Klinger:

Generalized Soft-Thresholding and Varying-coefficient Models

Sonderforschungsbereich 386, Paper 59 (1997)

Generalized Soft–Thresholding and Varying–coefficient Models

Artur Klinger

Institut für Statistik, Universität München
Ludwigsstr. 33, D–80539 München, Germany

email: artur@stat.uni-muenchen.de

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 “Statistische Analyse diskreter Strukturen, Modellierung und Anwendung in Biometrie und Ökonometrie.”

Summary

We propose a new method for estimation of unknown functions within the generalized linear model framework. The estimator leads to an adaptive economical description of the results in terms of basis functions. Our proposal extends the soft–thresholding strategy from ordinary wavelet regression to generalized linear models and multiple predictor variables. Several sets of basis functions, tailored to specific purposes, can be incorporated into our methodology. We discuss semiparametric statistical inference based on generalized soft–thresholding. An algorithm which produces a sequence of estimates corresponding to increasing model complexity is developed. Advantages of our approach are demonstrated by an application to German labour market data.

Some key words: Generalized additive models; Penalized Likelihood; Semiparametric models; Splines; Thresholding; Varying coefficients; Wavelets;

1 Introduction

During the last decade, developing flexible statistical models and methods to analyze them have been a topic of very active statistical research interest. There have been at least two main directions of methodological investigation. On the one side, smoothing procedures have been developed to allow for multiple predictors $x_1, \ldots, x_p$ and response variables y distributed according to a simple exponential family. For example, roughness penalty approaches are discussed in Hastie and Tibshirani (1990) (1993), Wahba (1990), Green and Silverman (1994) and Wahba, Wang, Gu, Klein and Klein (1995). One alternative is the principle of local likelihood estimation, which has been considered by Tibshirani and Hastie (1987), Staniswalis (1989), Fan, Heckman and Wand (1995) and Tutz and Kauermann (1997). As a common feature, smoothing methods incorporate a smoothing parameter that controls model complexity, i.e. the smoothness of the predictor functions. This bias–variance trade–off parameter acts continuously on the estimate. By continuous, we mean, that a small change of the smoothing parameter has only limited impact on the estimate.

On the other side, adaptive basis function approaches have been proposed in Friedman and Silverman (1989), Friedman (1991), and Stone, Hansen, Kooperberg and Troung (1997). Those techniques select an appropriate set of basis functions by forward selection – backward deletion strategies. For a given set of basis functions, corresponding coefficients are determined by least–squares or maximum likelihood estimation. The bias–variance trade–off is governed by the selection procedure, that controls the number of basis functions included.

Basis function approaches have several attractive features: They give a compact output in terms of few basis coefficients contributing to the estimate. Models reduce to simple parametric form, if the data suggest that such models are adequate. Due to their parsimonious representation, familiar quantities, such as correlation measures, can be transferred from classical parametric models. By specifying an appropriate set of basis functions, the procedure can easily be tailored to specific purposes. For example, basis functions allowing for jumps and breakpoints within the estimates might be supplied to the estimator.

There are also some disadvantages of adaptive basis function methods, as with many variable selection techniques. They tend to produce highly variable estimates. Moreover, a small change of the parameter governing the selection process may result in a rather different model. When interpreting the estimate, this variability might lead to substantially wrong conclusions. Since the selection process and the estimation is based on the same data set, the estimates can be seriously biased. Once a basis function is selected, its contribution tends to be overestimated.

In this paper, we propose a general method for estimating functions within the generalized linear model setup. The proposed estimator yields an adaptive economical description of the estimates in terms of basis functions. However it shares the stability of smoothing procedures. Our proposal is based on soft–thresholding estimators, which have become popular in the context of wavelet regression, compare Donoho and Johnstone (1994), Nason and Silverman (1994), Donoho, Johnstone, Kerkyacharian and Picard (1995) and Bruce and Gao (1996).

This work transfers the soft–thresholding idea to generalized linear models and multiple predictor variables. In contrast to variable selection, soft–thresholding provides a unified framework for selection of basis functions and estimation of corresponding coefficients. The trade–off parameter acts continuously on the estimate. As will be demonstrated in the subsequent sections, the generalized soft–thresholding methodology nicely combines the stability of smoothing procedures with the adaptivity and interpretability of basis function approaches.

1·1 *Varying–coefficient models*

Suppose, we observe a one- or multidimensional response variable y and a set of metrical and categorical explanatory variables X. We assume that y given X follows a simple exponential family with density function

$$f(y,\theta) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}, \tag{1·1}$$

where b and c are given functions with b twice continuously differentiable and the natural parameter $\theta = \theta(\eta(X))$ depends on the predictor values. The nuisance parameter ϕ is considered as fixed and may be estimated separately, if necessary.

Generalized linear models are discussed in detail in McCullagh and Nelder (1989) and in Fahrmeir and Tutz (1994), for multivariate responses. Models of this type include, for example, the logit and probit models for binomial responses, log–linear models for count data and cumulative logistic models for discrete ordinal responses.

Flexible extensions of generalized linear models start with a linear parametric specification

$$\eta(X) = \mu + \beta_1 x_1 + \cdots + \beta_p x_p$$

for the predictor. Weakening the stringent assumption of linearity we obtain generalized additive models or GAM's, (Hastie and Tibshirani (1990)). The predictor for GAM's have additive structure

$$\eta(X) = \mu + \beta_1(x_1) + \cdots + \beta_p(x_p), \tag{1·2}$$

with effects β_j varying smoothly in x_j, for $j = 1, \ldots, p$.

Combining GAM's with state–space extensions of generalized linear models (Fahrmeir and Kaufmann (1991), Fahrmeir and Tutz (1994), Ch. 8.) leads to varying–coefficient models as introduced by Hastie and Tibshirani (1993). This framework assumes the effects of covariates z_j, $j = 1, \ldots, p$, possibly constructed from basic covariates $x_l \in X, l \neq j$, as smooth functions $\beta_j(x_j)$, $x_j \in X$ $j = 1, \ldots, p$. Extending the predictor of generalized linear models to

$$\eta(X) = \beta_0(x_0) + \beta_1(x_1)z_1 + \ldots + \beta_p(x_p)z_p, \tag{1·3}$$

varying–coefficient models are a valuable tool for exploring interactions between covariates z_j and their effect-modifiers x_j. Semiparametric models, where $x_1 = \cdots = x_p = 1$, generalized linear models for time series or event history data and generalized additive models are obtained as important special cases of (1·3).

1·2 *Outline of the paper*

A basis function approach is used for estimating varying coefficients in (1·3). Each single function is described by

$$\beta_j(x_j) = \sum_k c_{jk}\varphi_k(x_j)$$

in terms of basis functions $\varphi_k(x_j)$. Basis coefficients c_{jk}, which are not evident from the data become thresholded to zero. Remaining coefficients contribute to the estimate $\hat{\beta}_j(x_j)$.

In preparation, we review the basic idea of soft–thresholding of wavelet coefficients in Section 2. To transfer the wavelet estimator to non–Gaussian situations, an alternative definition of soft–thresholding using estimating equations is introduced. For simplicity, we consider first estimation of a univariate response function in the generalized linear model framework.

The generalized soft–thresholding estimator proposed in Section 3·1 is derived by incorporating log–likelihood score functions into the estimating equations. By its general definition, the estimator can be used in connection with any set of basis functions. Orthogonality is not required and different sets of basis functions can even be combined to describe a varying coefficient. Based on an equivalence of generalized soft–thresholding and absolute penalized likelihood estimation we propose an analogue to spline smoothing. This analogue provides results similar to smoothing splines, by having an economical representation in terms of basis functions. Along with the methodological development we illustrate the finite sample performance of the estimator by presenting simulation studies within a log–linear Poisson model. Locally adaptive function estimation using one–sided spline basis and wavelets is discussed briefly.

In Section 4 the concept of generalized soft–thresholding is extended to allow for simultaneous estimation of several functions within the varying–coefficient model. Some attention is directed to keep the number of trade–off parameters small. We propose a scaling procedure in Subsection 4·2. This procedure determines the smoothness of the varying coefficients by employing score test statistics.

To obtain further insight into the model and its effects, we derive a quadratic approximation to maximum likelihood tests in Section 5. In a semiparametric fashion, this test can be used formally when basis functions used to test are specified in advance. Informally, we use the resulting test statistics to suggest presence of certain components in the model. The parsimonious form of the estimator allows to compute an inverse information matrix with respect to basis coefficients c_{jk}. Further insight into estimation results is provided by analyzing the corresponding correlation matrix.

We propose to look at the estimator as a function of the trade–off parameter. This parameter controls the complexity of all effects simultaneously. An efficient algorithm for computing a sequence of estimates corresponding to increasing model complexity is developed in Section 6.

The advantages of our approach in practical data analysis are demonstrated in Section 7, where we apply the proposed methodology to German labour market data. Our main interest is the effect of gender on the probability for leaving unemployment. Where possible, the output of generalized soft–thresholding is presented as a function of the trade–off parameter. Hereby we achieve more transparency in communicating results.

2 SOFT–THRESHOLDING ESTIMATES FOR GAUSSIAN ERRORS

To review the basic ideas of soft–thresholding, suppose we are given n observations (x_i, y_i) satisfying

$$y_i = \eta(x_i) + \varepsilon_i,$$

where the ε_i are independently distributed as $N(0, \sigma^2)$. To recover $\eta(x)$ from the data, let us assume that $\eta(x)$ can be well approximated by a few basis functions from a set of orthogonal basis functions $\{\varphi_k(x)\}_{k=1}^{n}$. If $\eta(x)$ is homogeneously smooth in the sense of the some squared derivative, orthogonal Demmler–Reinsch splines, as discussed in Subsection 3·2, yield a parsimonious approximation. More generally, a wide variety of functions, e.g. those that are piecewise smooth having some discontinuities and those having inhomogeneous smoothness properties can be parsimoniously approximated by the set of wavelet basis functions, see Donoho and Johnstone (1994) and Donoho et al. (1995) for details. Periodicity of $\eta(x)$ may easily be employed using orthogonal trigonometric polynomials as described in the example.

Let Z be a $n \times n$ matrix with i-th column created by evaluating $\varphi_1(x), \ldots, \varphi_k(x)$ at the i–th sample point. In case of Demmler–Reinsch splines Z is an orthonormal matrix. For wavelet functions, orthonormality of Z holds provided n is a power of 2, $x_i = i/n$, and appropriate boundary conditions are incorporated. Applying the orthogonal transform $\tilde{c} = Z'y$, for $y = (y_1, \ldots, y_n)'$ we obtain empirical coefficients $\tilde{c}_k$ of the basis functions satisfying

$$y_i = \sum_{k=1}^{n} \tilde{c}_k \varphi_k(x_i).$$

With $\eta(x_i) = \sum_{k=1}^{n} c_k \varphi_k(x_i)$ we have $\tilde{c}_k = c_k + \tilde{\varepsilon}_k$, where $\tilde{\varepsilon}_k$ are independently identically distributed $N(0, \sigma^2)$. Hence, if an empirical coefficient is small compared to σ, then it consists mainly of noise. Moreover, due to the parsimonious approximation of $\eta(x)$, we know that only a small fraction of the c_k's are substantially different from zero. This leads to the following continuous soft–thresholding estimator

$$\begin{aligned} \hat{c}_k &= \operatorname{sgn}(\tilde{c}_k) \max(0, |\tilde{c}_k| - \lambda\sigma) \\ &= \operatorname{sgn}(\tilde{c}_k)(|\tilde{c}_k| - \lambda\sigma)_+, \end{aligned} \tag{2·1}$$

where $\tilde{c}_k$ are pulled towards zero by $\lambda\sigma$, $\lambda > 0$. Empirical coefficients $\tilde{c}_k$ with absolute value smaller than the noise level $\lambda\sigma$ are exactly set to zero. The estimate $\hat{\eta}(x)$ of $\eta(x)$ is easily obtained by back transforming $\hat{\eta} = Z\hat{c}$, $\hat{\eta} = (\hat{\eta}(x_1), \ldots, \hat{\eta}(x_n))'$.

To extend soft–thresholding to more general models in section 3·1, it is convenient to express the estimator (2·1) in terms of estimating equations. Introducing $e_k = \tilde{c}_k - \hat{c}_k$, soft–thresholding $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_n)'$ implicitly is defined by

$$\begin{aligned} |e_k| &\leq \lambda\sigma & &\text{if } \hat{c}_k = 0, \\ e_k &= \lambda\sigma & &\text{if } \hat{c}_k > 0, \\ e_k &= -\lambda\sigma & &\text{if } \hat{c}_k < 0. \end{aligned} \tag{2·2}$$

Compared to normal equations from linear models, where $e_k = 0$, the estimating equations (2·2) allow for $e_k \in [-\lambda\sigma, \lambda\sigma]$. When the absolute e_k is smaller than the threshold $\lambda\sigma$ for $\hat{c}_k = 0$, we use $\hat{c}_k = 0$ as estimate. Otherwise, we chose from all $\hat{c}_k$ having $e_k \in [-\lambda\sigma, \lambda\sigma]$ that one which comes closest to 0.

Donoho and Johnstone (1994) study the risk of soft–thresholding of the form (2·1) measured by quadratic loss at the sample points. From their work we conclude that soft–thresholding has superior performance when only few basis functions $\varphi_k(x)$ contribute essentially to $\eta(x)$ as assumed above. In case of correlated errors, where $\tilde{c}$ is distributed as $N(0, V)$, Johnstone and Silverman

(1997) derive similar results for a coordinate–wise soft–thresholding with σ in (2·1) replaced by $(V_{kk})^{1/2} = (Var(\tilde{c}_k))^{1/2}$.

3 Estimation of a univariate regression function in generalized linear models

For simplicity, we consider first a model where $E(y_i|x_i) = h(\eta(x_i))$ for $i = 1, \ldots, n$, and h is a prespecified response function. Our aim is to recover $\eta(x)$ from the data. Assuming that y_i is distributed according to a given exponential family as in (1·1), the log–likelihood contributions of each observation have the form $l_i(\theta_i) = (y_i\theta_i - b(\theta_i))/\phi_i$, where θ_i is some function of the predictor i.e. $\partial b(\theta_i)/\partial\theta_i = h(\eta(x_i))$. Summing up over i yields the log–likelihood of $\eta(x)$ given the data:

$$l(\eta) = \sum_{i=1}^{n}\{y_i\theta(\eta(x_i)) - b(\theta(\eta(x_i)))\}/\phi_i. \tag{3·1}$$

Unrestricted maximum likelihood estimation of $\eta(x_i)$ is then obtained by equating the score functions

$$\begin{aligned} s_i(\eta) &= \partial l(\eta)/\eta(x_i) \\ &= D(\eta(x_i))/\sigma^2(\eta(x_i))\{y_i - h(\eta(x_i))\}, \end{aligned} \tag{3·2}$$

to zero. Here $\sigma^2(\eta(x_i))$ denotes the variance of y_i and $D(\eta(x_i)) = \partial h(\eta(x_i))/\eta(x_i)$. From (3·2) follows that an unrestricted maximum–likelihood estimator satisfies $h(\eta^{ML}(x_i)) = y_i$ when it exists and thus stochastic errors from the observations are not eliminated leading to large variances of $\eta^{ML}(x_i)$. To suppress the noise in the estimator, modifications of the maximum likelihood principle are necessary.

3·1 *Generalized Soft–thresholding*

Analogous to Section 2 let us assume that $\eta(x)$ can be parsimoniously represented by a set of basis coefficients for $\varphi_k(x)$ as $\eta(x) = \sum_{k=1}^{n} c_k\varphi_k(x)$ and let

$$\begin{aligned} s_k(c) &= \partial l(\eta)/\partial c_k \\ &= \sum_{i=1}^{n}\varphi_k(x_i)s_i(\eta) \end{aligned} \tag{3·3}$$

denote score functions for each basis coefficient. Now suppose that, $c_h = 0$ for some h, then by $\eta(x) = \sum_{k \neq h} c_k \varphi_k(x)$ and $E_\theta(s_i(\eta)) = 0$ we have $E_\theta(s_h(c)) = 0$. Thus we expect that $s_k(c)$ varies around zero, if c_k is not very distinct from zero. This gives a first intuition about the generalized soft–thresholding estimator. For generalized linear models with non increasing score functions $-\partial s_i(\eta)/\partial\eta(x_i) \geq 0$, the estimator $\hat{c} = (\hat{c}_1, \dots \hat{c}_n)'$ is defined by its components $\hat{c}_k$, satisfying simultaneously one of the following conditions

$$\begin{aligned} |s_k(\hat{c})| &\leq \lambda\gamma_k \qquad && \text{if } \hat{c}_k = 0, \\ s_k(\hat{c}) &= \lambda\gamma_k && \text{if } \hat{c}_k > 0, \\ s_k(\hat{c}) &= -\lambda\gamma_k && \text{if } \hat{c}_k < 0, \end{aligned} \tag{3·4}$$

with $\lambda > 0$ a given trade–off parameter. In the modified score equations (3·4), we have replaced the left side of the estimating equations (2·2) by the score functions $s_k(\hat{c})$. The definition (3·4) is general in the sense, that it applies to response variables distributed according to an arbitrary exponential family. Moreover, we no more assume orthogonality of a design matrix built up by point evaluations of basis functions. Basically, generalized soft–thresholding has two ingredients: A set of basis functions together with a sequence of possibly different thresholds, $\gamma_1, \dots, \gamma_n$. Since in general, the score functions $s_k(c)$ are not identically distributed random variables we allow for separate thresholds for each basis function as in the coordinate–wise thresholding of Johnstone and Silverman (1997). For $\gamma_k = 0$, the conditions (3·4) reduce to the common maximum likelihood score equation $s_k(\hat{c}) = 0$ for coordinate k. Possible specifications for γ_k will be discussed subsequently in the text.

Figure 1 (a) illustrates generalized soft–thresholding for a logit model with $n = 1$. The estimator corresponds to the intersection of the score functions with the step function $\lambda\gamma_k \text{sgn}(c_k)$. In Figure 1 (b) we plotted the generalized soft–thresholding estimate against maximum likelihood estimates. For the outer left and outer right intersection point, corresponding to $y = 0$ and $y = 20$, respectively, the maximum likelihood estimator diverges. The heuristic of generalized soft–thresholding is that, if a coefficient $\hat{c}_k$ in (3·4) is set to zero, its score function or slope of the log–likelihood $s_k(c)$ evaluated at $\hat{c}_k = 0$ is smaller than $\lambda\gamma_k$. Hence a maximum likelihood estimator c_k^{ML} given $\hat{c}_1, \dots, \hat{c}_{k-1}, \hat{c}_{k+1}, \dots, \hat{c}_n$ is also

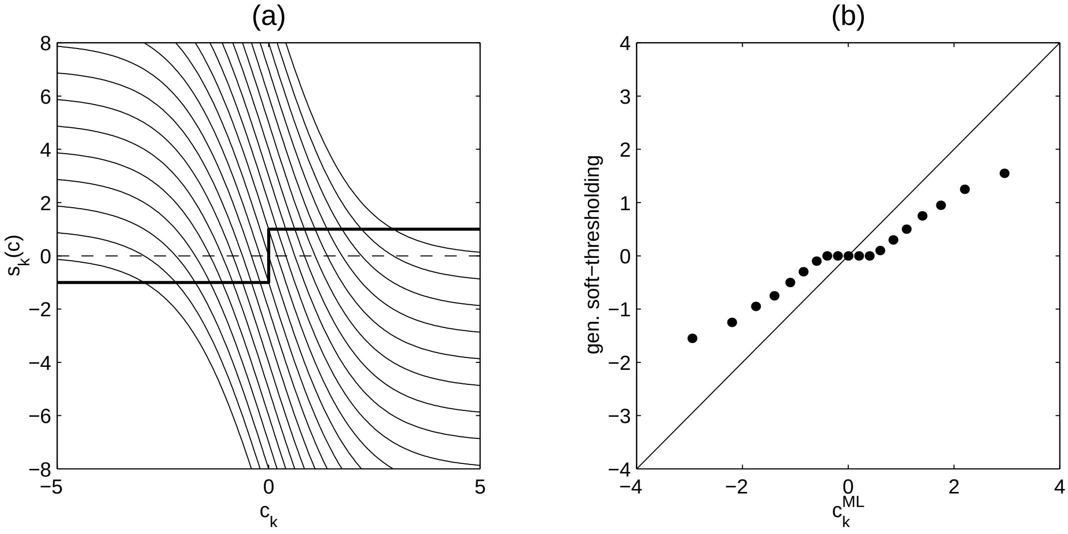

Figure 1: Univariate generalized soft–thresholding in a logit model. Score functions corresponding to a $B(20, p)$ distribution are drawn in (a) for $y = 0, \ldots, 20$. In (b) the generalized soft–thresholding estimator with $\lambda\gamma_k = 2.5$ is plotted against maximum likelihood estimates.

close to zero, or the likelihood is flat in this direction leading to a big variance of c_k^{ML}. Therefore including this coefficient cannot increase the likelihood more than inclusion of a covariate contributing mainly noise, and thus this coefficient is omitted. By adding a noise level $\lambda\gamma_k$ to the score function, non–zero coefficients are pulled towards zero compared to c_k^{ML}, which causes some bias. In (3·4), the bias variance trade–off is explicitly expressed by the parameter λ. We distinguish between two sources of bias: Some bias is due to the approximation of $\eta(x)$ by only some basis functions, regardless of the estimation procedure used. This kind of bias is referred to as approximation bias. Considering only the set of non–zero coefficients, additional bias is caused by equating the score functions as $s_k(\hat{c}) = \lambda\gamma_k \mathrm{sgn}(\hat{c}_k)$. In the following, this kind of bias is termed estimation bias. Both sources of possible bias are controlled by the trade–off parameter λ.

3·2 *Penalized likelihood estimation and spline smoothing*

In this subsection we discuss a specific set of basis functions together with a threshold sequence that mimics generalized spline smoothing. Within the penalized likelihood setting one tries to balance between fidelity to the data measured by the log–likelihood and roughness of the estimate. A popular penalized likelihood estimator is defined as the maximizer of

$$l(\eta) - \lambda \int (\eta^m(u))^2 du \tag{3·5}$$

over all functions in

$$\begin{aligned} W^m &= \{\eta : \eta \quad \text{has } m-1 \text{ absolutely continuous derivatives and} \\ &\quad \int (\eta^{(m)}(u))^2 du < \infty\}. \end{aligned}$$

O'Sullivan, Yandell and Raynor (1986) and Green and Silverman (1994) showed that the maximizer of (3·5) is a natural spline with knots at the design points $x_1, \ldots, x_n$. A specific basis for such smoothing splines was introduced by Demmler and Reinsch (1975), see also Eubank (1988), Ch.5. This orthogonal Demmler–Reinsch basis $\{\varphi_k(x)\}_{k=1}^n$ consists of natural splines satisfying

$$\begin{aligned} \sum_{i=1}^n \varphi_k(x_i)\varphi_j(x_i) &= \delta_{kj}, \\ \int \varphi_k^{(m)}(u)\varphi_j^{(m)}(u)du &= \delta_{kj}\gamma_k^2, \\ 0 = \gamma_1 = \cdots = \gamma_m < \gamma_{m+1} &\leq \cdots \leq \gamma_n, \end{aligned} \tag{3·6}$$

where $\delta_{kj} = I\{k = j\}$.

Figure 2 shows some of the Demmler–Reinsch functions computed by solving the corresponding eigenvalue problem as described in Eubank's book. The first basis functions $\varphi_1, \ldots, \varphi_m$ with $\gamma_1, \ldots, \gamma_m = 0$ span the space of polynomials of order m. For $k > m$, φ_k has exactly $k-1$ oscillations and its contribution to the penalty γ_k increases with k.

Assuming $\eta(x) = \sum_{k=1}^n c_k\varphi_k(x)$ together with (3·6) yields $\int(\eta^{(m)}(u))^2 du = \sum \gamma_k^2 c_k^2$ and the penalized likelihood criterion (3·5) can be written as

$$lp(c) = l(\eta) - \lambda \sum_{k=1}^n \gamma_k^2 c_k^2. \tag{3·7}$$

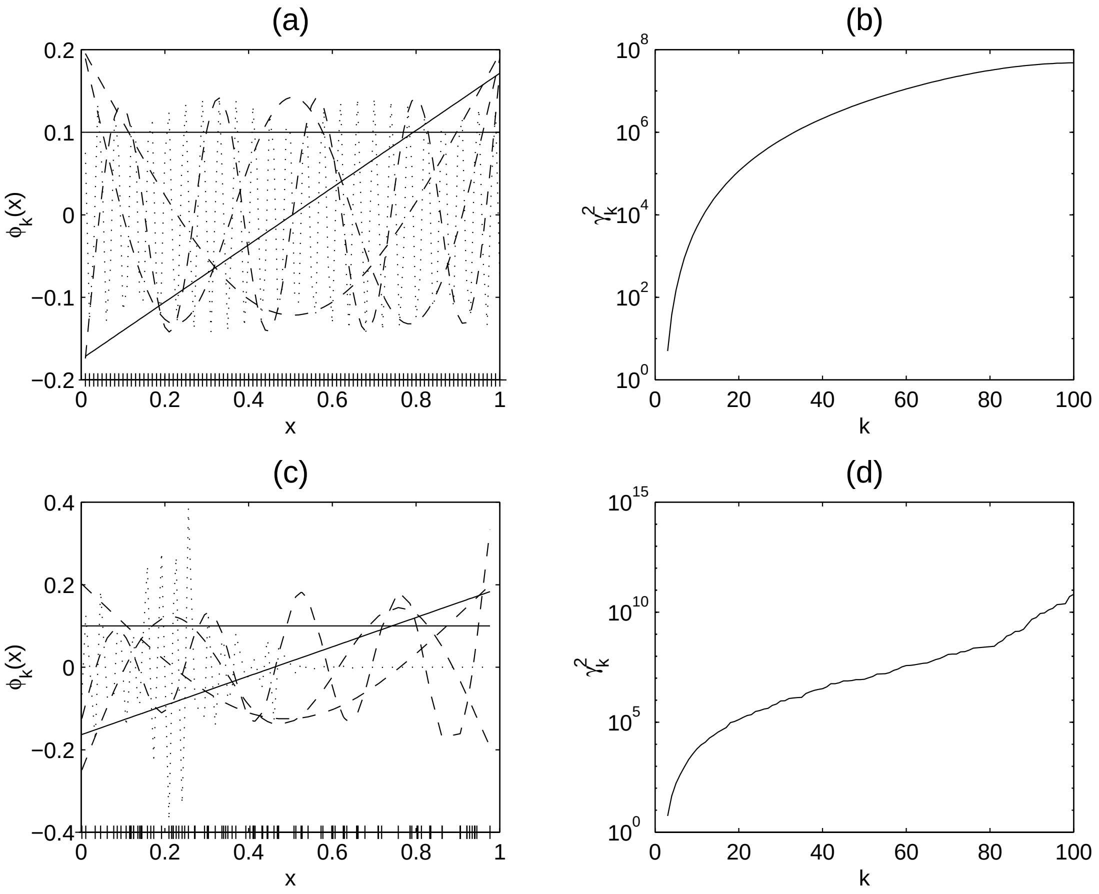

Figure 2: Demmler–Reinsch basis functions φ_1, φ_2 (solid) $\varphi_3, \varphi_5, \varphi_{10}$ (dashed) and φ_{50} (dotted) for $m = 2$, $n = 100$. (a) Equidistant design points. (b) Integrated squared curvature of φ_k for equidistant x_i. (c) Uniformly distributed design points. (d) Integrated squared curvature of φ_k for uniformly distributed x_i.

By (3·7) spline smoothing has the form of a generalized ridge estimator for the basis coefficients c_k, where no shrinkage applies to the null space spanned by polynomials of order m. Now, as inherent with smoothness, suppose that $\eta(x)$ is not too rough in the sense of $\int(\eta^{(m)}(u))^2 du$. Since γ_k^2 increases rapidly with k, it follows from (3·7) that most coefficients are near zero. As a consequence we get a parsimonious approximation of smooth $\eta(x)$ by only some of the first basis functions characterizing few sign changes or lower frequencies.

Figure 3 is typical for this situation: For the first function, having one maximum, the main systematic is described by the first three φ_k having up to two sign changes. The second, more complex shaped function is well approximated by the basis functions $\{\varphi_1, \ldots, \varphi_8\}$. In both situations, only few basis functions are necessary to keep the approximation bias reasonable small.

To recover systematics of the unknown function $\eta(x)$, we proceed by selecting only those basis functions which contribute essentially to $\eta(x)$ and estimate their coefficients c_k. This problem can be approached by introducing positive weights w_k in (3·7), leading to the weighted penalized likelihood criterion

$$lp(c,w) = l(\eta) - \lambda \sum_{k=1}^{n} \gamma_k^2 c_k^2 / w_k, \quad \sum_{k=1}^{n} w_k = 1. \tag{3·8}$$

In (3·8) a coefficient having small weight is strongly penalized, leading to $c_k = 0$ as $w_k \to 0$, whereas a coefficient with relatively big weight is less penalized compared to (3·7). Incorporating evidence from the data, we choose $\hat{w}_k$ as maximizer of $lp(c,w)$ over $w \in I\!R^n$. Langrangian calculus shows, that $\hat{w}_k$ becomes proportional to $\gamma_k |c_k|$ for $c_k \neq 0, \gamma_k > 0$. Substituting $w_k = \gamma_k |c_k|$ into (3·8) and demanding for a continuous penalty not penalizing $c_k = 0$, we obtain an absolute penalized likelihood estimator maximizing

$$lo(c) = l(\eta) - \lambda \sum_{k=1}^{n} \gamma_k |c_k|, \tag{3·9}$$

which is also considered in Tibshirani (1996) in the context of variable selection and shrinkage. The connection to soft–thresholding can be stated as follows:

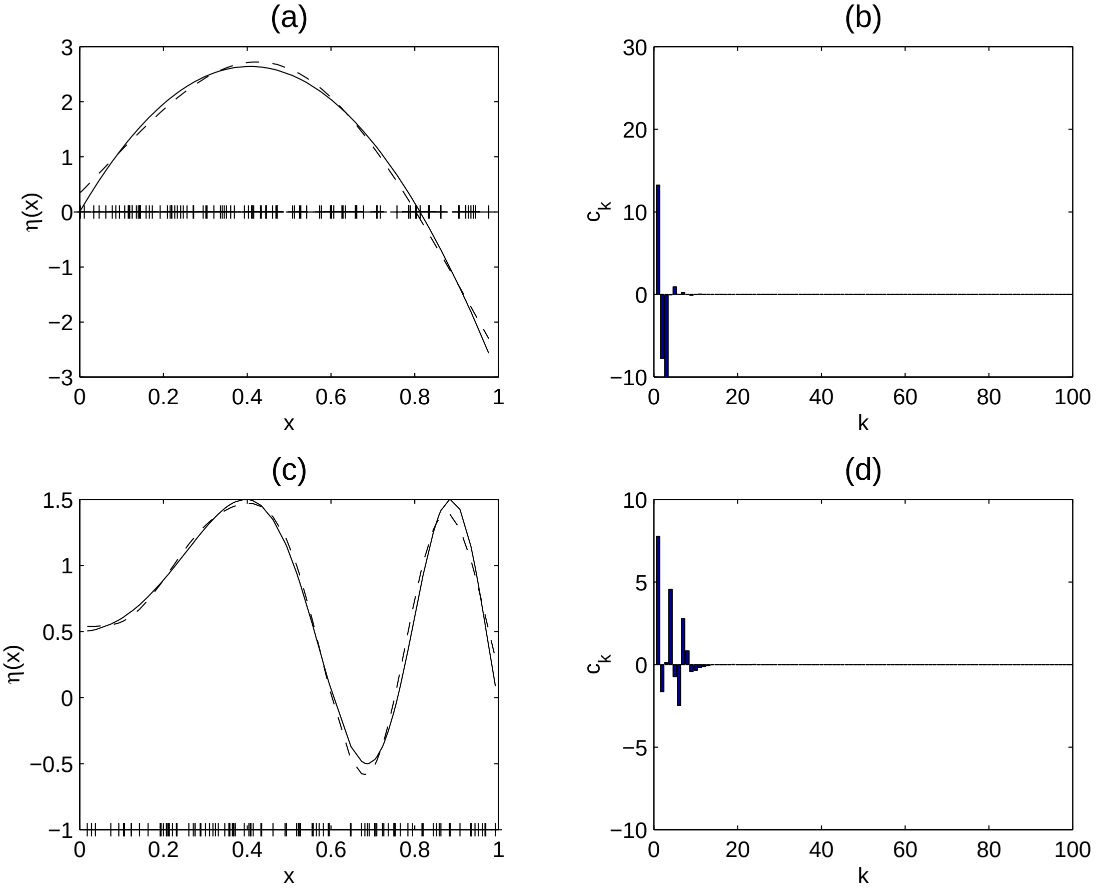

Figure 3: Coefficients of Demmler–Reinsch splines interpolating at hundred uniformly distributed sample points x_i. (a) $\eta_1(x) = x(11 - 16x)$ (solid) and approximation to $\eta_1(x)$ using $\varphi_1, \varphi_2, \varphi_3$ (dashed). (b) coefficients of $\eta_1(x)$ plotted versus number of oscillations. (c) $\eta_2(x) = \sin(10x^2) + 1/2$ and approximation to $\eta_2(x)$ using $\varphi_1, \ldots, \varphi_8$ (dashed). (d) coefficients of $\eta_2(x)$ plotted versus the number of oscillations.

Proposition 1: A maximizer of the absolute penalized log–likelihood (3·9) is necessarily a generalized soft–thresholding estimator as defined in (3·4). Moreover, let Z_B be a design matrix with columns composed by only those basis functions $\varphi_j(x_i)$ where $j \in \{k : \hat{c}_k \neq 0\}$ and let $H(\eta) = \frac{\partial^2 l(\eta)}{\partial \eta \partial \eta'}$ be the Hessian with respect to η. Then, if $-Z_B' H(\hat{\eta}) Z_B$ is positive definite at $\hat{c}$, generalized soft–thresholding is sufficient for a strict local maximum of the absolute penalized log–likelihood (3·9). (Proof in Appendix)

Since generalized soft–thresholding can be described as a penalized likelihood estimator incorporating a convex penalty function, existence and uniqueness is guaranteed also in cases where unrestricted likelihood estimation fails. As stated in Proposition 1, full rank is only required for a submatrix Z_B of the actual design matrix Z, which is defined by non–zero coefficients $\hat{c}_k \neq 0$. Therefore, it is even possible to supply a design matrix Z where columns are not independent and also different sets of basis functions can be combined in one design matrix. In practice, the non–zero pattern of the coefficient vector $\hat{c}$ depends in a complex way on the threshold sequence and on the actual data. Consequently, uniqueness conditions are difficult to check a priori and we recommend to watch convergence of the algorithm proposed in Section 6.

So far, we have restricted attention to spline–smoothing. If we were in favorite of an alternative smoothing operator, we can adopt the ideas in Hastie (1996) leading to pseudo splines. Basically, any linear smoother providing a symmetric smoothing matrix $S(\lambda)$ can be used in connection with generalized soft–thresholding. Within this framework, the point evaluations of the basis functions, $\varphi_k(x_i)$ correspond to the eigenvectors of $S(\lambda)$, and the threshold sequence is built up by the eigenvalues $\theta_k^\lambda \in (0,1]$ of $S(\lambda)$ as $\lambda\gamma_k = \left(1/\theta_k^\lambda - 1\right)^{1/2}$. Avoiding expensive eigendecompositions, Hastie gives an efficient algorithm for approximating the first eigenvalues and eigenvectors based only on applications of a given smoother. Computing the pseudo eigendecomposition of a specified smoother having desirable characteristics, generalized soft–thresholding can be customized in many ways. When many design points x_i are involved, computation of Demmler–Reinsch splines by expensive eigenvalue decompositions becomes too demanding. Then the pseudo spline algorithm provides an attractive alternative for approximating the first basis functions and thresholds needed.

To assess properties of the estimator, we compare it to spline–smoothing in a log–linear Poisson model. The observations y_i are distributed according to $y_i \sim \mathrm{Po}\left\{\exp(\eta(x_i))\right\}$ and hundred x_i were drawn from the uniform distribution $U(0, 1)$. Figure 4 shows the results computed from 1000 simulations using the 'true' functions $\eta_1(x)$ and $\eta_2(x)$ already considered in Figure 3. To neglect influences of the trade–off parameter in interpreting results, the smoothing parameter is chosen to minimize $\sum\left(\hat{\eta}_\lambda(x_i) - \eta(x_i)\right)^2$ over λ in each run. In Figure 4 (c) we can see, that for a function having constantly low second derivative, apart from the boundaries the bias is quite small for both methods. At the right boundary, soft–thresholding has a slightly lower bias compensated by a bigger variance, shown in Figure 4 (e). For the more wiggly function $\eta_2(x)$, Figure 4 (d) reflects the well–known fact, that the bias of cubic smoothing–splines is higher in areas with high curvature of $\eta_2(x)$, compare Figure 3 (c). This high curvature region at $x = 0.7$ mainly is described by φ_k with k between 4 and 8 having rather big coefficients c_k. Generalized soft–thresholding shows reduced bias there, because it penalizes those coefficients less. Considering the representation as weighted penalized likelihood estimator from (3·8), generalized soft–thresholding puts increased weights on those c_k contributing to the curvature at $x = 0.7$. Consequently, we observe a local reduction of bias and an increase of variance in this region, shown in Figure 4 (f). As conclusion we state that by reducing the explicit dimension, the soft–thresholding methodology produces estimates having about the same mean squared error than spline–smoothing. In the simulation shown, the median number of non–zero coefficients for estimating $\eta_1(x)$ was 6, whereas for $\eta_2(x)$ a median number of 10 coefficients were estimated unequal to zero.

3·3 *Locally adaptive function estimation*

In the last subsection, we considered $\eta(x)$ to be homogeneously smooth and obtained a parsimonious approximation by Demmler–Reinsch splines. Now, suppose we want to recover another though simply structured function of the form $\eta(x) = I\{x \geq x_k\}$ for some k. In the Demmler–Reinsch domain, such a $\eta(x)$

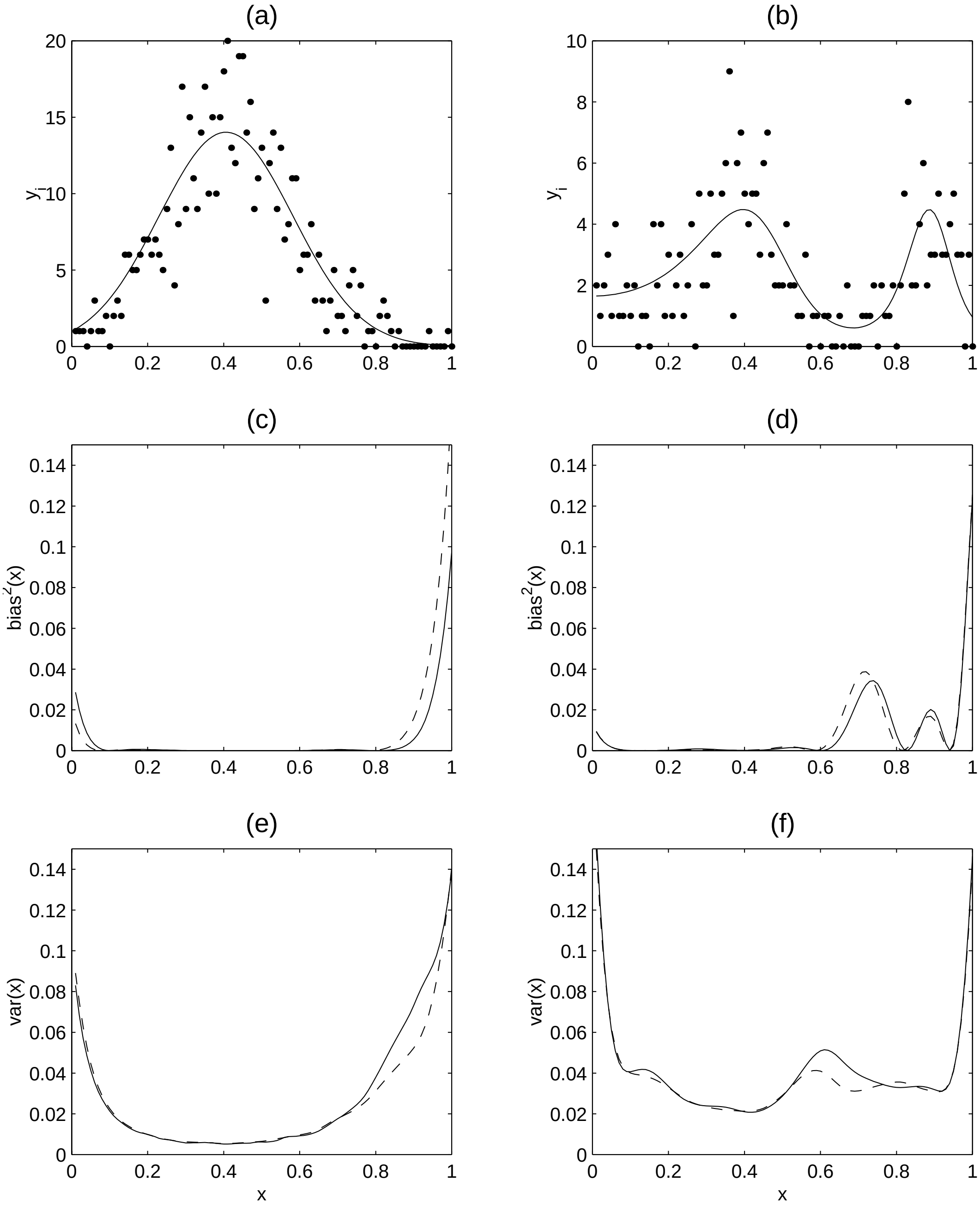

Figure 4: Squared bias and variance of generalized soft–thresholding and spline–smoothing computed from 1000 simulations. One simulated data set and true mean function $\exp(\eta_1(x))$ in (a) and $\exp(\eta_2(x))$ in (b), respectively. Squared bias of generalized soft–thresholding (solid) and spline–smoothing (dashed) are drawn in (c) for $\eta_1(x)$ and in (d) for $\eta_2(x)$. Corresponding variances of the estimates are showm in (e) and (f), respectively.

has a quite disadvantageous representation. Similarly to the approximation of a heaviside function by trigonometric polynomials, a high frequency component is needed to describe the jump at x_k, whereas for compensating this high frequency component outside of x_k many non–zero coefficients have to be employed.

Clearly, the set of indicator functions itself, $\varphi_k^I(x) = I\{x \geq x_k\}$ provides a most parsimonious approximation for such problems and one simply might supply $\{\varphi_k^I(x)\}, k = 1, \dots n$ to the generalized soft–thresholding estimator (3·4). By $\eta(x_i) = \sum c_k \varphi_k^I(x_i) = \sum_{k \leq i} c_k$ we have $c_i = \eta(x_i) - \eta(x_{i-1})$ and therefore, generalized soft–thresholding corresponds to a maximizer of the penalized log–likelihood criterion

$$lo(\eta) = l(\eta) - \lambda \sum_{i=2}^{n} \gamma_i |\eta(x_i) - \eta(x_{i-1})| - \lambda \gamma_1 |c_1|, \qquad (3\cdot 10)$$

in this situation. The threshold sequence may be chosen constantly, $\gamma_i = 1$, $i = 2, \dots, n$ or according to the distance of the design points, as $\gamma_i = (x_i - x_{i-1})^{-1}$, for example. For the first basis function $\gamma_1 = 0$ is a suitable choice, since $\varphi_1^I(x) \equiv 1$ corresponds to a global intercept term. Examining a representation as weighted penalized log–likelihood estimator as in (3·8), the adaptivity of estimator (3·10) becomes obvious. Compared to a discrete version of spline–smoothing, where the log–likelihood is penalized by $\lambda \sum_{i=2}^{n} (\eta(x_i) - \eta(x_{i-1}))^2$, soft–thresholding (3·10) implicitly incorporates weights varying over x proportional to $|\hat{\eta}(x_i) - \hat{\eta}(x_{i-1})|$. In the context of penalized least squares estimation, Mammen and van de Geer (1997) study general total variation penalties similar to (3·10) and derive essentially optimal rates of convergence in spatially inhomogeneous bounded variation function classes. The authors also propose general locally adaptive regression splines, where the total variation of the m-th derivative, $\eta^{(m)}(x)$ is penalized. In our framework, we analogously extend the concept of indicator functions to one–sided splines by supplying $\varphi_k(x) = (x - x_k)_+^m$, with $\varphi_k(x)^{(m)} = \varphi_k^I(x)$, together with non penalized polynomial terms up to order m. By its selective property, our estimator provides a spline function having adaptively chosen knot points. In this configuration, generalized soft–thresholding is an alternative to the adaptive regression spline methodology of Stone et al. (1997) for estimation in extended linear models. The ability of doing knot selection and parameter estimation simultaneously appears to be of particular attraction for splined models of this kind.

Suppose further, that $\eta(x)$ is well approximated by piecewise polynomials, where pieces are not smoothly joint together, i.e. $\eta(x)$ contains jumps. For this class of functions, one–sided splines as considered above are not a best choice, since $\varphi_k^I(x)$ is doing well in describing the jump, but badly in approximating the polynomial elsewhere, whereas for higher order splines the polynomial is approximated perfectly but many knots have to be employed for approximating the jump. In this situation, wavelet basis functions provide a parsimonious approximation as stated e.g. in Donoho and Johnstone (1994). Very briefly, wavelets refer to an orthogonal system of compactly supported basis functions. Their main contribution is to combine exact representation of polynomials and local support. By this property, a wide variety of functions, including piecewise polynomials, have a parsimonious representation in the wavelet domain, compare Daubechies (1992), Ch.9. When n is a power of two and the design points are equidistant, wavelet coefficients are extremely fast computable by the fast wavelet transform. In other cases, some extra interpolation has to be incorporated. We then replace $\eta(x_i)$ by the linear interpolant between $\eta(x_l)$ and $\eta(x_{l+1})$, where x_l and x_{l+1} are two adjacent neighbours in a dyadic grid on $[x_1, x_n]$. In principle non–equidistant design points can be handled in the same way. For very irregularly spaced x_i however, this procedure may degenerate and spline functions should be considered. Threshold sequences can be based on the dyadic structure of wavelet functions. Usually a coarse resolution level J_0 corresponding to some kind of trend is not penalized, i.e $\gamma_k = 0$ for $k = 1, \ldots, 2^{J_0} - 1$. For the remaining coefficients, one can use one global threshold or alternatively, one uses different thresholds according to the resolution level J, as e.g. 2^{-J}. The latter choice puts higher penalties on high frequencies and thus produces results of smooth appearance.

As will be demonstrated by the application in Section 7, one can make use of advantages that different sets of basis functions offer. By supplying them jointly to the estimator, appropriate basis functions from each set are selected. For example, smooth functions having only few jumps are well described by Demmler–Reinsch splines together with indicator functions. A similar strategy is proposed by Chen and Donoho (1995) for obtaining optimal signal decompositions. When configuring the estimator with basis functions from different sets, one has to account for their scaling. We allow for different scalings of basis functions, by adjusting the threshold values γ_k appropriately as described in Subsection 4·2. For ordinally scaled x_i, the set of indicator functions $\varphi_k^I(x)$ can be supplied, and generalized soft–thresholding trys to join adjacent categories to obtain a parsimonious representation.

4 Estimation of varying coefficients

Let $\eta_i = \eta(X_i)$ denote the predictor, connected by $E(y_i|X_i) = h(\eta_i)$ to an observation y_i which is distributed according to a specified exponential family. The varying coefficient–model assumes linearity of the predictor given the covariate values $x_{ij}, z_{ij} \in X_i$, $j = 1, \ldots, p$ as

$$\eta_i = \beta_0(x_{i0}) + \beta_1(x_{i1})z_{i1} + \ldots + \beta_p(x_{ip})z_{ip}. \tag{4·1}$$

Unrestricted maximum likelihood estimation of the coefficients $\beta_j(x_{ij})$ usually yields highly variable estimates as pointed out in the function estimation setting in Section 3. Hence further assumptions are incorporated. In our framework we assume, that each varying coefficient $\beta_j(x_j)$, $j = 1, \ldots, p$ can parsimoniously be well approximated by possibly different sets of basis functions $\{\varphi_{jk}\}$, $k = 1, \ldots, n_j$ as $\beta_j(x_{ij}) = \sum_k c_{jk}\varphi_{jk}(x_{ij})$.

Incorporating the multiplicative covariates z_{ij} and $z_{i0} = 1$, the basis coefficients are linked by

$$\eta_i = \sum_{j=0}^{p}\sum_{k=1}^{n_j} c_{jk}\varphi_{jk}(x_{ij})z_{ij}$$

to the predictor and the score functions for each basis coefficient, $s_{jk}(c) = \partial l(\eta)/\partial c_{jk}$ are given by

$$s_{jk}(c) = \sum_{i=1}^{n} z_{ij}\varphi_{jk}(x_{ij})s_i(\eta_i) \tag{4·2}$$

where $s_i(\eta_i) = \partial l_i(\eta)/\partial\eta_i$ are individual score contributions.

4·1 *Generalized Soft–thresholding*

For varying–coefficient models, the generalized soft–thresholding estimator from (3·4) extends to

$$\begin{aligned} |s_{jk}(\hat{c})| &\le \lambda\gamma_{jk} && \text{if } \hat{c}_{jk} = 0, \\ s_{jk}(\hat{c}) &= \lambda\gamma_{jk} && \text{if } \hat{c}_{jk} > 0, \\ s_{jk}(\hat{c}) &= -\lambda\gamma_{jk} && \text{if } \hat{c}_{jk} < 0 \end{aligned} \tag{4·3}$$

and estimates of the varying coefficients are obtained as $\hat{\beta}_j(x_j) = \sum_k \hat{c}_{jk}\varphi_{jk}(x_j)$. The threshold sequence γ_{jk} is based on thresholds for univariate function estimation, considered in the previous section. Since inclusion of multiplicative covariates z_{ij} effects the magnitude of the the score functions in (4·2), appropriate choice of γ_{jk} becomes more crucial for varying–coefficient models. In Subsection 4·2 we propose a scaling procedure to account for the covariate design.

The connection between generalized soft–thresholding and absolute penalized likelihood estimation, stated in Proposition 1, remains unchanged and the estimator corresponds to a maximizer of

$$lo(c) = l(\eta) - \lambda \sum_{j=0}^{p} \sum_{k=1}^{n_j} |c_{jk}|. \tag{4·4}$$

For a sufficiently large trade–off parameter λ, generalized soft–thresholding (4·3) becomes a maximum likelihood estimator of a common generalized linear model, where only coefficients $c_{jk} \in \mathcal{M}_0$ with $\mathcal{M}_0 = \{jk : \gamma_{jk} = 0\}$ are included. We refer to that model as the embedded model $\eta^{(0)}$ and assume that a maximum likelihood estimator for corresponding coefficients exists. This embedded model is contained in any generalized soft–thresholding estimate and represents a coarse frame of the varying coefficient model. Often the embedded model is set up by linear interaction terms as

$$\eta_i^{(0)} = \beta_{00} + \beta_{01}x_{i0} + \beta_{11}z_{i1} + \beta_{10}z_{i1} + \beta_{11}x_{i1}z_{i1} + \ldots + \beta_{p0}z_{ip} + \beta_{p1}x_{ip}z_{ip}. \tag{4·5}$$

When describing each varying effect by cubic Demmler–Reinsch splines, the model (4·5) corresponds to the null space of the penalty function $\lambda \sum_{j=0}^{p} \sum_{k=1}^{n_j} |c_{jk}|$, which is set up by all basis functions not penalized. In the case of purely additive terms, e.g. $\eta = \beta_1(x_1) + \beta_2(x_2)$, or when multiplicative covariates z_{ij} appear several times in the predictor, appropriate constant terms have to be removed from the set of basis functions to ensure identifiability of the embedded model (4·5). This strategy leads to centered estimates of $\beta_j(x_j)$ which are known from additive models, see e.g. Hastie and Tibshirani (1990).

4·2 *Scaling of the thresholds*

In the modified score equations (4·3), the variation of the score function $s_{jk}(\hat{c})$ depends on the scaling of the basis functions and the multiplicative covariates z_j as well as on the true predictor η_i. A simple way to make the estimator more invariant against different scalings of covariates and basis functions is to use standardized versions of $Z_{jk} = \{z_{1j}\varphi_{jk}(x_{1j}), \ldots, z_{nj}\varphi_{jk}(x_{nj})\}'$ having $\bar{Z}_{jk} = 0$ and $Z'_{jk}Z_{jk} = 1$. This strategy accounts for single covariates, but not for the global structure of the model. Therefore, additional information from the actual design is incorporated. We avoid blowing up the number of trade–off parameters by appropriate scaling of the threshold values γ_k as introduced in Section 3.

Our scaling procedure is based on connecting the modified score equations (4·3) to score tests for a null hypotheses $c_{jk} = 0$. We start with a maximum likelihood estimate of the embedded model $\hat{c}^{(0)}$ having design matrix Z_0 and consider a test for inclusion of another basis function φ_{jk}. This is done by using the score statistic

$$U_{jk} = \tilde{s}_{jk}(\hat{c}^{(0)})'\tilde{F}_{jk}(\hat{c}^{(0)})^{-1}\tilde{s}_{jk}(\hat{c}^{(0)}), \tag{4·6}$$

as approximation to the likelihood ratio. In (4·6), $\tilde{s}_{jk}$ denotes a score vector composed by all coefficients used in the embedded model together with one supplementary basis function to test on, i.e.

$$\tilde{s}_{jk} = (Z_0, Z_{jk})'\frac{\partial l(\eta^{(0)})}{\partial \eta}.$$

The matrix $\tilde{F}_{jk}$ is the matching Fisher information matrix

$$\tilde{F}_{jk}(c^{(0)}) = -(Z_0, Z_{jk})'E\left(\frac{\partial l(\eta^{(0)})}{\partial c \partial c'}\right)(Z_0, Z_{jk}).$$

Since $\tilde{s}_{jk}(\hat{c}^{(0)}) = 0$ for $jk \in \mathcal{M}_0$ the test statistic (4·6) reduces to $U_{sk} = s^2_{jk}(\hat{c}^{(0)})\sigma^{-2}_{jk}(\hat{c}^{(0)})$, where $\sigma^{-2}_{jk}(\hat{c}^{(0)})$ is the last diagonal element of $\tilde{F}_{jk}(\hat{c}^{(0)})^{-1}$. Substituting γ_{jk} in (4·3) by $\sigma_{jk}(\hat{c}^{(0)})$, the first modified score equation can be regarded as a test on $c_{jk} = 0$, where λ is some quantile of the standard normal distribution.

To adjust the thresholds, let $\{\gamma_k(j)\}$ denote the threshold sequence corresponding to the set of basis functions $\{\varphi_{jk}\}$ as in Section 3. When $\tilde{F}_{jk}$ is non–singular, the scaled threshold $\gamma_{jk} = \sigma_{jk}(\hat{\eta}^0)\gamma_k(j)$ is used in the modified score equations (4·3) to account for the variation of the score function s_{jk}. In the case of singular $\tilde{F}_{jk}$, the additional basis function explains variation already explained by the embedded model and thus, we remove φ_{jk} from the set of possible basis functions. In contrast to simple standardization, this strategy additionally accounts for correlations to the embedded model as a coarse frame of the varying–coefficient model.

When different sets of basis functions are used, additional considerations for appropriate scaling can become necessary. Consider for example, that $\{\varphi_{jk}\}$ is built up by the set of indicator functions φ^I_{jk} together with Demmler–Reinsch splines. For these splines, $\gamma_k(j)$ increases with the basis functions frequency and we adjust the thresholds $\gamma^I_k(j)$ for φ^I_{jk} on one $\gamma_k(j)$ corresponding to a specified number of sign changes. In principle, this strategy can be regarded as an additional trade–off between the coefficients frequency component and its tendency to have distinct breakpoints.

Some attention has to be drawn in choosing the embedded model. When splines are employed, it is quite natural to use polynomial terms. In the case of wavelet or Fourier representations for the varying effects, a proper choice of the embedded model becomes more crucial. For example, when wavelets are used, the choice of the coarse resolution level J_0 can have some impact on the estimates. For ordinary soft–thresholding of wavelet coefficients, this phenomenon has been studied in detail in Marron, Adak, Johnstone, Neumann and Patil (1995).

We close this section by demonstrating the finite sample performance of the estimator in a simulation study that will be continued in Section 5. In each of the 1000 runs, 200 observations were drawn according to $y_i \sim \mathrm{Po}(\exp(\eta_i))$, where $\eta_i = \beta_1(x_i) + \beta_2(x_i)z_i$. For x_i we used an equidistant grid $0.01, 0.02, \ldots, 1$ and for each grid point we simulated two observations y_i by setting $z_i = 0$ and $z_i = 1$, respectively. The varying effects are derived from the functions already

	$\beta_1(x)$				$\beta_2(x)$			
γ^I	Spline		Indicator		Spline		Indicator	
20	1.805	(0.581)	8.714	(2.280)	2.400	(0.720)	14.532	(2.488)
100	6.156	(1.436)	6.349	(1.915)	6.864	(1.145)	5.268	(1.878)
200	10.746	(1.643)	4.568	(1.508)	9.737	(1.523)	3.672	(1.541)

Table 1: Average number of coefficients for Demmler–Reinsch splines and indicator functions that are estimated unequal zero. The numbers in brackets correspond to the standard deviation computed from 1000 simulations.

considered in Figure 3 by adding two breakpoints on $\eta_1(x)$ and one breakpoint on $\eta_2(x)$. Both effects were estimated by combining cubic Demmler–Reinsch splines with the set of indicator functions φ_k^I. We removed the constant term from φ_k^I, since it is already contained in the set of orthogonal splines. The embedded model consists of four coefficients, representing linear terms for $\beta_1(x)$ and $\beta_2(x)$, respectively. As threshold for the indicator functions we used the values $\gamma^I = 20, 100, 200$, corresponding approximately to γ_6,γ_{12} and γ_{16} from the spline basis. The global threshold is set to $\lambda = 2/(3\gamma^I)$ resulting in about 30 basis functions used in total to represent the predictor. Table 1 shows, how the basis functions not contained in the embedded model are distributed over the estimates.

For all simulations, the breakpoints were found properly. When $\gamma^I = 20$ is used, the descent of $\beta_2(x)$ in $[0.5, 0.7]$ is represented by the indicator functions and the maximum at $x = 0.8$ is not recognized, see Figure 5 (b) and the last column in Table 1. Obviously, $\gamma^I = 100$, where about the same number of spline and indicator functions are used, is a better choice. For $\gamma^I = 200$ the estimates tend to be too wiggly, compare Figures 5 (e) (f). When mainly Demmler–Reinsch splines are used, the estimation error at the boundary for $\beta_2(x)$ at $[0.0, 0.2]$ is higher, see Figures 5 (d) (f).

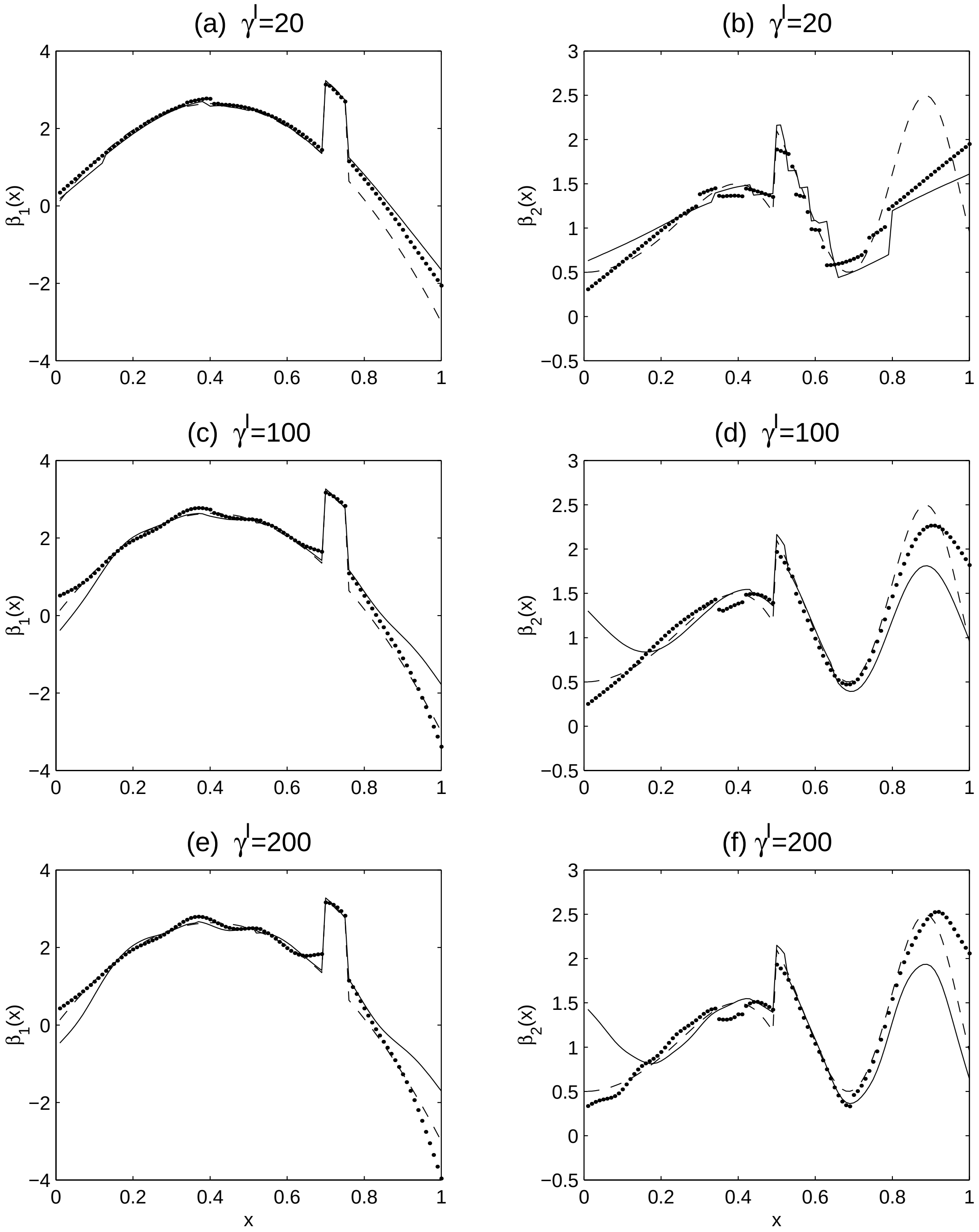

Figure 5: Estimation number 100 (dotted) and 900 (solid) from 1000 simulations ordered by the total estimation error $\sum_{ij}(\hat{\beta}_j(x_i) - \beta_j(x_i))^2$, when $\gamma^I = 100$ was used. The true functions $\beta_1 = \eta_1 + I\{x \in [0.7, 0.75]\}$ and $\beta_2 = \eta_2 + I\{x \geq 0.5\}$ are drawn as dashed lines.

5 Inference

Considering generalized soft–thresholding as a preprocessor for selecting an appropriate parametric model, analysis of deviance can be based on maximum likelihood estimates using only the obtained non–zero coefficients in a second step. In more complex situations, however, the maximum likelihood estimator becomes highly variable or even diverges due to the high number of parameters involved. Then, using a submodel consisting of very few basis functions only may lead to increased approximation bias caused by a too parsimonious approximation of $\beta_j(x_j)$.

To obtain more stability without increasing approximation bias, we propose to base inference directly on the generalized soft–thresholding estimate. More specifically, suppose to test for the hypothesis $c_{jk} = 0$ for some of the coefficients. This covers following interesting applications:

Situation 1: Test on any linear or nonlinear effect of covariate z_j: $c_{j1} = \ldots = c_{jn_j} = 0$.

Situation 2: Test on nonlinearity of $\beta_j(x_j)$: $c_{j3} = \ldots = c_{jn_j} = 0$ when cubic Demmler–Reinsch splines are used.

Situation 3: Test on a breakpoint of $\beta_j(x_j)$ in x_k: $c_{jk} = 0$ when indicator functions are used.

Situation 4: Semiparametric models: $c_{j1} = 0$ when only $\varphi_1^I(x_j) \equiv 1$ is supplied for covariate z_j

Formally, the hypothesis of the tests has to be fixed in advance, regardless of the soft–thresholding estimate. In this sense, our approach can be regarded as semiparametric. The coefficients under test are specified parametrically and the procedure accounts for not explicitly specified factors. Informally, we use test statistics based on estimated non–zero coefficients to suggest presence of specific effects.

In the following, we derive a quadratic approximation to the maximum likelihood test for a general linear hypothesis $Ac = 0$ comprising all four situations. First, assume that a good set of basis functions approximating the true varying coefficients is found by generalized soft–thresholding and the approximation bias becomes neglectable. Usually, this assumption can be fulfilled by using a reasonable small trade–off parameter λ leading to possible overfit of the data. In this setting it is sufficient to base inference only on the coefficients under test together with selected non–zero coefficients. Let $\mathcal{B}$ be the set of all coefficients estimated unequal zero and let $\mathcal{B}_A$ be the set of coefficients, that are used to formulate the hypothesis $Ac = 0$. Note that $\mathcal{B}_A$ is not necessarily a subset of $\mathcal{B}$. Suppose that the true model can be represented by basis functions from the set $\mathcal{B}_1 = \mathcal{B} \cup \mathcal{B}_A$ with corresponding coefficient vector c_1. Subsequently, we consider only coefficients with basis functions from the set $\mathcal{B}_1$. Soft–thresholding estimates are stored in the coefficient vector $\hat{c}_S$ composed by basis coefficients from $\{\hat{c}_{jk} : jk \in \mathcal{B}_1\}$. The test statistic is then derived using the quadratic approximation

$$Q(c) = l(\hat{c}_S) + s_1(\hat{c}_S)'(c_1 - \hat{c}_S) + \frac{1}{2}(c_1 - \hat{c}_S)'H_1(\hat{c}_S)(c_1 - \hat{c}_S) \tag{5·1}$$

to the log–likelihood. Here $H_1(c) = -\partial l(c)/(\partial c_1 \partial c_1')$ denotes the negative Hessian or observed information with respect to the basis coefficients and $s_1(c) = \partial s(c)/\partial c_1$. Maximizing the quadratic form $Q(c)$ over all coefficients in the set $\mathcal{B}_1$ under the restriction $Ac = 0$ and without restriction yields the following modified Wald test:

Proposition 2: Let $H_1(c)$ be the negative Hessian with respect to c_1 and let $H_1(\hat{c}_S)$ be positive definite, then

$$2\,[Q(\hat{c}_1) - Q(\hat{c}_0)] = (A\hat{c}_1)'[AH_1(\hat{c}_S)^{-1}A']^{-1}(A\hat{c}_1), \tag{5·2}$$

where

$$\begin{aligned} \hat{c}_1 &= \hat{c}_S + H_1(\hat{c}_S)^{-1}s_1(\hat{c}_S) \\ \hat{c}_0 &= \hat{c}_1 - H_1(\hat{c}_S)^{-1}A'[AH_1(\hat{c}_S)^{-1}A']^{-1}A\hat{c}_1 \end{aligned} \tag{5·3}$$

are estimates based on the quadratic form (5·1) satisfying $A\hat{c}_0 = 0$. (Proof in Appendix)

Generally, (5·2) can be regarded as a Wald test on corrected estimates $\hat{c}_1$. In the case when all coefficients to test on are estimated to zero, we have $\hat{c}_1 =$

γ^I	$\hat{\beta}_1(x)$		$\hat{\beta}_1^{cor}(x)$		$\hat{\beta}_2(x)$		$\hat{\beta}_2^{cor}(x)$	
20	0.102	(0.087)	0.056	(0.006)	0.131	(0.106)	0.114	(0.046)
100	0.059	(0.015)	0.103	(0.008)	0.093	(0.037)	0.137	(0.020)
200	0.103	(0.021)	0.174	(0.021)	0.134	(0.036)	0.199	(0.020)

Table 2: Average mean squared error for generalized soft–thresholding with bias correction. The numbers in brackets correspond to the averaged squared bias computed from 1000 simulations.

$H_1(\hat{c}_S)^{-1}s_1(\hat{c}_S)$ and the test statistic is similar to a score test. The correction of the estimates is equivalent to one step of a Fisher scoring iteration for a maximum likelihood estimate of c_1. As consequence, corrected estimates are closer to corresponding maximum likelihood estimators and estimation bias is decreased. For normally distributed data, $\hat{c}_1$ coincides with the least squares estimate of c_1 and might be regarded as a hard–threshold estimator. In Table 2 we report averaged mean squared errors for the parameter estimates shown in Figure 5 together with the corresponding bias part. For the estimation influenced mainly by indicator functions, the bias reduces drastically, resulting in a lower averaged mean squared error. The reduction of bias is smaller when orthogonal splines are dominant. Here the averaged mean squared error increases. Due to the increasing threshold sequence, more correction is done on high frequency basis functions representing less variation of the true $\beta_j(x_j)$. This causes an increase of variance for high frequency spline basis functions. Consequently, the bias corrected estimates tend to be more wiggly and are visually less favourable then generalized soft–thresholding estimates. In contrast to parameter estimation or recovery, where one focuses on mean squared error, bias has to be reduced for inferential purposes, as pointed out e.g. by Speckman (1988) in the context of semiparametric models.

Recall the definition of generalized soft–thresholding (4·3) based on the slope of log–likelihood and suppose that λ is sufficiently small. Then, following Subsection 3·1, one might argue, that measured by the log–likelihood, the estimator $\hat{c}_1$ is close to a maximum likelihood estimator of c_1. This encourages to approximate the distribution of the test statistic

$$T = (A\hat{c}_1)'[AF_1(\hat{c}_S)^{-1}A']^{-1}(A\hat{c}_1), \qquad \hat{c}_1 = \hat{c} + F_1^{-1}(\hat{c}_S)s_1(\hat{c}_S) \tag{5·4}$$

by the distribution of a corresponding maximum likelihood ratio test. In (5·4) the observed information is replaced by the expected information, as conventional in generalized linear models. Provided a rather small dimension of the null hypothesis (i.e. $\text{rank}(A) < 10$ for situations considered in this paper), we observed in simulation studies that a χ^2 distribution having $\text{rank}(A)$ degrees of freedom works well as approximation. When the number of parameters involved is bigger, or the main interest of investigation is testing, bootstrap approaches should be used to assess the distribution of T under the null hypothesis. For normally distributed response variables, of course, the χ^2 approximation is correct when no approximation bias occurs.

Biased estimation of coefficients contained in the hypothesis can also be due to correlated biased estimates of coefficients not formulated in the hypothesis. Therefore, one should also investigate in the matrix $F_1(\hat{c})^{-1}$ to detect possible correlations in the estimates. For test situations 1 and 2 a considerable increase of power can be obtained by imposing smoothness restrictions. Then, the hypothesis is set up only by coefficients of Demmler–Reinsch splines having up to a moderate number of zero crossings. For example, we use only the first 10 basis functions, regardless of the number of observation points.

Figure 6 shows results of a simulation study for test situation 3. Considering the model used in Figure 5, we tested the hypothesis $c_{2,150} = 0$ corresponding to an effect of the basis function $\varphi^I_{2,50}(x) = I\{x \geq 0.5\}$. We asses the approximative distribution under the null hypothesis by using $\beta_2(x) = \eta_2(x)$ and $\beta_1(x)$ as in Figure 5. In Figure 6 p-values gained from 1000 simulations are plotted versus quantiles of a uniform distribution. All lines are close to the diagonal in Figure 6 (a) indicating, that the χ^2 approximation works well. Figure 6 (b) shows quantiles for the alternative $\beta_2(x) = \eta_2(x) + I\{x \geq 0.5\}$. Considering a significance level of 0.1, for example, the test for no breakpoint in $x = 0.5$ rejects in about 90% of the cases for this true $\beta_2(x)$.

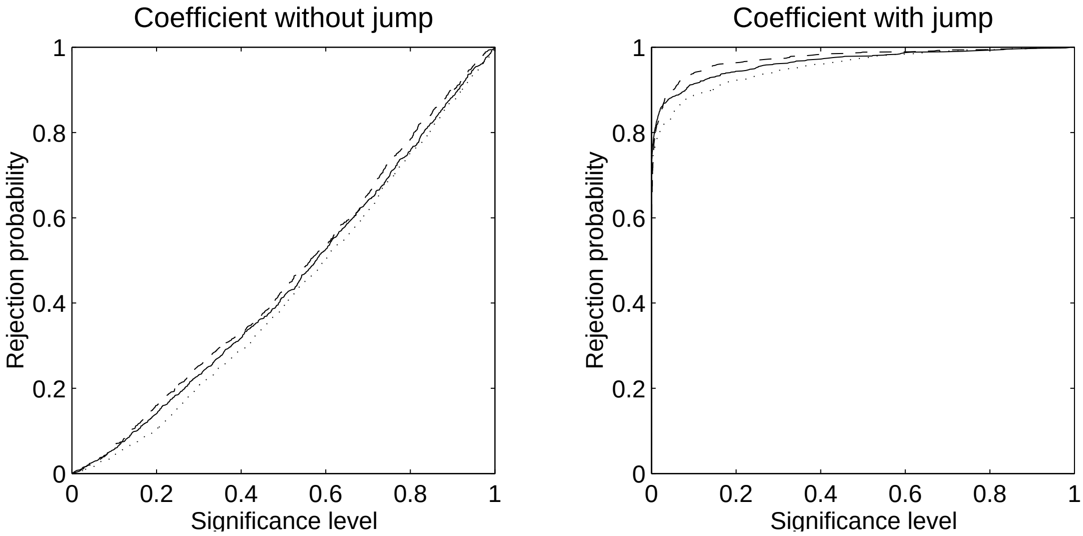

Figure 6: Rejection probabilities of the modified Wald test, using a χ^2 distribution. $\gamma^I = 20$ (dashed), $\gamma^I = 100$, (solid) $\gamma^I = 200$, (dotted)

6 An Algorithm for Generalized Soft–thresholding

In this section we propose an efficient algorithm which produces a sequence of estimates corresponding to a specified set of trade–off parameters λ. We start designing the algorithm by connecting generalized soft–thresholding to absolute penalized likelihood estimation as stated in Proposition 1. Following the proposal of Tishler and Zang (1982) in context of nonlinear L_1–norm estimation, we approximate the absolute penalty $\sum_{j=1}^{p}\sum_{k=1}^{n_j}|c_{jk}|$ in (4·4) by the continuously differentiable function

$$|c_{jk}| \approx g(c_{jk},\delta) = \begin{cases} -c_{jk}, & \text{if} \quad c_{jk} \leq -\delta \\ \frac{c_{jk}^2+\delta^2}{2\delta}, & \text{if} \quad -\delta < c_{jk} < \delta \\ c_{jk}, & \text{if} \quad c_{jk} \geq \delta \end{cases} . \tag{6·1}$$

For a moderate number of basis functions, this spline approximation allows a Newton type algorithm to compute an approximation c^* to $\hat{c}$ which maximizes

$$lo^*(c) = l(c) - \lambda \sum_{j=1}^{p}\sum_{k=1}^{n_j} g(c_{jk},\delta). \tag{6·2}$$

Working out the first derivatives of (6·2) yields a vector $s(c) - \lambda(d_1(c) + d_2(c))$ having components

$$\begin{aligned} \tfrac{\partial lo^*(c)}{\partial c_{jk}} &= s_{jk}(c) - \lambda\left(d_1(c_{jk}) + d_2(c_{jk})\right), \\ d_1(c_{jk}) &= \gamma_{jk}\mathrm{sgn}(c_{jk}) I\{|c_{jk}| \geq \delta\}, \\ d_2(c_{jk}) &= \tfrac{\gamma_{jk} c_{jk}}{\delta} I\{|c_{jk}| < \delta\}. \end{aligned}$$

The negative second derivative matrix of (6·2) is given by

$$\frac{\partial^2 lo^*(c)}{\partial c \partial c'} = H(c) + \lambda D(c, \delta), \qquad D(c, \delta) = \mathrm{diag}(\frac{\gamma_{jk}}{\delta} I\{|c_{jk}| < \delta\}).$$

In following modified Gauss–Newton or Fisher scoring procedure we replace the observed information $H(c)$ by its expectation $F(c)$ and simplify by $D(c, \delta)c = d_2(c, \delta)$.

Algorithm 1

1. Initialize the coefficient vector $c^{(m)}$, $m = 0$ and repeat
   (a) Compute the Fisher matrix $F(c^{(m)})$ and the score vector $s(c^{(m)})$ for the current coefficient vector $c^{(m)}$.
   (b) Solve the system
   $$[F(c^{(m)}) + D(c^{(m)}, \delta)]c^{(m+1)} = F(c^{(m)})c^{(m)} + s(\eta^{(m)}) - d_1(c^{(m)}, \delta)$$
   to obtain updated values $c^{(m+1)}$.
   (c) Trim steps crossing the zero:
   If $\{c_{jk}^{(m)} \neq 0$ and $\mathrm{sgn}(c_{jk}^{(m+1)}) \neq \mathrm{sgn}(c_{jk}^{(m)})\}$, set $c_{jk}^{(m+1)} = 0$.
2. Until the coefficients $c_{jk}^{(m)}$ do not change.

Trimming of coefficients in step 1(c) ensures that the coefficients c_{jk} do not alternate around $(-\delta, +\delta)$. For $c_{jk}^{(m)} \approx 0$ the quadratic approximation for $|c_{jk}| < \delta$ results in a rather small step length and therefore enables convergence to some $c_{jk}^* \in (-\delta, \delta)$. At convergence of Algorithm 1 we have

$$\begin{aligned} s_{jk}(c^*) = d_1(c_{jk}^*, \delta) = \lambda\gamma_{jk}\mathrm{sgn}(c_{jk}^*) \qquad & \text{if } |c_{jk}^*| \geq \delta \\ s_{jk}(c^*) = \lambda\gamma_{jk}c_{jk}^*/\delta < \lambda\gamma_{jk} \qquad & \text{if } |c_{jk}^*| < \delta \qquad (6{\cdot}3) \end{aligned}$$

and the conditions for the generalized soft–thresholding estimator (4·3) are fulfilled up to δ. From (6·3) we see, that the algorithm collects coefficients c^*_{jk}, having $|s_{jk}(c^*)| < \lambda\gamma_{jk}$ in the interval $(-\delta, \delta)$. The approximation c^* is improved and checked by removal of those coefficients. In the improved version, we set $\hat{c}_{jk} = 0$ when $|c^*_{jk}| < \delta$ and proceed with Newton type loops.

Algorithm 2 (Improved version)

1. Compute an approximation c^* to generalized soft–thresholding by Algorithm 1.

2. Let $\mathcal{M}$ be the set of basis functions φ_{jk}, defined by

$$\mathcal{M} = \{jk : |c^*_{jk}| > \delta\} \cup \mathcal{M}_0.$$

3. Compute improved estimates $\hat{c}$ by applying Algorithm 1 only to basis functions from $\mathcal{M}$. Use $\{c^*_{jk}, jk \in \mathcal{M}\}$ as initialization.

4. Check the results by verifying $|s_{jk}(\hat{c})| \leq \lambda\gamma_{jk}$ for all $jk \notin \mathcal{M}$.

Usually, Algorithm 2 adds only one extra iteration to Algorithm 1. If the check in Step 4 is passed, we have a generalized soft–thresholding estimator, satisfying the conditions (4·3) up to a prespecified termination criterion for the Newton–type iterations. Otherwise, when $|s_{jk}(\hat{c})| > \lambda\gamma_{jk}$ for some $jk \notin \mathcal{M}$, a slightly smaller value of δ helps to overcome this problem.

In varying–coefficient models the number of possible basis functions is often very large and direct use of Algorithm 2 becomes inefficient or even impossible due to linear dependencies. Based on the knowledge that the estimate consists of few non–zero coefficients only, we apply Algorithm 2 to an appropriate small fraction of basis functions. We start with a trade–off parameter $\lambda^{(0)}$ sufficiently big, so that the generalized soft–thresholding estimator $\hat{c}^{(0)}$ coincides with a maximum likelihood estimator for coefficients from the embedded model $\mathcal{M}_0$. Decreasing λ, we arrive at some $\lambda^{(1)} < \lambda^{(0)}$ where $|s_{jk}(\hat{c}^{(0)})| < \lambda^{(1)}\gamma_{jk}$ for some jk. Using the corresponding φ_{jk} together with the basis functions from the embedded model in Algorithm 2, we obtain a generalized soft–thresholding estimator $\hat{c}^{(1)}$

for $\lambda^{(1)}$. Continuing this principle leads to Algorithm 3, which computes the estimator for a sequence of threshold parameters $\lambda^{(0)} > \cdots > \lambda^{(l)} > \cdots > \lambda^{(L)}$. Since Algorithm 3 starts with estimation of the embedded model, we can easily incorporate the scaling of the threshold values as discussed in Subsection 4·2.

Algorithm 3:

1. Estimate the embedded model $c^{(0)}$ using only coefficients in $\mathcal{M}_0$ by maximizing the log–likelihood. Set $\mathcal{M} = \mathcal{M}_0$.
2. Select the threshold values γ_{jk} based on this estimate as described in Subsection 4·2 .
3. Do while $l \leq L$:
   (a) If $\exists jk \notin \mathcal{M} : |s_{jk}(\hat{c})| > \lambda^{(l)}\gamma_{jk}$ then add the index jk with $jk = \arg\max |s_{jk}(\hat{c})|/\gamma_{jk}$ to $\mathcal{M}$.
   (b) Compute current estimates $\hat{c}_{jk}$ by applying steps 1,2,3 of Algorithm 2 only to coefficients from $\mathcal{M}$.
   (c) Let $\mathcal{M} = \{jk : \hat{c}_{jk} \neq 0\} \cup \mathcal{M}_0$
   (d) If $|s_{jk}(\hat{c})| \leq \lambda^{(l)}\gamma_{jk}$ for all $jk \notin \mathcal{M}$:
   Keep the result $\hat{c}^{(l)} = \hat{c}$ as estimate for $\lambda^{(l)}$ and set $l = l + 1$.

Algorithm 3 adds successively basis coefficients to the set of non–zero coefficients. When the score function $s_{jk}(\hat{c})$ is smaller than the threshold value for all zero coefficients, we have an estimator for $\lambda^{(l)}$ and the algorithm proceeds with the next smaller $\lambda^{(l+1)} < \lambda^{(l)}$. Initializing Algorithm 1 in step 3 (b) with current estimates, only few Newton–type iterations are necessary. Due to the comparable small number of basis functions supplied in step 3 (b) to Algorithm 2, computation of the score vector in step 3 (d) is often the most expensive part of Algorithm 3. By employing efficient algorithms, specific for the set of basis functions used, computational cost is greatly reduced. For example, for wavelet basis functions computation of η is based on the inverse wavelet transform, whereas $s_{jk}(c)$ can be gained by the fast wavelet transform. In case of orthogonal splines it is sufficient to use only basis functions having up to a moderate number of sign changes.

Since Algorithm 3 produces a sequence of estimates for different values of λ it is particularly convenient for exploring the estimator as a function of the trade–off parameter λ.

For all computations shown in the paper we specified the approximation in Algorithm 1 by $\delta = 10^{-6}$ and used $\max|c_{jk}^{(m)} - \hat{c}_{jk}^{(m-1)}| < 10^{-6}$ as termination criterion for the Newton–type algorithm. Finally, we remark, that algorithms for generalized soft–thresholding can be based on most of the algorithms designed for nonlinear L_1-norm estimation. See Gonin and Money (1989), Ch. 2.3 for a survey of procedures, leading to alternatives for Algorithm 1.

7 Application to Unemployment Data

As an illustrative application of the proposed method, we investigate in the effect of gender on duration of unemployment periods. Our dataset consists of monthly unemployment periods from January 1983 through December 1992 recorded in the German socio–economic panel GSOEP (Hanefeld (1987)). Here we consider only spells starting with a transition from full–time employment to unemployed. An unemployment period ends, when the individual under study switches from unemployment to some different state such as part–time employment, housewife/husband, or to a full–time job.

To study the characteristics of unemployment we consider the terminations of each period as realizations of a stochastic process in calendar time t. We introduce an event indicator distinguishing between

$$y_i(t) = \begin{cases} 1, & \text{period } i \text{ ends with full–time employment at } t+1, \\ 2, & \text{period } i \text{ ends with anything but full–time employment at } t+1, \\ 0, & \text{period } i \text{ continues to } t+1, \end{cases}$$

and regard each process $y_i(t)$ as the outcome of a series of multinomial experiments. Thus, (conditional) probabilities $\pi_{ir}(t)$ of the disjunctive events $\{y_i(t) = r\}$, $r = 0, 1, 2$ are used to describe the dynamics of the labour market. A common choice of models relating those probabilities or time–discrete hazard functions to general event–specific predictors $\eta_r(X_i, t)$ is the multinomial logit model where

$$\begin{aligned}\pi_{ir}(t) &= h\{\eta_1(X_i, t), \ldots, \eta_m(X_i, t)\} \\ &= \frac{\exp\{\eta_r(X_i, t)\}}{1 + \sum_{q=1}^{m} \exp\{\eta_q(X_i, t)\}},\end{aligned} \qquad (7{\cdot}1)$$

see Allison (1982), Fahrmeir and Wagenpfeil (1996) and Fahrmeir and Knorr-Held (1997) for details. Furthermore, since censoring occurs, we also make use of a risk indicator

$$c_i(t) = \begin{cases} 1, & \text{period } i \text{ has been under study all the time until } t \\ 0, & \text{else,} \end{cases}$$

which masks unobserved transitions. Using this notation, the dataset is expressed by observed response variables $\tilde{y}_i(t) = c_i(t)y_i(t)$ and $c_i(t)$ together with a set of possibly time–varying covariates $x_{ij}(t) \in X_i(t)$. The model specification is completed by assuming multiplicative structure

$$P\{\tilde{y}_i(t) = r|\tilde{y}_i(t-1), X_i(s), c_i(s); s = 1, \ldots, t; i = 1, \ldots, n\} = c_i(t)\pi_{ir}(t),$$

for censoring mechanisms, as conventional in event history analysis, see e.g. Andersen, Borgan, Gill and Keiding (1993), Ch. 3.

Figure 7 gives a first summary of the data. The naive estimate of $\pi_{ir}(t)$, the ratio of the number of transitions to the number of individuals at risk, $\sum_i I\{y_i(t) = r\}/\sum_i c_i(t)$ for each subpopulation is plotted versus calendar time. Some periodicity of $\pi_{ir}(t)$ is evident from Figure 7 (a) and Figure 7 (d). Males seem to have lower propensity to leave unemployment to the "other" category than females have.

Also, the probability of leaving unemployment can be described by the unemployment duration time, $d = d_i(t)$. To take this into account we use a multinomial varying–coefficient model (7·1) with event–specific predictors

$$\eta_r(X_i(t), t) = \beta_{1r}(t) + \beta_{2r}(d) + \{\beta_{3r}(t) + \beta_{4r}(d)\} * gender, \qquad (7{\cdot}2)$$

where gender is 1 for females and -1 for males, respectively. In (7·2) the effect $\beta_{3r}(t)$ explores trends as well as seasonal aspects of female unemployment during the observation period, whereas $\beta_{4r}(d)$ distinguishes between long–term and

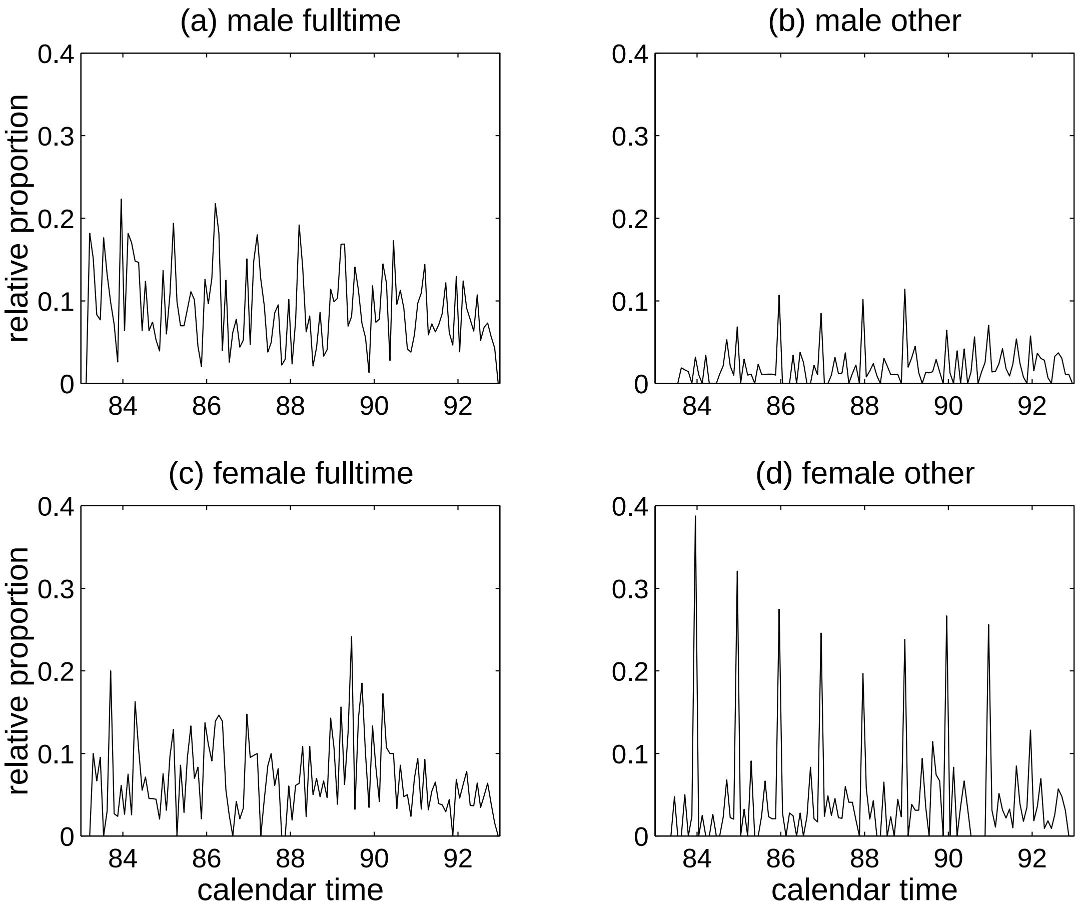

Figure 7: Relative proportions of observed transitions from the unemployment state. Proportions of (a) males and (c) females who found a full–time job to the number unemployed. Analogous proportions of (b) males and (d) females switched to the "other" category.

short–term unemployment with regard to gender. When interested mainly in duration effects, calendar time effects have to be incorporated to adjust for the current specific situation on the labour market. We avoid systematic bias due to omission of left–censored periods and include only 1781 periods terminating after January 1986. Furthermore, 30 periods lasting longer than 3 years are censored after 36 months.

Assuming the varying coefficients in (7·2) to be homogeneously smooth might cover important features. There are many reasons for possible abrupt changes in the propensity for re–employment, e.g. changes in labour legislation. Accounting for possible breakpoints, we decompose calendar time effects in

$$\beta_{jr}(t) = \beta_{jr}^{smo}(t) + \beta_{jr}^{jmp}(t) + \beta_{jr}^{per}(t).$$

Cubic Demmler–Reinsch splines are used for $\beta_{jr}^{smo}(t)$ and the set of indicator functions $\{\varphi^I(t)\}$ describes $\beta_{jr}^{jmp}(t)$. The periodical component, $\beta_{jr}^{per}(t)$ is based on trigonometric polynomials from the set

$$\{\cos(2\pi t/(12k)), \sin(2\pi t/(12k)), \quad k = 1, \ldots, 6\}$$

with period up to 12 months. Analogously, duration effects $\beta_{jr}(d)$ are decomposed into a smooth part, $\beta_{jr}^{smo}(d)$, and a part that modells jumps, $\beta_{jr}^{jmp}(d)$. Alltogether a catalogue of 984 basis functions are allowed to contribute to the predictor. The embedded model is set up by 12 parameters representing linear functions for calendar time and duration effects, respectively. We found that $\gamma^I = 100$ as threshold for indicator functions as well as for trigonometric functions provides a good trade–off between the smoothness, the jumps and the period.

Generalized soft–thresholding is carried out, starting Algorithm 3 with the threshold sequence $\lambda = 10^0, 10^{-0.0025}, \ldots, 10^{-2.5}$ by using grouped data. The output consists of 1001 estimates having between 14 and 211 coefficients contributing to the predictor $\hat{\eta}(\lambda)$. In Figure 8 (a) the deviance of the generalized soft–thresholding estimates $\hat{\eta}(\lambda)$ to the embedded model $D_0(\lambda) = -2\{l(\hat{\eta}^0) - l(\hat{\eta}(\lambda))\}$ is plotted as a function of the trade–off parameter λ. Figure 8 (b) displays the usual deviance $D(\lambda) = -2\{l(\hat{\eta}(\lambda)) - l(y)\}$ as the criterion for goodness of fit. We observe that goodness of fit increases monotonically with λ and is acceptable for

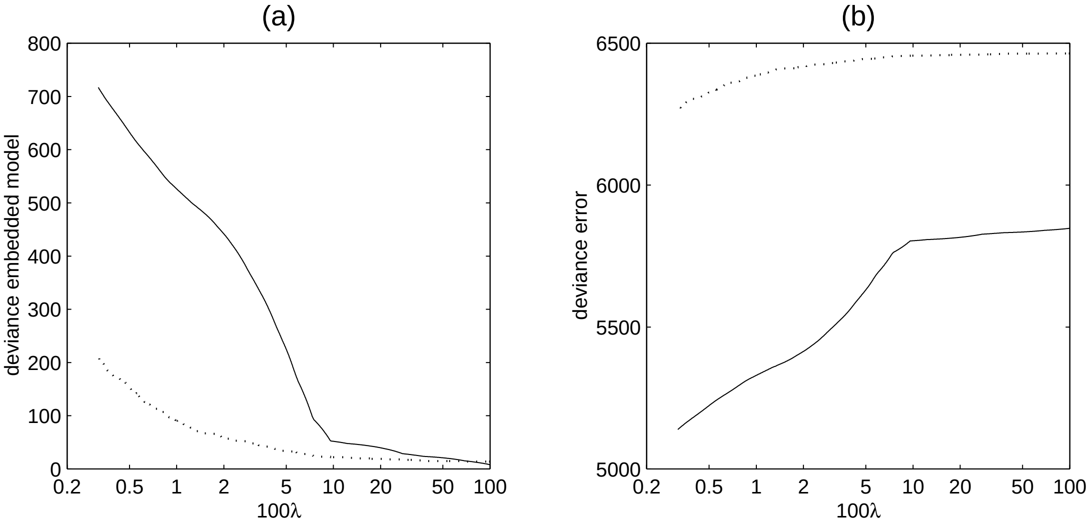

Figure 8: Deviance of generalized soft–thresholding estimates for unemployment data to embedded model (a) and deviance to the unrestricted log–likelihood (b). The dotted lines indicate the number of non–zero coefficients (a) in the model and the corresponding degrees of freedom for the error (b).

all estimates. Especially, thresholds with $\lambda < 0.02$ provide a good fit compared to the number of parameters involved.

Figures 9 and 10 plot estimates for multiple λ in the spirit of the family approach of Marron and Chung (1997). In models where several functions are estimated simultaneously, the family approach provides insight in how varying effects interact and the spread of the effects gives an idea of the precision of the estimate in specific regions. We show the results obtained by using a set of 11 trade–off parameters $\lambda \in [0.0049, 0.021]$. The corresponding estimates consist of 58 to 155 basis functions contributing to the predictor. For ease of interpretation all varying coefficients shown are centered to have mean zero.

To support the analysis, we computed test statistics according to the χ^2 approximation suggested in Section 5. We test for the hypotheses whether all included basis coefficients from a single set of basis functions are zero. This test is performed for each varying effect, separately. The hypothesis of those tests is based on estimated non–zero coefficients. Therefore, p-values obtained

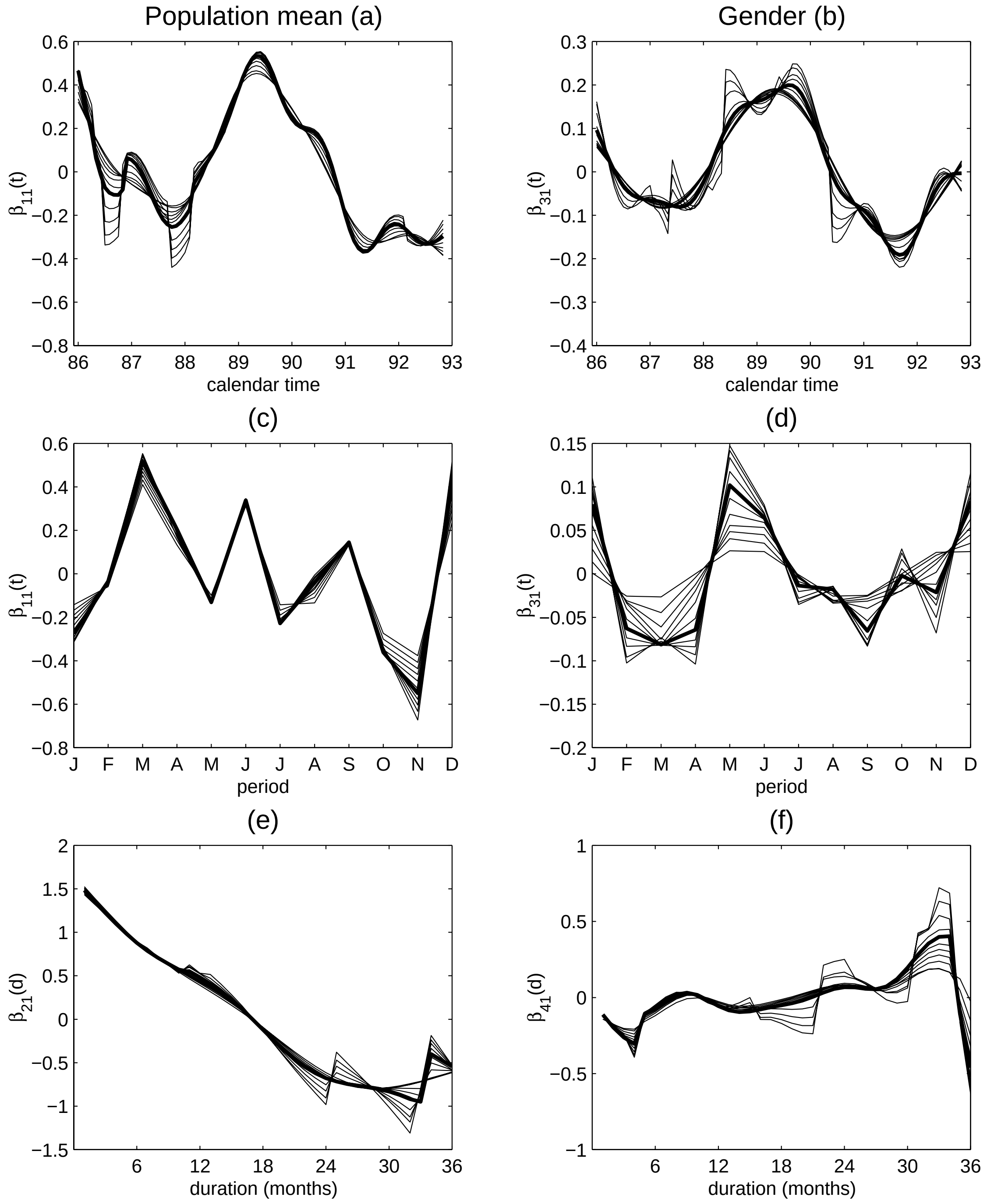

Figure 9: Centered effects modelling transitions to full–time employment. The thick line corresponds to $\lambda = 0.0089$. Remaining constants are -3.3283 for the population effect and -0.06219 for the effect of gender, respectively.

	Population mean			*Gender*		
Effect	*T*	*df*	*p–value*	*T*	*df*	*p–value*
Intercept	12.7689	1	0.000352	0.0285	1	0.866041
Calendar time						
Linear	0.0734	1	0.786510	0.6995	1	0.402945
Smooth	21.7923	8	0.005315	14.1839	5	0.014482
Jumps	17.1105	7	0.016697	—	0	—
Period	99.1432	8	0.000000	11.3446	5	0.044959
Duration						
Linear	17.7097	1	0.000026	1.4431	1	0.229634
Smooth	1.6791	3	0.641596	3.1638	5	0.674748
Jumps	2.3277	3	0.507238	7.5146	4	0.111066

Table 3: Tests on components of the varying coefficients modelling transitions to full–time employment.

are formally incorrect. However, we use them in an informal way to obtain an impression of the evidence of certain components in the model. In Tables 3 and 4 these tests are reported for the trade–off parameter $\lambda = 0.0089$ resulting in 99 non–zero coefficients. Corresponding estimates are plotted as thick line in Figures 9 and 10.

First, we discuss effects contributing to transitions for full–time employment. In Figure 9 (a) we observe a distinct maximum for getting re–employed in summer 1989 and two distinct lows at the beginning between July and October 1986 and between October 1987 and February 1988. The effect of gender over calendar time shown in Figure 9 (b) has a similar coarse structure. This supports the hypothesis, that in times of more pressure on the labour market it is even more difficult for female to find a full–time job. Referring to the p-value for smooth components of the calendar time effects in Table 3, this phenomenon is quite evident from the data. Clear periodicity of the probability of being re–employed is obvious from figure 9 (c), compare also Table 3. In Figure 9 (d) the periodic effect of gender is anticyclic during the first half of the year. This might be caused by fewer females working in the building trade, where often seasonal workers

	Population mean			*Gender*		
Effect	*T*	*df*	*p–value*	*T*	*df*	*p–value*
Intercept	52.8955	1	0.000000	3.9259	1	0.047548
Calendar time						
Linear	1.9528	1	0.162289	5.9979	1	0.014323
Smooth	15.5322	7	0.029751	4.3180	6	0.633728
Jump	9.2681	2	0.009715	5.0262	1	0.024967
Period	220.2585	8	0.000000	15.9773	6	0.013876
Duration						
Linear	0.0602	1	0.806226	3.1640	1	0.075279
Smooth	0.2051	2	0.902520	4.8205	4	0.306212
Jump	4.4297	2	0.109172	2.1756	1	0.140219

Table 4: Teststatistics for components of the varying coefficients modelling transitions to anything but full–time employment.

are employed. The population effect of duration on terminating unemployment is approximatively linear in Figure 9 (e). We conclude, that it becomes more difficult to find a full–time job the longer one is unemployed. Except for the linear component of the population effect, all other components modelling the effects of duration have rather high p-values in Table 3.

Considering the termination of unemployment for "other" reasons in Figure 10 (a), we observe a steep increase in the population effect during 1986. In contrast to transitions for full–time work, this effect stays at a high level after summer 1989 and shows an additional jump in June 1992. The effect of gender has a slightly negative trend over calendar time and shows a distinct jump in April 1989 just before East German labour participated in the panel; compare also the corresponding p-values in Table 4. The family plot in Figure 10 (b) displays a scattering effect in 1986 and during the second half of 1990. During those periods only very few transitions to other states were observed in total, compare Figure 7 (b) and (d). The sparseness of the data provides little information about the corresponding effect of gender, which causes scattering in the estimates. Clear periodicity is evident in Figure 10 (c) indicating that considered

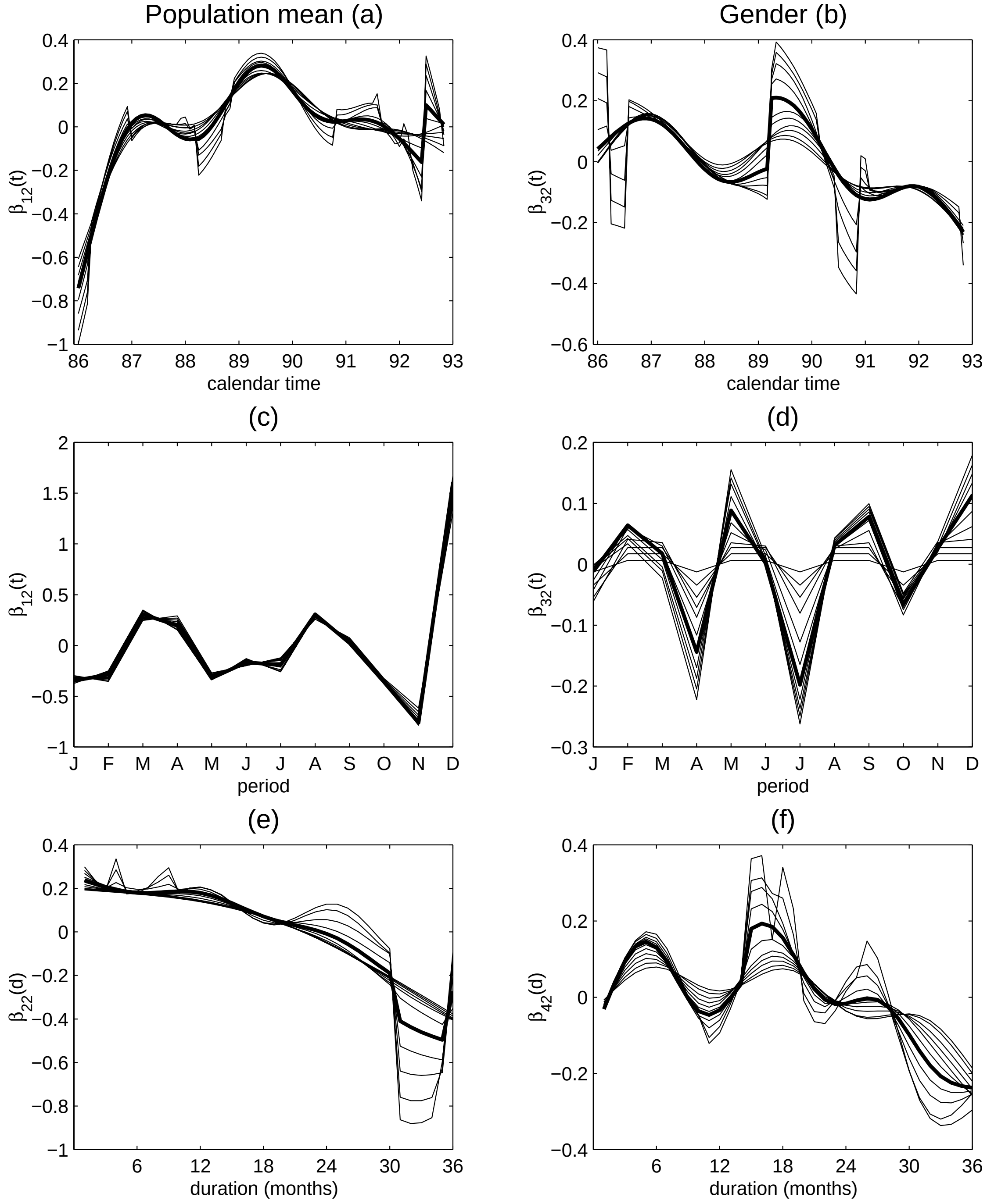

Figure 10: Centered effects modelling transitions to other terminations. The thick line corresponds to $\lambda = 0.0089$. Remaining constants are -3.6814 for the population effect and 0.2749 for the effect of gender, respectively.

periods terminate with high probability in December. A closer look at the data shows, that this effect is mainly influenced by transitions to housewife/husband. Therefore, it is even stronger for females (Figure 10 (d)). Again, p–values for duration effects indicate no evident influence of duration on transitions to other states.

To obtain further insight into the influence of duration, we assume duration effects to be smooth. We represent them by the first 8 orthogonal Demmler–Reinsch splines excluding the constant. The corresponding basis coefficients were estimated using Algorithm 3 together with the same threshold sequence as above. For each duration effect separately, a test on the hypothesis $c_{j1} = \cdots = c_{j8} = 0$ is performed. Resulting p-values are displayed in Figure 11 (a) as a function of the trade–off parameter λ. The p-value for the population effect, modelling transitions to full–time employment is not shown, since it is smaller 0.0001, regardless of λ. For thresholds bigger than 0.074 an effect of gender on transitions to a full–time job is significant up to 5%. With increasing model complexity this effect becomes less evident. In contrast, we observe that p-values for duration effects on other transitions are increasing with λ. Not accounting for nonlinearities over calendar time may cover a possible effect of duration here. In a follow–up paper we will stratify transitions to other states for a more refined investigation in the interaction of gender with duration.

Figure 11 (b) gives insight into correlations between the estimated basis coefficients. In the image plot we see the quantities of a correlation matrix computed from the inverse of the estimated Fisher matrix $F(\hat{c}_S)$ corresponding to the threshold $\lambda = 0.0089$. High correlations are visible particularly in the diagonal blocks. These intra–effect correlations correspond to basis functions contributing to the same varying effect. Apart from intra–effect correlations, we observe high correlations for coefficients contributing to the estimated duration effects $\hat{\beta}_{41}(d)$ and $\hat{\beta}_{42}(d)$. Both estimated effects consist of several basis coefficients correlated to corresponding calendar time effects. This helps explaining the variability of the p-values in Figure 11 (a). Let us pick out correlations between $\hat{\beta}_{41}(d)$ and $\hat{\beta}_{31}(t)$. The increase of $\hat{\beta}_{41}(d)$ is mainly caused by females unemployed for a long period who found full–time work in times of the good global job situation between

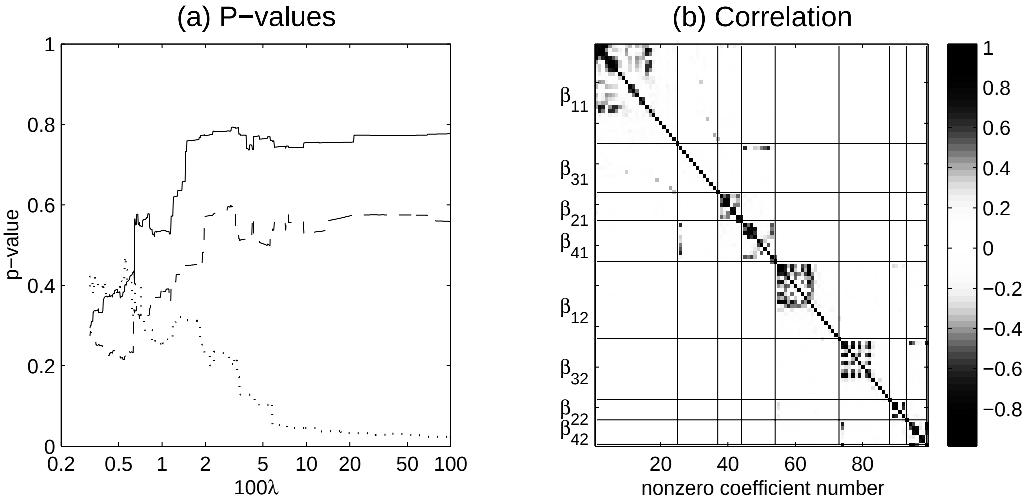

Figure 11: (a) P–values for a test on smooth effects of the duration of an unemployment period. Lines correspond to "gender – full–time" (dotted); "population – other" (dashed); "gender – other" (solid). (b) Estimated correlation between basis coefficients.

1988 and 1990. As consequence, we cannot reject the hypotheses of no interaction between duration and gender. A simpler parametric model, not accounting for more complex calendar time effects, would yield a different conclusion.

8 Discussion

We believe that generalized soft–thresholding should belong to every statisticians toolbox. It produces compact results which can be handled and analyzed similarly to more classical parametric approaches. The estimator can be customized simply by supplying different sets of basis functions. Estimated effects inherit their properties from the basis functions. The technique is more stable than approaches based on variable selection and the smoothing parameter acts continuously on the estimated coefficients.

Allthough we embedded generalized soft–thresholding in the framework of the varying–coefficient models, it can be used in connection with many other structures. Future research will include interactions between metrical covariates.

Those interactions can be modelled by bivariate functions, which are described using tensor product basis functions.

The selective property of generalized soft–thresholding also allows for simultaneous variable selection within the varying–coefficient model. Results, stated in Tibshirani (1996) for the lasso procedure, strongly suggest this extension.

Appendix

Proof of Proposition 1

The proof proceeds similarly to conditions for nonlinear L_1–norm estimation, compare e.g. Gonin and Money (1989), Ch. 2. Let $c_k^+ = \max(0, c_k)$, $c_k^- = \max(0, -c_k)$ and Z a design matrix consisting of the point evaluations $\varphi_k(x_i)$. Using $c_k = c_k^+ - c_k^-$ and $|c_k| = c_k^+ + c_k^-$, we rewrite the absolute penalized log–likelihood criterion $lo(c)$ from (3·9) as

$$lo(c^+ - c^-) = l(c^+ - c^-) - \lambda\gamma'(c^+ + c^-),$$

where $c^+ = (c_1^+, \ldots, c_n^+)$, $c^- = (c_1^-, \ldots, c_n^-)$ and $\gamma = (\gamma_1, \ldots, \gamma_n)$. Now, we can reformulate the maximum absolute penalized likelihood estimator as

$$\text{maximize} \quad lo(c^+ - c^-) \quad \text{subject to} \quad c_k^+ \geq 0, c_k^- \geq 0, \quad k = 1, \ldots, n. \tag{A·1}$$

and derive Kuhn–Tucker necessary conditions as stated e.g. in Gill, Murray and Wright (1981), Ch. 3. The concept is based on the active constraint sets $A^+ = \{k : c_k^+ = 0\}$, $A^- = \{k : c_k^- = 0\}$, their complements C^+ and C^-, respectively and the partial derivatives:

$$\begin{aligned} \partial lo(c^+ - c^-)/\partial c_k^+ &= s_k(c) - \lambda\gamma_k, \\ \partial lo(c^+ - c^-)/\partial c_k^- &= -s_k(c) - \lambda\gamma_k. \end{aligned}$$

According to Kuhn–Tucker, at a maximum $\hat{c}^+$, $\hat{c}^-$ of (A·1), the set $A = \{k : \hat{c}_k = 0\} = A^+ \cap A^-$ is formed by all k corresponding to non–positive partial derivatives, i.e. $s_k(\hat{c}) - \lambda\gamma_k \leq 0$ and $-s_k(\hat{c}) - \lambda\gamma_k \leq 0$, which results in $A = \{k : |s_k(\hat{c})| \leq \lambda\gamma_k\}$. The set of positive coefficients necessarily satisfies $s_k(\hat{c}) - \lambda\gamma_k = 0$ and $-s_k(\hat{c}) - \lambda\gamma_k \leq 0$ leading to $C^+ \cap A^- = \{k : s_k(\hat{c}) = \lambda\gamma_k\}$. For $c_k < 0$ we analogously arrive at $C^- \cap A^+ = \{k : s_k(\hat{c}) = -\lambda\gamma_k\}$.

To derive sufficient conditions, let e_k be a n dimensional unit vector and let E_{A^+} be a matrix with rows $e_k', k \in A^+$, E_{A^-} respectively. Then, the set of active constraints matches with the rows of

$$E_A = \begin{pmatrix} E_{A^+} & 0 \\ 0 & E_{A^-} \end{pmatrix},$$

and the projected Hessian on the null space of active constraints whose columns form a basis for the set of vectors orthogonal to the rows of E_A can be created from $B = \{e_k : k \in (C^+ \cap A^-) \cup (C^- \cap A^+)\}$. Since $(C^+ \cap A^-) \cap (C^- \cap A^+) = \emptyset$, we can use the actual coefficients c_k instead of c_k^+, c_k^- and form a matrix E_B having column vectors in B. The sufficient condition known from constrained optimization, requires $-E_B'H(c)E_B$ to be positive definite, where $H(c) = \partial ls(c)/\left(\partial c \partial c'\right)$. By $E_B'H(c)E_B = E_B'Z'H(\eta)ZE_B$ and $ZE_B = Z_B$, this is equivalent to the condition formulated in Proposition 1.

Proof of Proposition 2

To simplify notation in the proof, we skip the index 1 and S in s_1, H_1, c_S and refer to as s, H and $\hat{c}$. First we derive the maximizer of the quadratic approximation

$$Q(c) = l(\hat{c}) + s(\hat{c})'(c - \hat{c}) + \frac{1}{2}(c - \hat{c})'H(\hat{c})(c - \hat{c}) \tag{A·2}$$

without restriction ($\hat{c}_1$) and under the restriction $Ac = 0$ denoted by $\hat{c}_0$. Equating

$$\partial Q(c)/\partial c = s(\hat{c}) - H(\hat{c})c + H(\hat{c})\hat{c}$$

to zero yields

$$\hat{c}_1 = \hat{c} + H(\hat{c})^{-1}s(\hat{c}).$$

Introducing Langrangian multipliers as $Q_0(c) = Q(c) - \lambda' Ac$ results in

$$\partial Q_0(c)/\partial c = s(\hat{c}) - H(\hat{c})c + H(\hat{c})\hat{c} - A'\lambda$$

and we obtain

$$\hat{c}_0 = \hat{c} + H(\hat{c})^{-1}[s(\hat{c}) - A'\lambda] \tag{A·3}$$

as maximizer of (A·2) under the restriction $Ac = 0$. Multiplying both sides in (A·3) by A yields

$$A'\lambda = G(\hat{c})[\hat{c} + H(\hat{c})^{-1}s(\hat{c})],$$

where

$$G(\hat{c}) = [AH(\hat{c})^{-1}A']^{-1}$$

and the maximizer is obtained as

$$\begin{aligned}\hat{c}_0 &= \hat{c} + H(\hat{c})^{-1}s(\hat{c}) - H(\hat{c})^{-1}G(\hat{c})[\hat{c} + H(\hat{c})^{-1}s(\hat{c})] \\ &= \hat{c}_1 - H(\hat{c})^{-1}G(\hat{c})\hat{c}_1.\end{aligned}$$

Straight forward calculus shows, that

$$\begin{aligned}Q(\hat{c}_1) &= l(\hat{c}) + \frac{1}{2}s(\hat{c})'H(\hat{c})^{-1}s(\hat{c}) \\ Q(\hat{c}_0) &= l(\hat{c}) + \frac{1}{2}s(\hat{c})'H(\hat{c})^{-1}s(\hat{c}) - \frac{1}{2}\hat{c}_1'G(\hat{c})'H(\hat{c})^{-1}G(\hat{c})\hat{c}_1\end{aligned}$$

and therefore,

$$\begin{aligned}2[Q(\hat{c}_1) - Q(\hat{c}_0)] &= \hat{c}_1'G(\hat{c})'H(\hat{c})^{-1}G(\hat{c})\hat{c}_1 \\ &= \hat{c}_1'A'[AH(\hat{c})^{-1}A']^{-1}A\hat{c}_1\end{aligned}$$

by definition of $G(\hat{c})$.

References

Allison, P. D. (1982). Discrete–time methods for the analysis of event histories, *in* S. Leinhardt (ed.), *Sociological Methodology*, Jossey–Bass, San Francisco, pp. 61–89.

Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer, New York.

Bruce, A. G. and Gao, H. (1996). Understanding WaveShrink: Variance and bias estimation, *Biometrika* **83**, 727–745.

Chen, S. and Donoho, D. (1995). Basis Pursuit, *Technical report*, Department of Statistics, Stanford University.

Daubechies, I. (1992). Ten lectures on wavelets, *CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 61, SIAM, Philadelphia.

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

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaption by wavelet shrinkage, *Biometrika* **81**, 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia (with discussion)?, *J.R. Statist. Soc. B* **57**, 301–369.

Eubank, R. (1988). *Spline smoothing and nonparametric regression*, Marcel Dekker, New York.

Fahrmeir, L. and Kaufmann, H. (1991). On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression, *Metrika* **38**, 37–60.

Fahrmeir, L. and Knorr-Held, L. (1997). Dynamic discrete–time duration models, *in* A. Raftery (ed.), *Sociological Methodology*, Jossey–Bass, San Francisco, pp. ???–???

Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*, Springer, New York.

Fahrmeir, L. and Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risk models, *J. Amer. Statist. Assoc.* **91**, 1584–1594.

FAN, J., HECKMAN, N. E. AND WAND, N. P. (1995). Local polynomial kernel regression for generalized linear models and quasi–likelihood functions, *J. Amer. Statist. Assoc.* **90**, 141–150.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion), *Ann. Statist.* **19**, 1–141.

FRIEDMAN, J. H. AND SILVERMAN, B. W. (1989). Flexible parsimounious smoothing and additive modelling (with discussion), *Technometrics* **31**, 3–39.

GILL, P. E., MURRAY, W. AND WRIGHT, M. H. (1981). *Practical Optimization*, Academic Press, San Diego.

GONIN, R. AND MONEY, A. (1989). *Nonlinear L_p–norm estimation*, Marcel Dekker, New York.

GREEN, P. AND SILVERMAN, B. (1994). *Nonparametric regression and generalized linear models*, Chapman and Hall.

HANEFELD, U. (1987). *Das sozio–ökonomische Panel*, Campus, Frankfurt.

HASTIE, T. (1996). Pseudo Splines, *J.R. Statist. Soc. B* **58**, 379–396.

HASTIE, T. AND TIBSHIRANI, R. (1990). *Generalized additive models*, Chapman and Hall, London.

HASTIE, T. AND TIBSHIRANI, R. (1993). Varying-coefficient Models (with discussion), *J.R. Statist. Soc. B* **55**, 757–796.

JOHNSTONE, I. M. AND SILVERMAN, B. W. (1997). Wavelet threshold estimators for data with correlated noise, *J.R. Statist. Soc. B* **59**, 319–352.

MAMMEN, E. AND VAN DE GEER, S. (1997). Locally adaptive regression splines, *Ann. Statist.* **25**, ???–???

MARRON, J. S., ADAK, S., JOHNSTONE, I. M., NEUMANN, M. H. AND PATIL, P. (1995). Exact risk analysis of wavelet regression, *Statistics research report*, Australian National University.

MARRON, J. S. AND CHUNG, S. S. (1997). Presentation of smoothers: The family approach, *Preprint*, University of North Carolina, Chapel Hill.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized linear models 2nd ed.*, Chapman and Hall.

NASON, G. P. AND SILVERMAN, B. W. (1994). The discrete wavelet transform in S, *J. of Comp. and Graph. Statist* **3**, 163–191.

O'Sullivan, R., Yandell, B. S. and Raynor, W. J. (1986). Automatic smoothing of regression functions in generalized linear models, *J. A. Statist. Assoc.* **81**, 96–103.

Speckman, P. (1988). Kernel smoothing in partial linear models, *J.R. Statist. Soc. B* **50**, 413–436.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood–based models, *J. Amer. Statist Assoc.* **84**, 276–283.

Stone, C. J., Hansen, M., Kooperberg, C. and Troung, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling, *Ann. Statist.* **25**, ???–???

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J.R. Statist. Soc. B* **58**, 267–288.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation, *J. Amer. Statist. Assoc.* **82**, 559–568.

Tishler, A. and Zang, I. (1982). An absolute deviations curve–fitting algorithm for nonlinear models, *in* S. Zanakis and J. Rustagi (eds), *Optimization in Statistics, TIMS Studies in Managenment Science*, Vol. 19, North Holland, Amsterdam.

Tutz, G. and Kauermann, G. (1997). Local estimators in multivariate generalized linear models with varying coefficients, *Comput. Statist.* **12**, 193–207.

Wahba, G. (1990). Spline Models for Observational Data, *CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 59, SIAM, Philadelphia.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy, *Ann. Statist.* **23**, 1865–1896.