stability and prediction. Moreover, shrinkage procedures are continuous processes since variable selection is carried out in one single step as opposed to subset selection and stepwise algorithms (Hesterberg *et al.*, 2008).

For the additive model case, subset selection and stepwise procedures can be carried out using, for instance, the Akaike Information Criteria (e.g. Greven and Kneib, 2009; Wager *et al.*, 2007). A number of hypothesis testing approaches have also been proposed, which do model selection in terms of either choosing between linear and more general smooth term alternatives or dropping unimportant components from the model (Cantoni and Hastie, 2002; Hastie an Tibshirani, 1990; Kauermann *et al.*, 2009; Kauermann and Tutz, 2001; Scheipl *et al.*, 2008; Wood, 2006). Despite the fact that some testing methods have been introduced in the GAM context (Hastie an Tibshirani, 1990; Wood, 2006), a *general* reliable distributional theory for the smooth terms of a GAM has not been developed to date. Shrinkage methods for linear models and GLMs, which simultaneously address estimation and variable selection, have been proposed (e.g. Breiman, 1995; Efron *et al.*, 2004; Tibshirani, 1996; Tutz and Binder, 2007; Yuan and Lin, 2006; Zou, 2006). Some algorithms have also been introduced to achieve component selection within additive models (Avalos *et al.*, 2007; Belitz and Lang, 2008; Buhlmann and Yu, 2003; Cantoni *et al.*, 2010; Lin and Zhang, 2006; Xue, 2009) and GAMs (see Zhang and Lin (2006) and references therein). However, for the GAM case, the boosting technique of Tutz and Binder (2006) and a generalization of the approach of Belitz and Lang (2008) seem to be the only fitting procedures available.

In this chapter, we focus on smooth component selection when dealing with GAMs by pursuing a shrinkage approach. As mentioned earlier on, this approach is appealing since it has the properties of stability and predic-

tion, and variable selection can be carried out in one single step. Furthermore, it avoids having to use testing methods for which there is not a general distributional theory. We propose two effective shrinkage methods and extend the nonnegative garrote estimator to achieve component selection within GAMs. Their empirical performance is compared to that of some available methods via an extensive simulation study. The procedures are also illustrated by analysing data on plasma beta-carotene levels from a cross-sectional study conducted in the United States. Note that we concentrate throughout on the case in which we need to select from a small to moderate sized set of potential predictors. In part this is due to method constraints. However, we also believe that in practice it is not very common that the modeller *does not* know which of a very large number of predictors is important, but *does* know that an additive structure gives an appropriate model.

5.2 Methods

Smoothing parameter estimation can select between models of different complexity, but it does not usually remove a smooth term from the model altogether. This is because the usual penalty of a spline basis does not allow for the shrinkage of the functions that are in the penalty null space (and for the most useful smoothers the null space has a dimension greater than zero). The proposals in Sections 5.2.1 and 5.2.2 are based on the idea that the space of a spline basis can be decomposed in the sum of two component, one associated with the functions in the penalty null space and the other with the penalty range space. The smoothing penalty shrinks functions in the range space (to zero if the smoothing penalty is high enough), but leaves the function component in the null space untouched. So to have the possibility of shrinking the whole spline term to zero, it is necessary to penalize the null space. As an alternative approach, the method introduced in Section 5.2.4 does not require the use of such a decomposition, and is based on the idea of shrinking the smooth function estimates obtained from a standard fitted GAM. The proposed methods have the properties of subset selection, but with the advantage that variable selection can be achieved in one single step. Section 5.2.5 presents some of the available alternatives, whereas Section 5.2.6 briefly discusses multiple smoothing parameter estimation which is crucial for the variable selection methods to work well.

5.2.1 Double penalty approach

The generic smoothing penalty matrix $\mathbf{S}_j$ associated with a smooth term of a GAM can be decomposed as

$$\mathbf{U}_j \Lambda_j \mathbf{U}_j^{\mathsf{T}}, \tag{5.1}$$

where $\mathbf{U}_j$ is an eigenvector matrix associated with the j^{th} smooth function, and Λ_j the corresponding diagonal eigenvalue matrix. The fact that a part of the spline basis space deals with the penalty null space implies that Λ_j contains zero eigenvalues. This may be problematic if variable selection has to be carried out. For instance, let us assume that the j^{th} smooth component is a nuisance function, and that we use a penalty matrix as defined above during the model fitting process. Even if λ_j goes to infinity there will not be any guarantee that the smooth term will be suppressed completely (i.e. estimated as zero).

In order to circumvent this difficulty, we can produce an extra penalty which penalizes only functions in the null space of the penalty, so that a smooth component can be completely removed. Specifically, let us consider decomposition (5.1). An extra penalty can be formed as follows

$$\mathbf{S}_j^* = \mathbf{U}_j^* \mathbf{U}_j^{*\top},$$

where $\mathbf{U}_j^*$ is the matrix of eigenvectors corresponding to the zero eigenvalues of Λ_j. So a GAM can be fitted subjecting each component function to a double penalty of the form

$$\lambda_j \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j \boldsymbol{\beta} + \lambda_j^* \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j^* \boldsymbol{\beta}. \tag{5.2}$$

where both λ_j and λ_j^* will now have to be estimated. By introducing a penalty for the null space, smoothing parameter estimation (that is part of GAM fitting) can completely remove terms from the model.

To re-iterate the basic idea, any spline type smoother can be decomposed into two component functions: a component in the null space of the penalty, and a component in the range space of the penalty. The first term in (5.2) penalizes only function components in the range space, but can shrink these to zero, while the second term in (5.2) penalizes only function components in the null space, but can shrink these too to zero. For example, in the case of the usual cubic spline penalty, the second term in (5.2) would penalize straight line components to zero, while the first term would penalize (towards zero) function components representing departure from straight line behaviour.

This approach can be employed by setting the argument `select=TRUE` in

the `gam` function of the `R` package `mgcv`.

5.2.2 Shrinkage approach

As an alternative approach which avoids doubling the number of smoothing parameters to estimate, we can replace the smoothing penalty matrix $\mathbf{S}_j$ with

$$\tilde{\mathbf{S}}_j = \mathbf{U}_j \tilde{\Lambda}_j \mathbf{U}_j^{\mathsf{T}},$$

where $\tilde{\Lambda}_j$ is the same as Λ_j except for the zero eigenvalues which are set to ϵ, a small proportion of the smallest strictly positive eigenvalues of $\mathbf{S}_j$. This is exactly equivalent to fixing $\lambda_j^* = \epsilon\lambda_j$ in (5.2) and forces the eigenvalues of $\tilde{\mathbf{S}}_j$ associated with the penalty null space to be different from zero: hence smoothing parameter selection can remove a smooth component from the model altogether. We choose ϵ to be a small proportion of the smallest strictly positive eigenvalues of $\mathbf{S}_j$ to ensure that $\boldsymbol{\beta}^{\mathsf{T}}\tilde{\mathbf{S}}_j\boldsymbol{\beta} \approx \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta}$ for all the regression spline coefficients except those in or "close to" the null space of $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta}$. A value for ϵ equal to $1/10$ yields good results in terms of goodness of fit and shrinkage (see Section 5.3).

This approach can be employed specifying the GAM formula of `mgcv` as a function of shrinkage smoothers. Two classes are implemented: `cs` and `ts`, based on cubic and thin plate regression spline smoothers, respectively.

5.2.3 Shrinkage penalty interpretation

Estimation by penalized likelihood with GCV or REML type smoothing parameter estimation can be viewed as an empirical Bayes procedure, with the penalties corresponding to (usually improper) Gaussian priors on the spline coefficients (the basic idea goes back to Kimeldorf and Wahba (1970)). In this case the $\mathbf{S}_j$ are viewed as prior precision matrices. It is the lack of full rank in the $\mathbf{S}_j$ that makes the prior improper, and this impropriety is a consequence of having a null space of functions that are treated as 'completely smooth' according to the penalty.

The proposals in sections 5.2.1 and 5.2.2 both remove the impropriety from the prior, since both $\lambda_j\mathbf{S}_j + \lambda_j^*\mathbf{S}_j^*$ and $\tilde{\mathbf{S}}_j$ are full rank. The double penalty approach of section 5.2.1 makes no prior assumption about the how much to penalize the null space relative to the range space for a term, but allows the smoothing parameter estimation to determine this from the data. On the other hand, the single penalty section 5.2.2 approach assumes that the null space

should be penalized less than the range space. This is a natural approach to take in some cases. In a cubic spline case for example, this would mean that as the smoothing parameter increases we would first penalize towards a straight line, and then shrink the line towards zero. However there is an inevitable arbitrariness about exactly how to weight the penalization of the two components, and if the data suggest penalizing the null space more heavily than the range space there is often no compelling reason for not doing so (perhaps the data contain no overall trend, for example).

In the work reported here we have employed the simplest null space penalties that will remove all impropriety from the priors and hence allow terms to be completely removed from the model. We have not considered whether some null space components should be penalized more than others. If the null space itself allows moderately complicated functions (e.g. the null space of a multi-dimensional thin plate spline penalty based on moderately high derivatives) then the modeller might want to impose some hierarchy within the null space basis coefficients penalizing some more than others. However it seems likely that the improvements achievable by doing this would be rather modest, and we will not pursue this further here. In any case, for 1 dimensional smooths using a cubic spline penalty, the null space is only one dimensional after the imposition of identifiability constraints on the GAM components, so the issue does not arise.

Finally, it is worth noting that after the imposition of standard identifiability constraints some smoothers have a zero dimensional null space corresponding to a *proper* prior for the coefficients, and can therefore be selected out of the model without the methods of sections 5.2.1 and 5.2.2. The obvious example is a spline, $f(x)$, based on the penalty functional $\int f'(x)^2 dx$. The null space of this penalty is the space of constant functions, which is eliminated from the space of the estimates by the identifiability constraint on f. Hence within the space of the identifiability constraint the penalty has full rank, and the corresponding prior of f is proper. Such terms can be used in `R` package `mgcv`, and priors of this sort have been used in a fully Bayesian context also (e.g. Chib and Greenberg (2007) use a random walk prior which shrinks towards the constant functions, and to zero after constraint). The difficulty with such terms is that the required low order penalization typically results in poor mean square error performance, in part because of undesirable properties such as tending to a constant function at the boundaries of the data.

5.2.4 Nonnegative garrote component selection

In order to identify the important smooth components of an additive model, Cantoni *et al.* (2010) and Yuan (2007) suggest employing the nonnegative garrote estimator, first proposed by Breiman (1995) in the linear model context, which has the properties of shrinkage and stability. The idea behind this is as follows. In a first step we obtain the original regression coefficient or smooth function estimates, depending on whether we are in a parametric or nonparametric context. We then shrink the model components by solving a constrained optimization problem.

The method presented here generalizes the nonnegative garrote estimator for additive models proposed by the authors above to the GAM context. First, we obtain some initial estimates for the smooth components of a model, $[\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots]$. Second, we solve the problem

$$\text{minimize } D(\boldsymbol{\eta}) \text{ w.r.t } \mathbf{d} \text{ subject to } \mathbf{d} \geq \mathbf{0} \text{ and } \mathbf{1}^{\mathsf{T}}\mathbf{d} = \gamma, \tag{5.3}$$

where $\boldsymbol{\eta} = \hat{\mathbf{F}}\mathbf{d}$, and $\hat{\mathbf{F}} = [\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots]$. The parameter vector $\mathbf{d}$ contains the shrinking coefficients, and γ is a tuning parameter. D is the usual model deviance defined as $2\phi\{l_{\text{sat}} - l(\boldsymbol{\eta})\}$, where $l(\boldsymbol{\eta})$ is the log-likelihood of the model with linear predictor $\boldsymbol{\eta}$ and l_{sat} the maximum value for the log-likelihood of the model with one parameter per datum.

For a given $\hat{\mathbf{F}}$ and γ, the estimated shrinking coefficients allow us to do variable selection. That is, if $\hat{d}_j = 0$ then the j^{th} component is viewed as uninformative and hence removed from the model. The shrinking coefficients also give information about the importance of each component in the model since some terms can be shrunk by some proportion $\hat{d}_j$, left unchanged (if $\hat{d}_j = 1$) or magnified (if $\hat{d}_j > 1$). The j^{th} final smooth component estimate is given by $\hat{\mathbf{f}}_j^* = \hat{\mathbf{f}}_j\hat{\mathbf{d}}$.

A small value for γ shrinks the d_j to zero and vice versa, hence affecting the final estimates. In fact this parameter has to be selected with a certain degree of accuracy. As suggested by Cantoni *et al.* (2010) and Yuan (2007), a 5-fold cross validation gives satisfactory results in terms of achieving a good balance between bias and variance, and it can be implemented using the following practical algorithm:

1. Split the data into subsets denoted by $I_1, \ldots, I_b, \ldots, I_B$, where b represents the subset considered and B the maximum number of subsets used for cross validation. In this case, $B = 5$.

2. Choose an equally spaced grid of values for γ in the interval $[0, n_c]$, where n_c indicates the total number of covariates used to fit the model.

3. For each value γ in the interval $[0, n_c]$

   (a) For each value of b

      i. Fit a standard GAM (employing `mgcv` or any other available smoothing package) using the sample containing all the observations except those in I_b. Then store the resulting smooth function estimates in $\hat{\mathbf{F}}^{[-I_b]}$.

      ii. Using the subset of observations as in i., solve (5.3) via iterative minimization of the problem

$$\|\sqrt{\mathbf{W}^{[k]}}(\mathbf{z}^{[k]} - \boldsymbol{\eta})\|^2 \text{ subject to } \mathbf{d} \geq \mathbf{0} \text{ and } \mathbf{1}^{\mathsf{T}}\mathbf{d} = \gamma,$$

      where k is the iteration index, $z_i^{[k]} = \eta_i^{[k]} + g'(\mu_i^{[k]})(y_i - \mu_i^{[k]})$, $\boldsymbol{\eta}^{[k]} = \hat{\mathbf{F}}^{[-I_b]}\mathbf{d}^{[k]}$, $\mu_i^{[k]} = g^{-1}(\eta^{[k]})$ and $W_{ii}^{[k]} = g'(\mu_i^{[k]})^{-2}V(\mu_i^{[k]})^{-1}$. In practice, this can be achieved by replacing, in the inner loop of `glm.fit` in `R`, the function `fit` with the function `pcls` for quadratic programming available in `mgcv`.

      iii. For each observation i in I_b, obtain $D_i(\hat{\eta}_i^{[-I_b]})$ where $\hat{\boldsymbol{\eta}}^{[-I_b]} = \hat{\mathbf{F}}\hat{\mathbf{d}}$, with parameter vector $\hat{\mathbf{d}}$ obtained in the previous two steps, and D_i is the contribution to the "full data" deviance that is associated with the i^{th} datum.

   (b) Calculate the cross validation predictive deviance

$$V_B(\gamma) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_b}\sum_{i\in I_b} D_i(\hat{\eta}_i^{[-I_b]}),$$

   where n_b represents the sample size for the subset I_b.

4. Obtain final smooth component estimates by repeating steps i. and ii. but using the whole sample, with value for γ selected to minimize $V_B(\gamma)$.

5.2.5 Some available alternatives

For the sake of comparison in our simulation study, we briefly describe some of the alternative available methods for carrying out component selection in a regression spline context.

Backward selection

A classic backward selection procedure for variable selection within GAMs can be employed. In order to implement the procedure we need to use some p-value definition. Here, we follow the approach of Wood (2006) described in Section 2.5.

A backward selection procedure using the p-value definition discussed in this section can be implemented by extracting the p-values for the smooth components of a GAM from the function `summary.gam` in `mgcv`.

GAM boosting

Binder and Tutz (2008) found that when a subset of a large number of predictors has to be selected and the degree of smoothness for the smooth components has to be chosen, generalized additive modeling by likelihood based boosting can achieve these two goals simultaneously (Tutz and Binder, 2006). They also provide simulation evidence that GAM boosting can be much better than alternative methods in very data poor settings, with many spurious covariates. This procedure iteratively fits a GAM by applying a 'weak learner' on the residuals of smooth components. The number of boosting steps is determined by a stopping rule such as cross-validation or an information criterion.

The `R` package `GAMBoost` can be used for fitting GAMs by likelihood based boosting, using 2^{nd} degree B-splines with 1^{st} order difference penalty as the default settings suggest. The function `optimGAMBoostPenalty` can be employed to select the optimal number of boosting steps.

Modified backfitting

Belitz and Lang (2008) developed an elegantly simple method for simultaneously estimating a model and selecting which components to include, based on a modification of backfitting, with computationally efficient sparse smoothers. As with backfitting, smooths are estimated by iteratively smoothing partial residuals, but at each step, rather than using a single fixed degrees of freedom smoother (as in classical backfitting), Belitz and Lang compute a number of alternative smoothers, corresponding to different degrees of freedom *plus* the null function corresponding to dropping the term altogether. To choose between these alternatives they compare a whole model GCV or AIC score for each alternative (using the current best estimate for the rest of the linear predictor). The method gains efficiency by using sparse smoothers (B-splines +

discrete penalties), and by using an additive approximation for the effective degrees of freedom for the whole model, required by the GCV or AIC score. The latter approximation is perhaps the method's main potential weakness: the approximation will deteriorate as covariate correlation increases, which has the potential to cause method performance to suffer.

Belitz and Lang (2008) only present the method in the additive context, but as they point out the extension to *generalized* additive models is straightforward, and is available in BayesX (`www.statistik.lmu.de/~bayesx/`), the command line version of which can be called from within `R`.

Parsimonious additive models

This approach, introduced by Avalos *et al.* (2007) for additive models, consists of separating the parametric and nonparametric parts of the smooth functions, and then fitting the parametric bit using a LASSO regression (Tibshirani, 1996) and the nonparametric part by solving a penalized least squares problem. A modified version of this approach can be implemented as follows:

1. Using thin plate regression splines (Wood, 2003, 2006), set up a matrix $\mathbf{X}^*$ containing the terms of the smooth functions which deal with the penalty null space.

2. Store the coefficients ($\hat{\boldsymbol{\alpha}}_{ols}$) obtained by fitting a linear regression of $\mathbf{y}$ on $\mathbf{X}^*$, where $\mathbf{y}$ represents the response vector.

3. By using the library `lars`, compute the lasso coefficients by minimization of the problem

$$\|\mathbf{y} - \mathbf{X}^*\boldsymbol{\alpha}\|^2 + \theta \sum_j^p |\alpha_j| \text{ w.r.t. } \boldsymbol{\alpha}.$$

   The tuning parameter θ is selected by K-fold cross validation using the function `cv.lars` in `lars`. Default settings suggest to set $K = 10$.

4. Compute the adjusted variable $\mathbf{y}^* = \mathbf{y} - \mathbf{X}^*\hat{\boldsymbol{\alpha}}_{ols}$, in order to ensure orthogonality between linear and nonlinear fits.

5. Set up a matrix $\mathbf{X}^+$ containing the terms of the smooth functions that deal with the penalty range space. Then, by using for instance `mgcv`, solve the

penalized least squares problem

$$\|\mathbf{y}^* - \mathbf{X}^+\boldsymbol{\beta}_+\|^2 + \sum_j \lambda_j \boldsymbol{\beta}_+^{\mathsf{T}} \mathbf{S}_j^+ \boldsymbol{\beta}_+ \text{ w.r.t. } \boldsymbol{\beta}_+,$$

where the $\mathbf{S}_j^+$ are the smoothing penalty matrices associated with the penalty range space.

6. Combine the results obtained in steps 3 and 5 and work out the final smooth function estimates.

This approach may look appealing since the LASSO regression can yield, via the use of a l_1 penalty, $\hat{\boldsymbol{\alpha}} = \mathbf{0}$. This means that, provided the smoothing parameters associated with the nuisance functions go to infinity, such a procedure can produce parsimonious additive models. However, as discussed in Hesterberg *et al.* (2008), the main drawback is that a linear term may be shrunk to zero while keeping the corresponding higher order components.

5.2.6 Smoothness selection

In order to implement the methods discussed in the previous sections, some multiple smoothing parameter selection procedure is needed. Importantly, for the methods to perform well it is crucial to use some stable and reliable computational method.

As explained in Section 2.3, multiple smoothing parameter estimation can be achieved via the computational methods of Wood (2006, 2008, 2010) implemented in `mgcv`. The simulation study in the next section will shed light on which criteria yield the best results.

5.3 Simulation study

A simulation study was conducted to compare the practical performance of the methods discussed in the previous section. Under a wide variety of settings, and employing a number of test functions, the procedures were compared in terms of shrinkage properties and a measure of fit.

5.3.1 Design and model fitting settings

The three linear predictors used for the simulation study are defined as

$$\eta_{1i} = \sum_{j=1}^{6} f_j(x_{ji}), \quad \eta_{2i} = f_7(x_{7i}, x_{8i}) + f_8(x_{9i}, x_{10i}) + f_9(x_{11i}, x_{12i})$$

$$\text{and } \eta_{3i} = f_1(x_{1i}) + f_3(x_{3i}) + f_4(x_{4i}).$$

The functions are displayed in Figure 5-1 and defined in Table 5.1. Uniform covariates on (0, 1) with equal correlations were obtained using the algorithm from Gentle (2003), as illustrated in the previous chapters. This procedure was employed to obtain correlation among all covariates involved in the linear predictor. The cases in which ρ was set to 0 and 0.9 were considered. The functions were scaled to have the same range and then summed. Data were simulated under the four error model - link function combinations detailed in Table 4.2, at each of three signal to noise ratio levels. 100 replicate data sets were then generated at each distribution and error level combination.

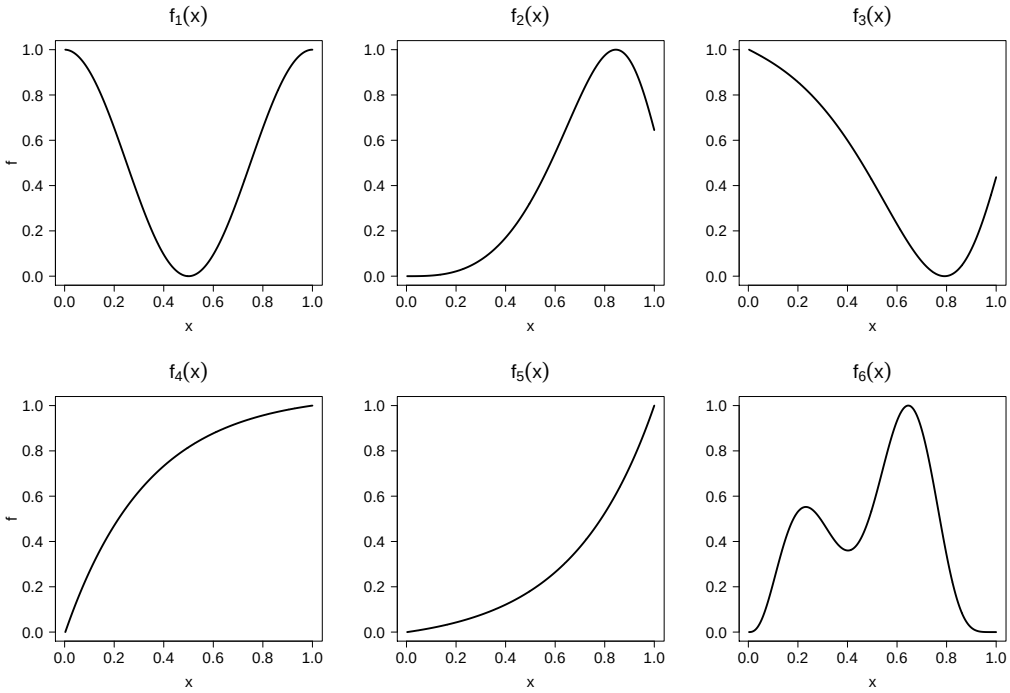

Figure 5-1: The test functions used to generate the datasets.

To maintain computational feasibility and because of limitations applying to some methods, the simulation study did not employ a completely factorial design. Instead it was conducted in the following four phases. In each phase 100 replicates of each combination of conditions were used, 3 noise levels were

$$
\begin{aligned}
f_1(x) &= 2\sin(\pi x) \\
f_2(x) &= e^{x^2} \\
f_3(x) &= -x \\
f_4(x) &= x^{11}\{10(1-x)\}^6 + 10(10x)^3(1-x)^{10} \\
f_5(x) &= 0.5\{x^3 + \sin(\pi x^3)\} \\
f_6(x) &= \cos(2\pi x) + \sin(\pi x) \\
f_7(x,z) &= 0.7e^{-\{(-3x+3)^2 + 0.7(3z-3)^2\}/5} \\
f_8(x,z) &= 0.39e^{\{-\frac{(x-0.3)^2}{0.25} - \frac{(z-0.3)^2}{0.25}\}} + 0.20e^{\{-\frac{(x-0.8)^2}{0.25} - \frac{(z-0.8)^2}{0.25}\}} \\
f_9(x,z) &= 0.16e^{\{-\frac{(x-0.3)^2}{0.25} - \frac{(z-0.3)^2}{0.25}\}} + 0.20e^{\{-\frac{(x-0.8)^2}{0.25} - \frac{(z-0.8)^2}{0.25}\}}
\end{aligned}
$$

Table 5.1: Test function definitions. f_1 - f_9 are plotted in Figure 5-1.

considered at each of $\rho = 0$ and 0.9:

1. Gaussian identity link models were compared for all methods, for η_1 with 6 nuisance covariates. Both REML and GCV smoothness selection were compared, and the sample size was 200. This phase suggested eliminating the Lasso&Splines method and the Belitz&Lang approach from the subsequent phases (the published versions of these do not treat the generalized case, although for Belitz and Lang, this is not a serious problem).

2. The other three distribution-link models were compared for η_1 with 6 nuisance covariates using all remaining methods. Again REML and GCV were compared where appropriate, and the sample size was 200. This phase suggested that GAM boosting is not competitive, at least for low numbers of nuisance variables. The combination of phases 1 and 2 suggested dropping GCV selection.

3. All distribution link models were compared for η_2 plus 6 nuisance covariates, using all remaining methods except GAM boosting. Smoothness selection was by REML for those methods were there is a choice. The sample size was 200.

4. All distribution link models were compared for all remaining methods including GAM boosting using η_3 and either 11 or 27 nuisance covariates. Sample sizes were 200 for 11 nuisance and 400 for 27. GAM boosting was re-considered here as this situation is the one were it is expected to be competitive.

For phases 1,2 and 4, all procedures, except for GAM boosting and the Belitz&Lang approach, were implemented using TPRSs (Wood, 2003) based

on second-order derivatives and with basis dimensions equal to 10. For phase 3 TPRSs with basis dimensions equal to 20, 20 and 50 were used.

The methods were compared in terms of shrinkage, and mean squared error (MSE) in predicting the linear predictors. To assess the shrinkage properties we used the false negative rate (i.e. rates at which influential covariates are not selected) for the variables in the linear predictors, and false positive rate (i.e. rates at which spurious terms are selected) for non influential covariates. The rates were calculated according to the MSEs rounded up to 7 digits. Notice that using as a criterion edf ≈ 0 led to the same results. Backward selection was carried out at the 5% significance level.

5.3.2 Results

To save space, only some of the most important examples are shown. The displayed plots have been chosen to be representative and to convey enough information to draw some general conclusions. Additional plots are available upon request.

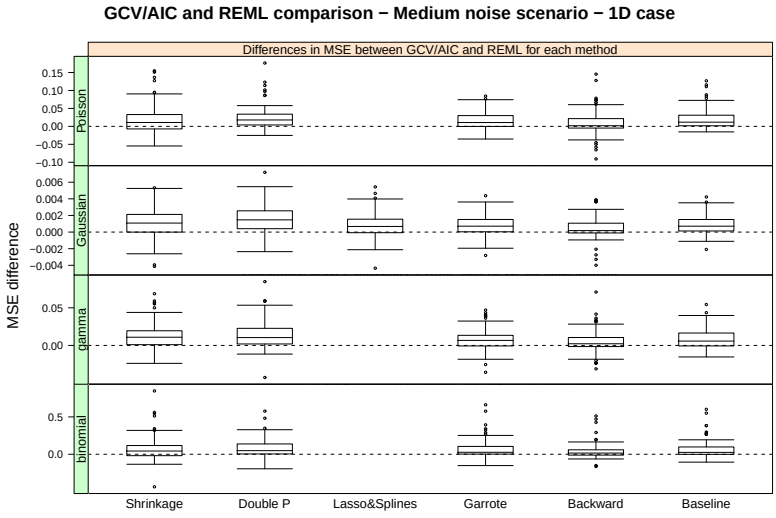

Figure 5-2: MSE comparisons between GCV/AIC and REML for four error distributions and methods discussed in Section 5.2, when using linear predictor η_1. Covariate correlation is 0 and the signal to ratio level is medium. Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 5.3.1. Boxplots show the distributions of differences in mean squared error between GCV/AIC and REML. In all cases a Wilcoxon signed rank test indicates the REML has lower MSE than GCV/AIC (p-value $< 10^{-2}$).

Figure 5-2 shows the difference in MSE between the same models estimated by GCV/AIC and by REML, for each error model and method combination, from phases 1 and 2 of the simulation study (employing linear predictor η_1).

Covariate correlation is 0. Missing box plots within the figures are because the method described in Section 5.2.5 only deals with additive models (and was anyway not competitive in phase 1). REML outperforms GCV/AIC smoothness selection; this suggests that REML allows for better smoothing parameter estimation, hence smooth term estimates are more accurate than when using GCV/AIC. The plots for cases in which $\rho = 0.9$ are omitted since they lead to the same conclusions. In the subsequent plots, we only report the results obtained when using REML.

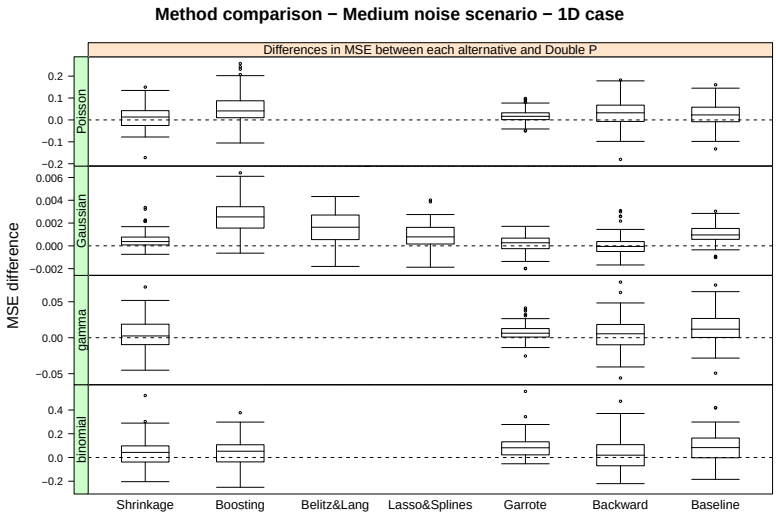

Figure 5-3: MSE results between the methods discussed in Section 5.2 and the double penalty approach for four error distributions and linear predictor η_1. REML estimation is employed for all methods except for GAM boosting and the Belitz&Lang approach. Covariate correlation is 0 and the signal to ratio level is medium. Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 5.3.1. Boxplots show the distributions of differences in mean squared error between each method and the double penalty approach. In all cases a Wilcoxon signed rank test indicates that double penalty has lower MSE than the competing methods (p-value $< 10^{-6}$), except for the Backward method in the Gaussian and Binomial cases where there is no significant difference (p-value > 0.10).

Figure 5-3 compares the MSE performance of all the methods discussed in the chapter, relative to the double penalty approach of section 5.2.1, from phases 1 and 2 of the study (above the zero line indicates performance worse than the double penalty method). Notice that `GAMBoost` supports only canonical link functions, hence MSE results for the gamma case are not available. Our results indicate that, overall, the double penalty approach performs significantly better than the competing methods in terms of MSE.

Figures 5-4 and 5-5 show the false positive rates for the methods considered here and the four error models, at each signal to noise ratio level. GAM boosting is not competitive. This result is in agreement with the findings of Cantoni

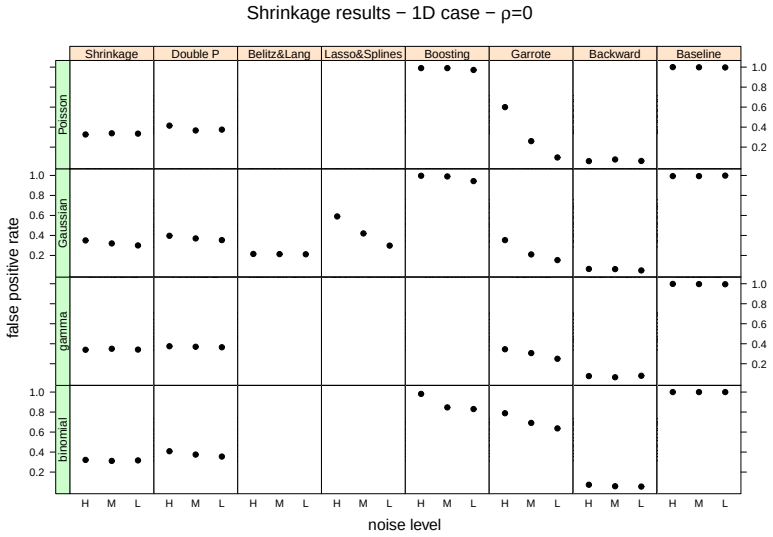

Figure 5-4: Shrinkage results for the methods discussed in Section 5.2, for four error distributions and linear predictor η_1. REML estimation is employed for all methods except for GAM boosting and the Belitz&Lang approach. Covariate correlation is 0 and H, M and L stand for high, medium and low signal level. Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 5.3.1. False positive rates give the proportion of times spurious terms are selected. Vertical lines show ± 2 standard error bands.

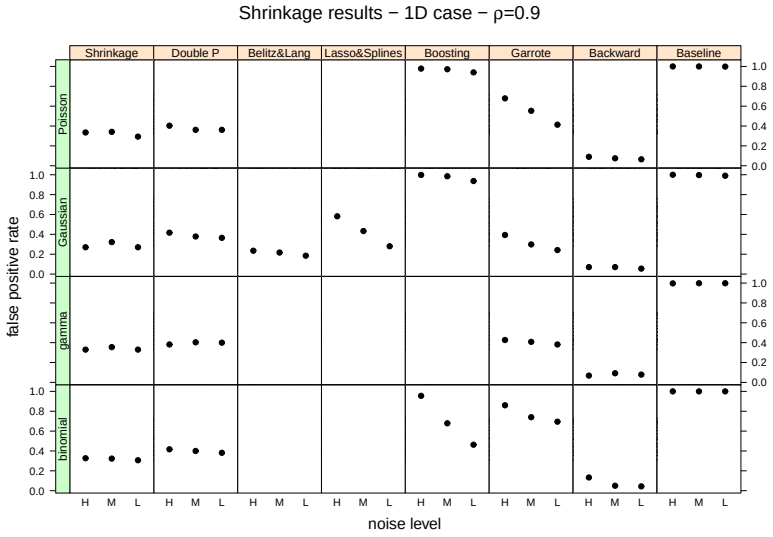

Figure 5-5: Shrinkage results for the methods discussed in Section 5.2, for four error distributions and linear predictor η_1. REML estimation is employed for all methods except for GAM boosting and the Belitz&Lang approach. Covariate correlation is 0.9. Further details are given in the caption of Figure 5-4.

et al. (2010). Shrinkage and double penalty are competitive as compared to the alternatives. The nonnegative garrote estimator is also competitive but not for the binomial case. As covariate correlation increases, the nonnegative garrote performance worsens. Backward selection yields the best results, but false negative rates are about 0.4, 0.3, and 0.1. These increase by about 0.1 point when covariate correlation is 0.9. So, if the data have high information content then Backward selection may be preferred over the competing methods, otherwise our proposals yield the most reliable results.

The Belitz&Lang approach yields good false positive rates. However, false negative rates (plots not shown here) indicate that this method eliminates influential covariates with rates about 0.25, 0.18, and 0.09 for the high, medium and low noise cases, respectively. This also explains its MSE performance as compared to the other approaches (see Figure 5-3). False negative rates are about 0.60, 0.29 and 0.17 when covariate correlation is 0.9. The combination of high false negative rates and relatively high MSE led us to drop the Belitz&Lang method after phase 1 of the study.

The poor false positive rate performance of `GAMBoost` is because the procedure typically retains predictors whose estimated curves are close to the zero line and that have been selected in a small number of boosting steps. These two facts could be combined into a procedure to improve false positive rates. However this presents us with some difficulty in obtaining fair false positive rate criteria for comparison with other methods, so is not pursued here. Alternative boosting procedures such as those documented in Bühlmann and Hothorn (2010) and Shafik and Tutz (2009) should also improve performance, but in the absence of public domain software we do not pursue these approaches here.

Figure 5-6 compares the MSE of the methods considered in phase 3 (linear predictor η_2) to the MSE of the double penalty approach. Results are again given by error model and method. The results confirm the finding that the double penalty approach yields overall the smallest MSEs. Similar conclusions were obtained when $\rho = 0.9$. Figure 5-7 indicates that shrinkage and double penalty yield, overall, reasonable false positive rate results. The nonnegative garrote estimator also produces reasonable results but not for the binomial case. As before, false negative rates (not reported here) indicate the Backward selection should be preferred over the other methods if the data have high information content.

Finally, Figures 5-8 and 5-9 show MSE comparisons by error model and method for linear predictor η_3 from phase 4 of the simulations, when mod-

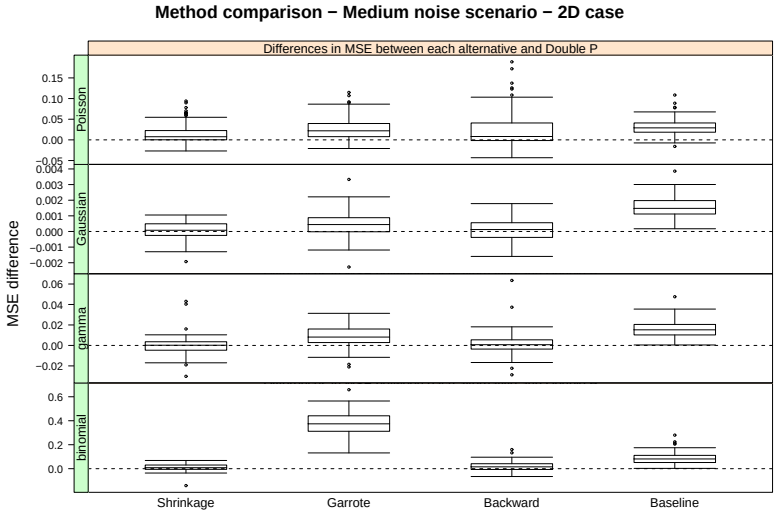

Figure 5-6: MSE comparisons between some of the methods discussed in Section 5.2 and the double penalty approach for four error distributions, when REML estimation and linear predictor η_2 are employed. Covariate correlation is 0 and the signal to ratio level is medium. Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 5.3.1 and in the caption of Figure 5-3. In all cases a Wilcoxon signed rank test indicates that double penalty has lower MSE than the competing methods (p-value $< 10^{-6}$).

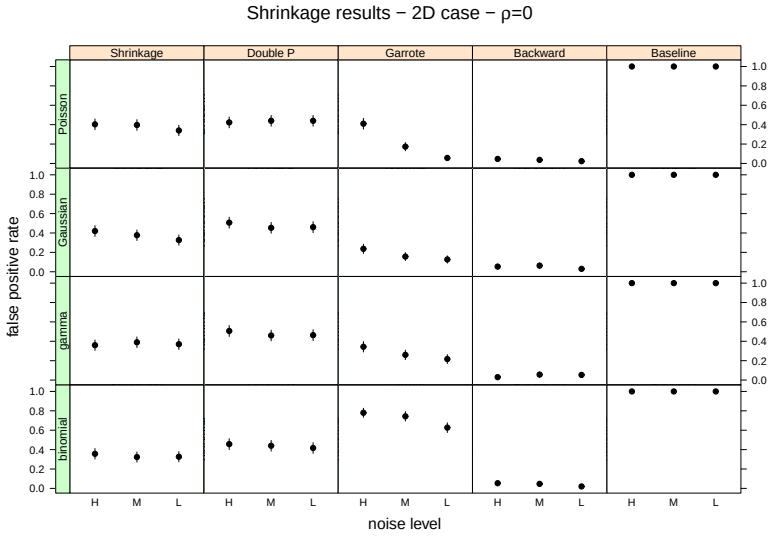

Figure 5-7: Shrinkage results for some of the methods discussed in Section 5.2 and four error distributions, when REML estimation and linear predictor η_2 are employed. Covariate correlation is 0. Further simulation details are given in Section 5.3.1 and in the caption of Figure 5-4.

els are fitted using fourteen and thirty covariates, respectively, most of which are nuisance variables. The results confirm the findings of this section, with the only difference that double penalty does not outperform the shrinkage approach. The overall shrinkage performance of the methods for the two scenarios was in line with that reported in the previous plots (overall false positive rates about 0.33, 0.37, 0.84, 0.51, 0.09, and 1 for Shrinkage, Double penalty, Boosting, Garrote, Backward and Baseline, respectively).

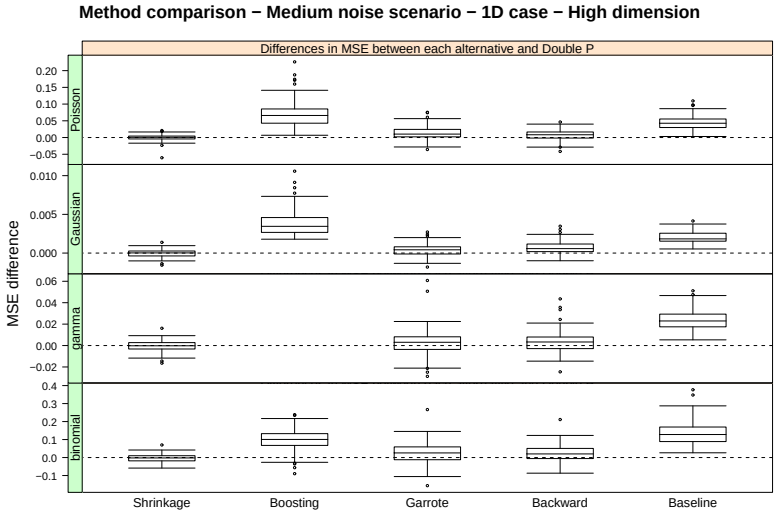

Figure 5-8: MSE comparisons between some of the methods discussed in Section 5.2 and the double penalty approach for four error distributions and linear predictor η_3. REML estimation is employed for all methods except for GAM boosting. Models are fitted using fourteen covariates, eleven of which are not influential. Covariate correlation is 0 and the signal to ratio level is medium. Further simulation details are given in Section 5.3.1 and in the caption of Figure 5-3. In all cases a Wilcoxon signed rank test indicates that double penalty has lower MSE than the competing methods (p-value $< 10^{-5}$), except for the Shrinkage approach where there is no significant difference (p-value > 0.29).

5.4 Real data example

In this section, we show the results obtained by applying the methods discussed in the chapter to a real dataset on plasma beta-carotene levels.

5.4.1 Beta-carotene data

The data are from a cross-sectional study conducted in the United States. The aim of the analysis was to investigate the relationship between beta carotene plasma concentrations and personal characteristics as well as dietary variables of subjects who had a biopsy examination or removed lesions of the

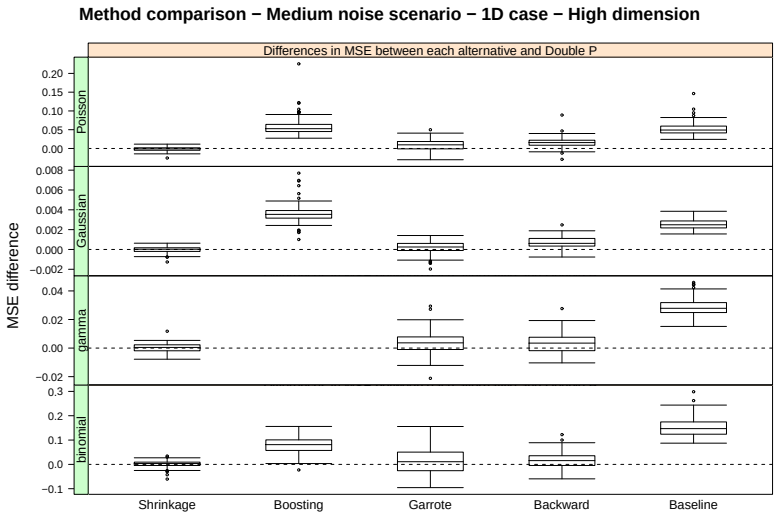

Figure 5-9: MSE comparisons between some of the methods discussed in Section 5.2 and the double penalty approach for four error distributions and linear predictor η_3. REML estimation is employed for all methods except for GAM boosting. Models are fitted using thirty covariates, twenty-seven of which are spurious. Covariate correlation is 0 and the signal to ratio level is medium. Further simulation details are given in Section 5.3.1 and in the caption of Figure 5-3. In all cases a Wilcoxon signed rank test indicates that double penalty has lower MSE than the competing methods (p-value $< 10^{-6}$), except for the Shrinkage approach where there is no significant difference (p-value > 0.33).

lung, colon, breast, skin, ovary or uterus that were not found to be cancerous (Nierenberg *et al.*, 1989; Marra and Radice, 2010). The dataset was obtained from the StatLib-Datasets Archive website (`http://lib.stat.cmu.edu/datasets/Plasma_Retinol`) and is made up of 315 individuals. The dataset contains a number of continuous variables.

The covariates considered were *age* (in years), *Quetelet index* (which is a measure of obesity defined as weight divided by the square of height), number of *calories* consumed per day, *plasma beta-carotene* (ng/ml), grams of *fat* consumed per day, grams of *fiber* consumed per day, *cholesterol* (mg per day), and *dietary beta-carotene* (mcg per day).

5.4.2 Results

The aim was to fit a nonparametric model and perform variable selection. As pointed out in Marra and Radice (2010), plasma BC levels strongly exhibit a positively skewed distribution. Therefore, a gamma distribution with a log link function between the linear predictor and the mean was employed. Notice that GAM boosting, the Lasso&Splines approach and Belitz&Lang method were not applied on this dataset given the results of the previous section. We applied the remaining methods using each of GCV and REML with model

fitting settings as discussed in Section 5.3.1. Pearson correlations among the covariates were in the range $[0.05, 0.9]$, and the squared correlation coefficient between μ_i (calculated using a standard GAM) and y_i was about 0.25, suggesting that the noise in this dataset is high.

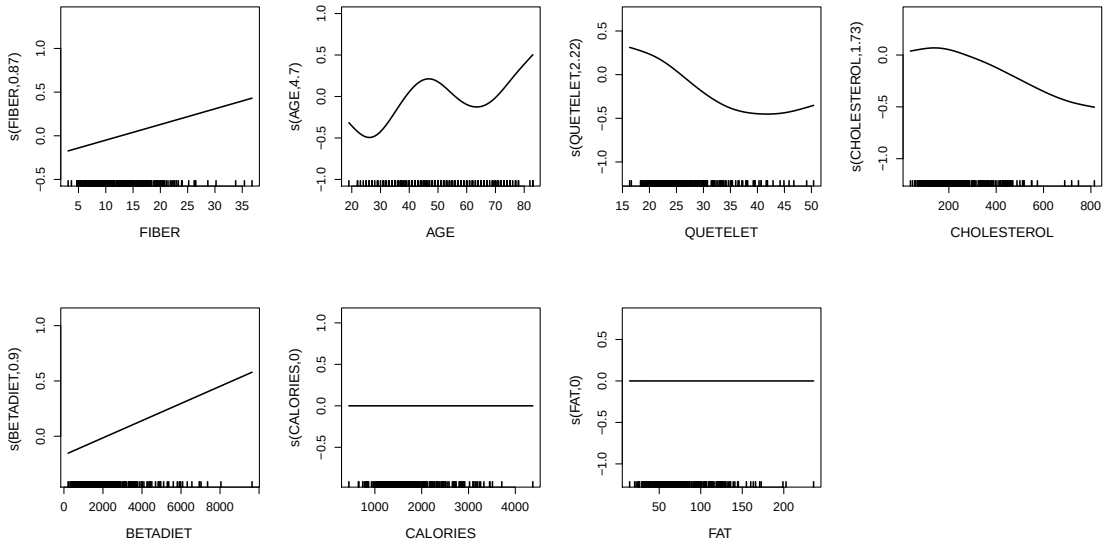

Figure 5-10: Smooth function estimates obtained by applying the double penalty approach with REML estimation on the plasma beta-carotene dataset described in Section 5.4.1. The results are reported on the scale of the linear predictor. The numbers in brackets in the y-axis captions are the edf of the smooth curves. The 'rug plot', at the bottom of each graph, shows the covariate values.

According to the shrinkage, double penalty and nonnegative garrote approaches the variables *calories* and *fat* were not influential, hence removed from the model. Backward selection removed *calories*, *fat* and *fiber*, suggesting the elimination of an important covariate. In fact, this was consistent with our simulation study which showed that if the data do not have high information content then Backward selection eliminates influential predictors.

Figure 5-10 shows the smooth function estimates obtained by applying the double penalty approach on the plasma beta-carotene dataset. Similar results were obtained by using the shrinkage and nonnegative garrote methods (plots not reported here). The estimated functions reveal the presence of non-linear relationships between the outcome and the selected regressors. This allows the researcher to gain more insights into the phenomenon of plasma beta-carotene in comparison to using a fully parametric approach. The smooths of *dietary beta-carotene* and *fiber* exhibit a linear behaviour, hence these terms can enter the model in a parametric manner. Finally, we repeated 5-fold cross validation 100 times, and then calculated prediction risk estimates. The results, dis-

played in Figure 5-11, are consistent with the findings of our simulation study; overall, REML outperforms GCV and the double penalty approach performs significantly better than the competing methods in terms of prediction.

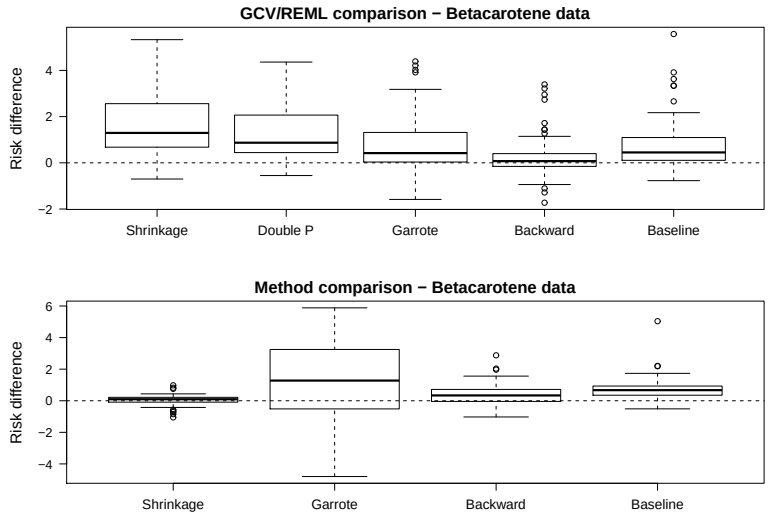

Figure 5-11: The top boxplots report prediction risk comparisons (in units of 10^3) between GCV and REML for some of the methods discussed in Section 5.2 when using the beta-carotene dataset (see details in Section 5.4). The plots show the distributions of differences in prediction risk estimate between GCV and REML, which were obtained repeating 5-fold cross validation 100 times. In all cases a Wilcoxon signed rank test indicates the REML yields lower risk estimates as compared to GCV (p-value $< 10^{-19}$), except for Backward where this evidence is less strong (p-value < 0.022). The bottom boxplots report prediction risk comparisons between the four shrinkage methods used for the beta-carotene dataset and the double penalty approach, when REML estimation is employed. The plots show the distributions of differences in prediction risk estimate between each method and double penalty. In all cases a Wilcoxon signed rank test indicates that double penalty produces lower risk estimates as compared to the competing methods (p-value $< 10^{-18}$), except for Shrinkage where this evidence is less strong (p-value < 0.017).

To complete the analysis, following e.g. Bühlmann and Hothorn (2010), we looked at a synthetically enlarged problem. Specifically, we generated ten uniform variables with correlations approximately equal to 0.5 (see Section 5.3.1) and included them in the model containing the real predictors. This allowed us to check how many ineffective variables would be selected. The proposed approaches performed satisfactorily in that, overall, seven variables out of ten were eliminated. This was consistent with the simulation results.

We now discuss the conclusions obtained using a classic GAM with the confidence intervals of Chapter 4, for model selection purposes. Figure 5-12 reports the smooth function estimates obtained when we employ a standard GAM without any shrinkage smoother. As we can see, although the impact of some variables appears to be weak, all regressors are retained in the model. If we use confidence intervals to carry out variable selection, we may conclude

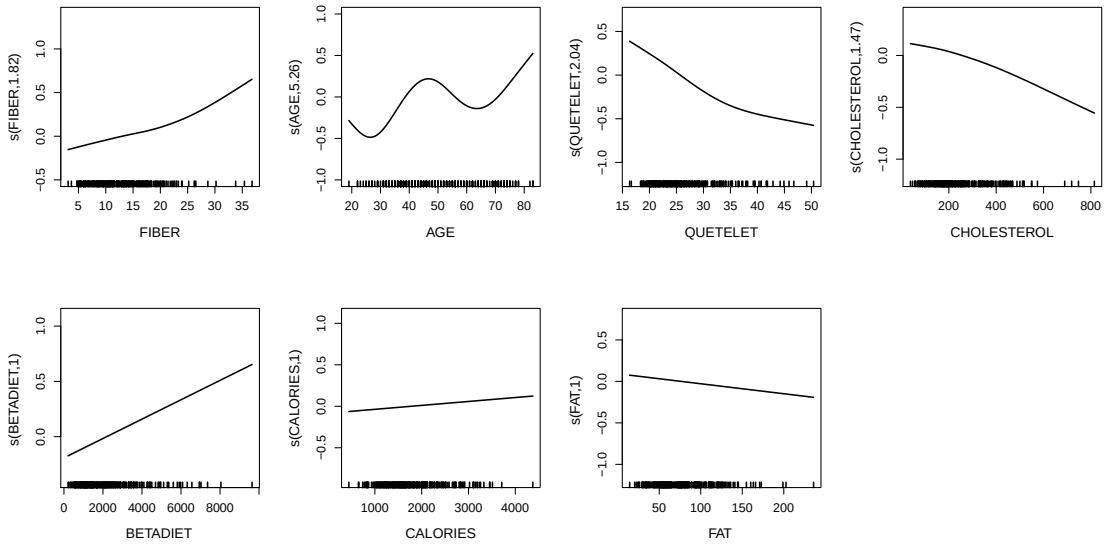

Figure 5-12: Smooth function estimates obtained by fitting a standard GAM with REML estimation on the plasma beta-carotene dataset described in Section 5.4.1. Further details are given in the caption of Figure 5-10.

that *calories*, *fat*, *fiber*, and *cholesterol* are not influential since a straight flat line can be drawn within their variability bands (see Figure 5-13). However, as discussed in Section 4.4, although such intervals are a useful guide to help model selection, one can not completely rely on them alone to infer features in the smooth estimates. The results in Figure 5-10 suggest that only *calories* and *fat* should be removed from the model.

Based on the example discussed in Section 4.4, we can be confident that the true effect of *age* is nonlinear. However, the fact that a straight line can be drawn within the variability bands for the smooth components of *Quetelet index* and *cholesterol* does not imply that their true relationship is linear. This question may be addressed using the approach described in Section 2.6. Here, we can test the null hypothesis that a simple model (the impact of *Quetelet index* and *cholesterol* is assumed to be linear) is true against the alternative relating to a more complex model (the effect *Quetelet index* and *cholesterol* is nonlinear). In this case, the resulting p-value is 0.068 which indicates to reject the null at the 10% significance level. In other words, the increase in the log-likelihood due to replacing the smooths with straight lines is somewhat significant. This result is consistent with Figure 5-10 which suggests that *Quetelet index* and *cholesterol* have a nonlinear impact.

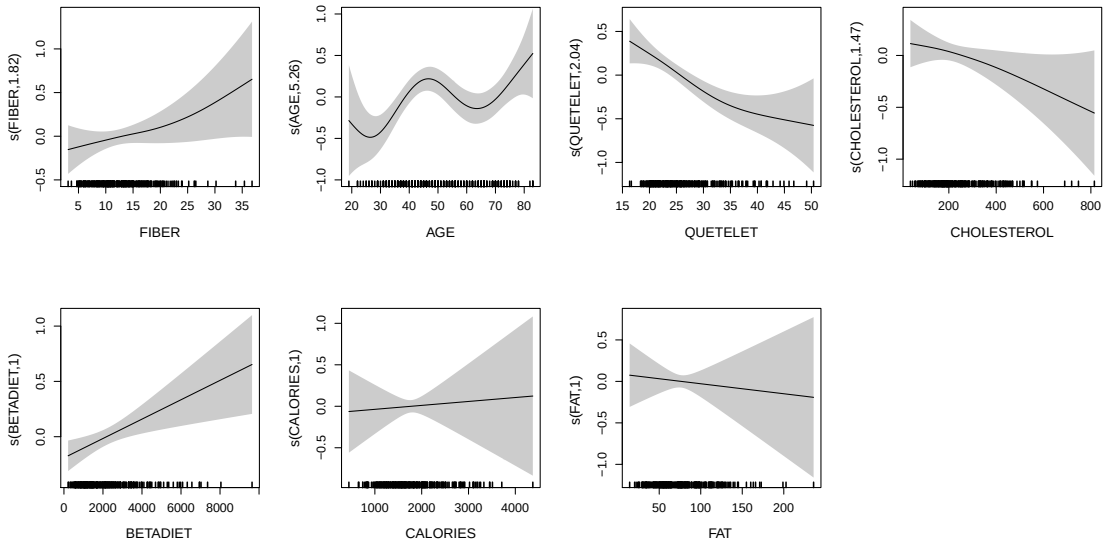

Figure 5-13: The same smooth function estimates as those reported in Figure 5-12. The shaded regions represent 95% Bayesian confidence intervals discussed in Chapter 4.

5.5 Discussion

In this chapter, we have proposed two effective shrinkage methods and extended the nonnegative garrote estimator to achieve component selection within GAMs, for situations in which there are moderate numbers of spurious covariates which it would be beneficial to eliminate. We have compared the empirical performance of the proposals to that of some available techniques via an extensive simulation study, and illustrated some of the procedures analysing data on plasma beta-carotene levels from a cross-sectional study conducted in the United States.

Our results show that, overall, the proposed shrinkage approaches perform significantly better than the competing methods in terms of predictive ability. As for the variable selection performance, the shrinkage and double penalty approaches are competitive as compared to the alternatives. The nonnegative garrote estimator is also competitive but not for the binomial case. As covariate correlation increases, the nonnegative garrote performance worsens. Matters improve when REML is employed for smoothing parameter estimation. Backward selection yields the best false positive rates. However, false negative rates indicate that this method eliminates influential covariates, especially for the low signal to noise ratio level-high covariate correlation scenario, worsening its predictive ability. GAM boosting performs comparatively poorly in the scenarios considered here, although its shrinkage performance could almost

certainly be improved by changing the way variables are selected.

If the data have high information content then backward selection may be preferred over alternatives, otherwise our proposals yield the most reliable results. The main limitation of all the methods discussed here, except for GAM boosting and Belitz and Lang (2008), is that they can not deal with situations in which $n < p$ (n is sample size and p is the number of predictors) and indeed require $n \geq kp$ where k is the average basis size used for the smoothers. For $p > n$, GAM boosting and the Belitz&Lang method appear to be the only methods available when an additive structure is considered appropriate. It is notable that both these methods rely on prediction error criteria such as AIC and GCV for smoothness selection, while for the other methods we found REML smoothness selection to generally yield superior results. It would be interesting to see whether REML based extensions to Belitz and Lang (2008) or GAM boosting could be produced which would improve the performance of these methods.

Summary

In this thesis we have discussed some theoretical and practical aspects of penalized regression spline smoothing. This technique is becoming among the most useful and used of statistical methods in many fields. We provided a brief overview of GAMs, based on the penalized likelihood framework with regression splines, by discussing some aspects that are relevant to this thesis. We then concentrated on three main issues.

In Chapter 3, we tackled the problem of unobservable confounding, issue which can not be neglected especially in observational studies when the researcher is interested in evaluating the effects of one or more predictors of interest on a response variable. The IV approach represents a valid means to account for unobservables. This technique, first proposed in econometrics, only recently has received some attention in the applied statistical literature. We have proposed a flexible procedure to carry out IV analysis within the GAM context. Our proposal is backed up with an extensive simulation experiment whose results confirmed that the proposed procedure represents a flexible theoretically sound means of obtaining consistent curve/parameter estimates in the presence of unmeasured confounding. We have also introduced a Bayesian interval correction procedure for the intervals of the proposed two-step approach, which performed well in simulation in terms of coverage probabilities.

In Chapter 4, we have shown by simulation and extension of Nychka's analysis, that the Wahba/Silverman type Bayesian intervals for the components of a penalized regression spline based GAM have generally good frequentist properties, across-the-function. The exception occurs when components estimated subject to identifiability constraints have interval widths vanishing somewhere as a result of heavy smoothing. Coverage probabilities can be improved if intervals are only obtained for unconstrained quantities, such as a smooth component plus the model intercept. The theoretical results also allow us to define alternative intervals when a frequentist approach is adopted. The results make a novel contribution in extending Nychka's argument to the GAM component case thereby pinpointing the circumstances in

which the intervals will and will not work, and explaining the role of smoothness selection as well as smoothing parameter uncertainty. The findings are backed up with quite extensive simulation testing of the finite sample performance of the three types of confidence intervals considered in this work.

In Chapter 5, we have proposed two simple but effective shrinkage methods and extended the nonnegative garrote estimator to achieve component selection within GAMs. We have compared the empirical performance of the proposals to that of some available techniques via an extensive simulation study, and illustrated some of the procedures analysing data on plasma beta-carotene levels. Our results show that, overall, the proposed approaches performs significantly better than the competing methods in terms of predictive ability and shrinkage. The performance of the methods improves when REML is employed for smoothing parameter estimation. If the data have high information content then Backward selection may be preferred over the competing methods, otherwise our proposals yield the most reliable results.

An interesting area for future work is to develop flexible simultaneous equation estimation methods dealing with the issues of (i) unmeasured/unobservable confounding (recall that a solution based on two-stage estimation has been presented in Chapter 3), (ii) heterogeneity, (iii) sample selection and (iv) covariate-response nonlinear relationships. A non-exhaustive list of observational economic and biostatistical studies affected by issues (i) and (ii) includes the study of the effect of private health insurance on medical care utilization, impact of diabetes on employment, effect of physical activity on obesity, and association between various reported risky sexual behaviors and sexually transmitted infection. As for (iii), in many applications, the dependent variable can take any nonnegative real value but has positive probability of a zero outcome. For instance, when each observation is a record of the daily rainfall, many days may have no rainfall. In household expenditure and medical cost studies, some households spend nothing on certain goods, and a portion of the population have zero medical expense. As for point (iv), the functional shape of any covariate-response relationships is rarely known a priori and the outcome may depend on the predictors in a complicated manner. The estimators typically employed for applied work neglect at least one of the problems above, hence yielding biased and inconsistent estimates of the relationship of interest. This is clearly not desirable because it may cause, e.g., policy makers to focus on wrong policy decisions. Estimation methods based on simultaneous equations and smoothing splines seem to be a promising alternative to the single equation estimation techniques available in the literature.

Bibliography

[1] AI, C., CHEN, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71, pp. 1795–1843.

[2] AKAIKE, H., 1973. Information Theory and An extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory*, eds. B. Petran and F. Csaaki, Budapest, pp. 267–281.

[3] AMEMIYA, T., 1974. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2, pp. 105–110.

[4] AVALOS, M., GRANDVALET, Y., AMBROISE, C., 2007. Parsimonious additive models. *Computational Statistics and Data Analysis*, 51, pp. 2851–70.

[5] BELITZ, C., LANG, S., 2008. Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis*, 51, pp. 6044–6059.

[6] BINDER, H., TUTZ, G., 2008. A comparison of methods for the fitting of generalized additive models. *Statistics and Computing*, 18, pp. 87-99.

[7] BECHER, H., 1992. The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, 11, pp. 1747-1758.

[8] BECK, C.A., PENROD, J., GYORKOS, T. W., SHAPIRO, S., PILOTE, L., 2003. Does aggressive care following acute myocardial infarction reduce mortality? Analysis with instrumental variables to compare effectiveness in Canadian and United States patient populations. *Health Services Research*, 38, pp. 1423–1440.

[9] BENEDETTI, A., ABRAHAMOWICZ, M., 2004. Using generalized additive models to reduce residual confounding. *Statistics in Medicine*, 23, pp. 3781–801.

[10] BOUND, J., JAEGER, D. A., BAKER, R. M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90, pp. 443–450.

[11] BREIMAN, L., 1995. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37, pp. 373-84.

[12] BRUGIAVINI, A., JAPPELLI, T., WEBER, G., 2002. The survey of health, aging and wealth. Mimeo, Universita' di Salerno, Italy.

[13] BUCHMUELLER, T. C., GRUMBACH, K., KRONICK, R., KAHN, J. G., 2005. Book review: The effect of health insurance on medical care utilization and implications for insurance expansion: A review of the literature. *Medical Care Research and Review*, 62, pp. 3–30.

[14] BÜHLMANN, P., HOTHORN, T., 2010. Twin Boosting: improved feature selection and prediction. *Statistics and Computing*, 20, pp. 119–138.

[15] BÜHLMANN, P., Yu, B., 2003. Boosting With the L2 Loss: Regression and Classification. *Journal of the American Statistical Association*, 98, pp. 324-339.

[16] CANTONI, E., FLEMMING, J., RONCHETTI, E., 2010. Variable selection in additive models by nonnegative garrote. *Statistical Modelling*, to appear.

[17] CANTONI, E., HASTIE, T., 2002. Degrees of freedom tests for smoothing splines. *Biometrika*, 89, pp. 251-263.

[18] CHIB, S., GREENBERG, E., 2007. Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16, PP. 86–114.

[19] CRAVEN, P., WAHBA, G., 1979. Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, 31, pp. 377–403.

[20] DAS, M., 2005. Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics*, 124, pp. 335–361.

[21] DUCHON, J., 1977. Splines minimizing rotation-invariant semi-norms in solobev spaces. In W. Schemp and K. Zeller (Eds.), *Construction Theory of Functions of Several Variables*, pp. 85-100, Berlin, Springer.

[22] EFRON B., HASTIE T., JOHNSTONE I., TIBSHIRANI, R., 2004. Least angle regression. *The Annals of Statistics*, 32, pp. 407-451.

[23] FAHRMEIR, L., KNEIB, T., LANG, S., 2004. Penalized Structured Additive Regression for Space-Time Data: A Bayesian Perspective. *Statistica Sinica*, 14, pp. 731–761

[24] FAHRMEIR, L., LANG, S., 2001. Bayesian Inference for Generalized Additive Mixed Models based on Markov Random Field Priors. *Journal of the Royal Statistical Society Series C*, 50, pp. 201–220.

[25] FROSINI, B. V., 2006. Causality and causal models: A conceptual perspective. *International Statistical Review*, 74, pp. 305–334.

[26] GENTLE, J. E., 2003. *Random Number Generation and Monte Carlo Methods*, London: Springer-Verlag.

[27] GREENLAND, S., 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29, pp. 722–729.

[28] GREVEN, S., KNEIB, T., 2009. On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models. Johns Hopkins University, Department of Biostatistics Working Papers, Paper 179.

[29] GU, C., 1992. Penalized Likelihood Regression - A Bayesian Analysis. *Statistica Sinica*, 2, pp. 255–264.

[30] GU, C., 2002. *Smoothing Spline ANOVA Models*, London: Springer-Verlag.

[31] GU, C., WAHBA, G., 1993. Smoothing Spline ANOVA with Component-Wise Bayesian Confidence Intervals. *Journal of Computational and Graphical Statistics*, 2, pp. 97-117.

[32] GUISAN, A., EDWARDS, T. C., HASTIE, T., 2002. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, 157, pp. 89–100.

[33] HALL, P., 1992. Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. *Annals of Statistics*, 20, pp. 675–694.

[34] HALL, P., HOROWITZ, J. L., 2005. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33, pp. 2904–2929.

[35] HÄRDLE, W., BOWMAN, A. W., 1988. Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. *Journal of the American Statistical Association*, 83, pp. 102–110.

[36] HÄRDLE, W., HALL, P., MARRON, J. S., 1988. How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?. *Journal of the American Statistical Association*, 83, pp. 86–95.

[37] HÄRDLE, W., HUET, S., MAMMEN, E., SPERLICH, S., 2004. Bootstrap Inference in Semiparametric Generalized Additive Models. *Econometric Theory*, 20, pp. 265–300.

[38] HÄRDLE, W., MARRON, J. S., 1991. Bootstrap Simultaneous Error Bands for Nonparametric Regression. *The Annals of Statistics*, 19, pp. 778–796.

[39] HARMON, C., NOLAN, B., 2001. Health insurance and health services utilization in Ireland. *Health Economics*, 10, pp. 135–145.

[40] HAUSMAN, J. A., 1978. Specification tests in econometrics. *Econometrica*, 46, pp. 1251–1271.

[41] HAUSMAN, J. A., 1983. Specification and Estimation of Simultaneous Equations Models. In Griliches, Z., Intriligator, M.D. eds. *Handbook of Econometrics*, Amsterdam: North Holland, pp. 391–448.

[42] HASTIE, T., TIBSHIRANI, R., 1990. *Generalized Additive Models*, London: Chapman & Hall.

[43] HECKMAN, J., 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46, pp. 931–59.

[44] HESTERBERG, T., CHOI, N. H., MEIER, L., FRALEY, C., 2008. Least angle and l_1 penalized regression: A review. *Statistics Surveys*, 2, pp. 61–93.

[45] HOFTER, R. H., 2006. Private health insurance and utilization of health services in Chile. *Applied Economics*, 38, pp. 423–439.

[46] JOHANSEB, D., GRONBAEK, M., OVERVAD, K., SCHNOHR, P., ANDERSEN, P. K., 2005. Generalized Additive Models applied to analysis of the relation between amount and type of alcohol and all-cause mortality. *European Journal of Epidemiology*, 20, pp. 29–36.

[47] JOHNSTON, K. M., GUSTAFSON, P., LEVY, A. R., GROOTENDORST, P., 2008. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, 27, pp. 1539–1556.

[48] KAUERMANN, G., CLAESKENS, G., OPSOMER, J. D., 2009. Bootstrapping for Penalized Spline Regression. *Journal of Computational and Graphical Statistics* 18, pp. 126–146.

[49] KAUERMANN, G., KRIVOBOKOVA, T., FAHRMEIR, L., 2009. Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B*, 71, pp. 487–503.

[50] KAUERMANN, G., OPSOMER, J. D., 2003. Local Likelihood Estimation in Generalized Additive Models. *Scandinavian Journal of Statistics*, 30, pp. 317–337.

[51] KAUERMANN, G., TUTZ, G., 2001. Testing generalized linear and semiparametric models against smooth alternatives. *Journal of the Royal Statistical Society B*, 63, pp. 147–166.

[52] KIMELDORF, G., WAHBA, G., 1970. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41, pp. 495–502.

[53] LEIGH, J. P., SCHEMBRI, M., 2004. Instrumental variables technique: Cigarette price provided better estimate of effects of smoking on SF-12. *Journal of Clinical Epidemiology*, 57, pp. 284–293.

[54] LIN, Y., ZHANG, H. H., 2006. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34, pp. 2272-2297.

[55] LINDEN, A., ADAMS, J. L., 2006. Evaluating disease management programme effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12, pp. 148–154.

[56] MADDALA, G., 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

[57] MARRA, G., RADICE, R., 2010. Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, 19, pp. 107–125.

[58] MARRA, G., RADICE, R., 2011. Do We Adequately Control for Unmeasured Confounders when Estimating the Short-Term Effect of Air Pollution on Mortality? *Water, Air, & Soil Pollution*, in press.

[59] MARRA, G., RADICE, R., 2012. A Flexible Instrumental Variable Approach. *Statistical Modelling*, in press.

[60] MARRA, G., WOOD, S. N. Coverage Properties of Confidence Intervals for Generalized Additive Model Components. Submitted.

[61] MARRA, G., WOOD, S. N. Practical Variable Selection for Generalized Additive Models. Submitted.

[62] McCULLAGH, P., NELDER, J. A., 1989. *Generalized Linear Models*, London: Chapman & Hall.

[63] NEWEY, W. K., POWELL, J. L., 2003. Instrumental variable estimation of nonparametric models. *Econometrica*, 71, pp. 1565–1578.

[64] NIERENBERG, D. W., STUKEL, T. A., BARON, J. A., DAIN, B. J., GREENBERG, E. R., 1989. The Skin Cancer Prevention Study Group. Determinants of Plasma Levels of Beta-carotene and Retinol. *American Journal of Epidemiology*, 130, pp. 511–21.

[65] NYCHKA, D., 1988. Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Association*, 83, pp. 1134–1143.

[66] REIDPATH, D. D., CRAWFORD, D., TILGNER, L., GIBBONS, C., 2002. Relationship between body mass index and the use of healthcare services in Australia. *Obesity research*, 10, pp. 526–531.

[67] REISS, P. T., OGDEN, R. T., 2009. Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society B*, 71, pp. 505–524.

[68] ROYSTON, P., 2005. Polynomial Regression. In: Armitage P, Colton T eds. *Encyclopedia of Biostatistics*, Wiley.

[69] ROYSTON, P., ALTMAN, D. G., 1994. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with Discussion). *Applied Statistics*, 43, pp. 429–467.

[70] RUPPERT, D., WAND, M. P., CARROLL, R. J., 2003. *Semiparametric Regression*. London: Cambridge University Press.

[71] SCHEIPL, F., GREVEN, S., KUCHENHOFF, H., 2008. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics and Data Analysis*, 52, pp. 3283–3299.

[72] SILVERMAN, B. W., 1985. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society Series B*, 47, pp. 1–52.

[73] STAIGER, D., STOCK, J. H., 1997. Instrumental variables regression with weak instruments. *Econometrica*, 65, pp. 557–586.

[74] TERZA, J. V., BASU, A., RATHOUZ, P. J., 2008. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27, pp. 531–543.

[75] TIBSHIRANI, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58, pp. 267-288.

[76] TUTZ, G., BINDER, H., 2006. Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting. *Biometrics*, 62, pp. 961–71.

[77] TUTZ, G., BINDER, H., 2007. Boosting ridge regression. *Computational Statistics and Data Analysis*, 51, pp. 6044–6059.

[78] WAGER, C., VAIDA, F., KAUERMANN, G., 2007. Model Selection for Penalized Spline Smoothing using Akaike Information Criteria. *Australian and New Zealand Journal of Statistics*, 49, pp. 173–190.

[79] WAHBA, G., 1983. Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline. *Journal of the Royal Statistical Society Series B*, 45, pp. 133–150.

[80] WAHBA, G., 1985. A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *The Annals of Statistics*, 13, pp. 1378–1402.

[81] WAHBA, G., 1990. *Spline models for observational data*, Philadelphia: SIAM.

[82] WANG, Y., WAHBA, G., 1995. Bootstrap Confidence Intervals for Smoothing Splines and Their Comparison to Bayesian Confidence Intervals. *Journal of Statistical Computation and Simulation*, 51, pp. 263–279.

[83] WOOD, S. N., 2000. Modelling and Smoothing Parameter Estimation With Multiple Quadratic Penalties. *Journal of the Royal Statistical Society Series B*, 62, pp. 413–428.

[84] WOOD, S. N., 2003. Thin Plate Regression Splines. *Journal of the Royal Statistical Society Series B*, 65, pp. 95–114.

[85] WOOD, S. N., 2004. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association*, 99, pp. 673–86.

[86] WOOD, S. N., 2006a. *Generalized Additive Models: An Introduction with R*, London: Chapman & Hall.

[87] WOOD, S. N., 2006b. On Confidence Intervals for Generalized Additive Models based on Penalized Regression Splines. *Australian & New Zealand Journal of Statistics*, 48, pp. 445–64.

[88] WOOD, S. N., 2008. Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models. *Journal of the Royal Statistical Society Series B*, 70, pp. 495–518.

[89] WOOD, S. N., 2010. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B*, DOI: 10.1111/j.1467-9868.2010.00749.x.

[90] WOOLDRIDGE, J. M., 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

[91] XUE, L., 2009. Consistent variable selection in additive models. *Statistica Sinica*, 19, pp. 1281–1296.

[92] YUAN, M., 2007. Nonnegative Garrote Component Selection in Functional ANOVA Models. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, March 21-24, San Juan, Puerto Rico.

[93] YUAN, M., LIN, Y., 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68, pp. 49-67.

[94] ZANIN, L., MARRA, G., 2011. Rolling Regression Versus Time-Varying Coefficient Modelling: An Empirical Investigation of the Okun's Law in Some Euro Area Countries. *Bulletin of Economic Research*, in press.

[95] ZHANG, H. H., LIN, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, 16, pp. 1021-1041.

[96] ZOU, H., 2006. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101, pp. 1418-1429.