Klinger:

# Inference in High Dimensional Generalized Linear Models based on Soft-Thresholding

Sonderforschungsbereich 386, Paper 165 (1999)

Online unter: http://epub.ub.uni-muenchen.de/

# Inference in High Dimensional Generalized Linear Models based on Soft–Thresholding

ARTUR KLINGER

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

email: artur@stat.uni-muenchen.de

## Summary

We propose a new method for estimation of a high number of coefficients within the generalized linear model framework. The estimator leads to an adaptive selection of model terms without substantial variance inflation. Our proposal extends the soft–thresholding strategy from Donoho and Johnstone (1994) to generalized linear models and multiple predictor variables. Furthermore, we develop an estimator for the covariance matrix of the estimated coefficients, which can even be used for terms dropped from the model. Used in connection with basis functions, the proposed methodology provides an alternative to other generalized function estimators. It leads to an adaptive economical description of the results in terms of basis functions. Specifically, it is shown how adaptive regression splines and qualitative restrictions can be incorporated. Our approach is demonstrated by applications to solvency prognosis and rental guides.

*Keywords:* Generalized additive models; Generalized linear models; Penalized likelihood; Shrinkage; Splines; Thresholding;

# 1 Introduction

Let $\eta$ denote the linear predictor of a generalized linear model(GLM) (McCullagh and Nelder, 1989; Fahrmeir and Tutz, 1994), which is linked by

$$\mathrm{E}(y_i) = \mu_i, \qquad \eta_i = h^{-1}(\mu_i), \qquad , i = 1, \cdots, n$$

to a response variable $y_i$ with known distribution function. This paper focuses on GLM, having many terms in the linear predictor,

$$\eta_i = z_{i1}\beta_1 + \cdots + z_{ip}\beta_p, \qquad i = 1, \cdots, n \tag{1}$$

where $z_{ij}$, $j = 1, \cdots, p$ are possibly transformed explanatory variables. We introduce the term HDGLM (High dimensional generalized linear model) for those models incorporating many parameters compared to the sample size $n$. HDGLM occur in many observational studies, where explanatory variables have to be included to account for effects not controlled by the experimental design. Another application of HDGLM are situations, where no given functional form for the influence of a metrical covariate $x_j$ can be assumed in advance. In this setting, the $z_{ij}$ are defined by point evaluations of appropriate basis–functions $\psi_j(x_{ij})$, (e.g. spline–functions) leading to functional terms of the form

$$f_j(x_j) = \sum_{k=1}^{n_j} \psi_{jk}(x_j)\beta_{jk}$$

in (1) Used in connection with basis functions, HDGLM are a powerful alternative to nonparametric smoothing procedures as used in the generalized additive model framework of Hastie and Tibshirani (1990), where a predictor of the form

$$\eta = f_1(x_1) + \cdots f_p(x_p).$$

is assumed.

It is already known since Stein (1956), that maximum likelihood and least squares principles are unsuitable for handling many parameters. Putting much emphasis on bias, these estimators lead to unreasonably high variance of the estimates. Classical remedies for high dimensional problems can be classified into two groups.

The first group consists of approximately linear estimators, including the ridge–estimator (Nyquist, 1990; Marx et al., 1992; Segerstedt, 1992) and many of the smoothing principles (O'Sullivan et al., 1986; Hastie and Tibshirani, 1990; Staniswalis, 1989; Fan et al., 1995). Generally, estimators of this class reduce variance globally by introducing some bias. The amount of bias depends on the true parameter vector $\beta = (\beta_1, \ldots, \beta_p)'$ and is not limited in $\mathbb{R}^p$. Reasonable bias can only be guaranteed by incorporating prior knowledge about the magnitude of the parameters or about the smoothness of each predictor function. While approximately linear estimators are theoretically well understood, they are not able to adapt onto distinct parametrization or on a smoothness class, compare Donoho and Johnstone (1995). Therefore, additional nonlinear decision rules have to be incorporated, in practice.

The second group of estimators comprises numerous variable selection strategies. These nonlinear estimators lead to an explicit reduction of dimensionality. They are adaptively selecting model terms and, by using appropriate basis functions, they are also able to detect single jumps or breakpoints in the predictor functions. As a drawback, these nonlinear strategies lead to considerable variance inflation and, in conjunction with the maximum likelihood principle, to selection bias.

In this paper we suggest a compromise between these two classical remedies, the generalized soft–threshold (GSoft) estimator. The naming is due to the soft–threshold strategy, introduced by Donoho and Johnstone (1994) in the case of normally distributed errors and orthogonal covariate design. GSoft is closely related to the LASSO of Tibshirani (1996), but is further developed in several aspects.

As the ridge estimator maximizes the loglikelihood in an elliptical region, GSoft can be regarded as a maximizer of the loglikelihood in an angular region. This alternative restriction allows for adaptive selection of basis–functions and model terms without substantial variance inflation. GSoft leads to a parsimonious decomposition of the predictor, which is often demanded in practice but in contrast to common variable selection techniques, GSoft is a smooth procedure without selection bias. Even terms dropped from the model can be judged by their variances or derived test statistics. In connection with spline basis-functions or wavelets, GSoft allows for locally adaptive smoothing within the generalized linear model framework. It leads to a parsimonious representation of the predictor functions in terms of basis functions, which can be further studied by analyzing the corresponding covariance matrix. In addition, qualitative information about non–negative coefficients or the monotonicity of predictor functions can simply be incorporated.

### 1·1 *Outline*

The estimator is defined in section 2 and it is shown how inequality constraints can be incorporated into the procedure. Section 3 deals with problems arising from linear transformation of the predictor variables. It is shown how GSoft can be made invariant to linear transformations of the covariates. Section 4 gives approximations to the variance and the bias of the estimate. These approximations are further used to develop an estimator for the corresponding covariance matrix. In section 5, we demonstrate the small sample properties of the estimator in a simulation study. The extension to nonparametric function estimation in generalized linear models is made in section 6, where GSoft is connected to spline smoothing. Applications of the methodology to prognosis of solvency and rental guides are given in section 7.

## 2 Generalized Soft–Thresholding

Given a sequence of thresholds $\gamma_1, \cdots, \gamma_p \geq 0$ and a global threshold $\lambda > 0$, GSoft (Generalized Soft–Thresholding) is defined as a maximizer of the penalized likelihood criterion

$$lp(y;\beta,\lambda) = l(y;\beta) - \lambda \sum_{j=1}^{p} \gamma_j |\beta_j|, \quad \gamma_j \geq 0, \qquad (2)$$

where $l(y;\beta)$ is the loglikelihood function, given the data. Following Tibshirani (1996), the estimator can be interpreted as a constrained maximum likelihood estimator, or in a Bayesian context, as a maximum a posteriori estimator. The additional thresholds $\gamma_j$ account for the scaling of each covariate and can also be used to incorporate prior knowledge about the relevance of each term. The next theorem characterizes the estimator by estimating equations:

**Theorem:** Let $s_j(y;\beta) = \partial l(y;\beta)/\partial\beta_j$ denote the score–function with respect to $\beta_j$ and $H(\eta) = -\partial l(y;\eta)/(\partial\eta\partial\eta')$ the negative Hessian with respect to the linear predictor $\eta$. Conditions (3) and (4) give a necessary and sufficient condition for a local maximizer $\hat{\beta}$ of $lp(y;\beta,\lambda)$.

$$\begin{array}{rcrlcl} |s_j(y;\hat{\beta})| & \leq & \lambda\gamma_j & \text{if } \hat{\beta}_j & = & 0 \\ s_j(y;\hat{\beta}) & = & \lambda\gamma_j & \text{if } \hat{\beta}_j & > & 0 \\ s_j(y;\hat{\beta}) & = & -\lambda\gamma_j & \text{if}\hat{\beta}_j & < & 0 \end{array} \tag{3}$$

$$Z_\lambda' H(\hat{\eta}) Z_\lambda \text{ is positive definite,} \tag{4}$$

where $Z_\lambda$ a design matrix, consisting of all $z_j$ satisfying $|s_j(y;\hat{\beta})| = \lambda\gamma_j$. (Proof in appendix)

From the implicit definition (3) we see that the estimator has point mass on $\hat{\beta}_j = 0$ for $\gamma_j > 0$. Depending on $s_j(y;\hat{\beta})$, it adaptively selects terms from the model. In the case of a normal error distribution and orthogonal design, it reduces to the explicitly defined soft–threshold estimator

$$\hat{\beta}_j = \text{sgn}(z_j'y)(|z_j'y| - \lambda\gamma_j)_+$$

of Donoho and Johnstone (1994). When using the threshold $\gamma_j = 0$, the equations (3) reduce to the common maximum likelihood score equations $s_j(y;\beta) = 0$.

—- Figure 1 about here —-

Fig. 1 (a) illustrates GSoft in a logit–model. It corresponds to the abscissa of the intersection between the score–function $s_j(y,\beta)$, $y = 0,\ldots,20$ and the step function $\lambda\gamma_j\text{sgn}(\beta_j)$ in (a). Fig 1 (b) shows GSoft as a function of the maximum likelihood estimator, which doesn't converge for $y = 0$ and $y = 20$.

### 2·1 *Existence and uniqueness*

Since GSoft can be regarded as a penalized likelihood estimator incorporating a convex penalty, it exists even in situations, where the maximum likelihood principle diverges, compare Fig. 1 (a). However, in special cases of collinearity, solutions to the estimating equations (3) are not unique.

Consider the example

$$\eta = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3, \quad z_3 = \frac{z_1 + z_2}{2}, \tag{5}$$

where the linear predictor alternatively can be written as

$$\eta = \left(\beta_1 + \frac{1}{2}\beta_3\right) z_1 + \left(\beta_2 + \frac{1}{2}\beta_3\right) z_2. \tag{6}$$

For $\mathrm{sgn}(\beta_1) = \mathrm{sgn}(\beta_2) = \mathrm{sgn}(\beta_3)$ we have

$$|\beta_1| + |\beta_2| + |\beta_3| = |\beta_1 + \frac{1}{2}\beta_3| + |\beta_2 + \frac{1}{2}\beta_3|,$$

and both $\eta$ and the penalty are equivalent for the models (5) and (6). Therefore, no unique GSoft estimation exists in this example. However, using slightly different thresholds $\gamma_j$ would lead to a unique estimator. Since $Z_\lambda$ in (3) depends on the observations $y$ and on the thresholds $\gamma_j$, the uniqueness condition (4) is difficult to verify in practice.

### 2·2 *Nonnegativity constraints*

By a simple modification, GSoft can be extended to incorporate nonnegativity or nonpositivity constraints of the form

$$\beta_j \geq 0 \qquad \text{or} \qquad \beta_j \leq 0. \tag{7}$$

Let $(\beta_j)_+ = \max(\beta_j, 0)$ denote the nonnegative part and $(\beta_j)_- = \max(-\beta_j, 0)$ denote the nonpositive part of a coefficient $\beta_j$. Then we have

$$\beta_j = (\beta_j)_+ - (\beta_j)_- \qquad \text{and} \qquad |\beta_j| = (\beta_j)_+ + (\beta_j)_-, \tag{8}$$

and by

$$\begin{aligned} lp(y; \beta, \lambda) &= l\{y; (\beta_j)_+ - (\beta_j)_-\} - \lambda \sum_{j=1}^{p} \gamma_j \{(\beta_j)_+ + (\beta_j)_-\}, \\ & (\beta_j)_+ \geq 0, \quad (\beta_j)_- \geq 0, \end{aligned}$$

GSoft can alternatively be regarded as a constrained penalized likelihood estimator for the positive and the negative part of $\beta$. Different penalization of $(\beta_j)_+$ and $(\beta_j)_-$ yields the restrictions (7). We extend the target function of GSoft to

$$\begin{aligned} lp(y; \beta, \lambda) &= l\{y, (\beta_j)_+ - (\beta_j)_-\} - \lambda \sum_{j=1}^{p} \{\gamma_j^+ (\beta_j)_+ + \gamma_j^- (\beta_j)_-\} \\ & (\beta_j)_+ \geq 0, \quad (\beta_j)_- \geq 0, \end{aligned} \tag{9}$$

by introducing separate thresholds $\gamma_j^+$ and $\gamma_j^-$ for $(\beta_j)_+$ and $(\beta_j)_-$, respectively. This leads to the estimating equations

$$\begin{array}{rcccccl} -\lambda\gamma_j^- & \leq & s_j(y; \hat\beta) & \leq & \lambda\gamma_j^+ & \text{if } \hat\beta_j & = & 0 \\ & & s_j(y; \hat\beta) & = & \lambda\gamma_j^+ & \text{if } \hat\beta_j & > & 0 \\ -\lambda\gamma_j^- & = & s_j(y; \hat\beta) & & & \text{if } \hat\beta_j & < & 0 \end{array} \tag{10}$$

$$Z_\lambda' H(\hat\eta) Z_\lambda \text{ is positive definite.} \tag{11}$$

for a local maximum of (9). The result follows immediately from the proof in the appendix.

— Figure 2 about here —

In the following, we focus on the nonnegativity constraint $\hat{\beta}_j \geq 0$, which is in complete analogy to $\hat{\beta}_j \leq 0$. The third equation in (10) can be neglected, if $\gamma_j^- = C$ with $C$ sufficiently large, such that $-\lambda C < s_j(y;\hat{\beta})$ for $\hat{\beta}_j = 0$, $j = 1,\ldots,p$. Then the first or the second condition of (10) are always fulfilled and we have $\hat{\beta}_j \geq 0$. For $C \to \infty$, the estimating equations for $\hat{\beta}_j$ result in

$$\begin{array}{rcll} s_j(y;\hat{\beta}) & \leq & \lambda\gamma_j^+ & \text{if } \hat{\beta}_j = 0 \\ s_j(y;\hat{\beta}) & = & \lambda\gamma_j^+ & \text{if } \hat{\beta}_j > 0, \end{array}$$

as illustrated in fig. 2. There, GSoft under the constraint $\beta_j \geq 0$ corresponds to the intersection of $s_j(y;\beta)$ with the angular in Fig. 2 (a). By moving this angular, general constraints of the form $\beta_j \geq c$ can be handled, as well. In the limit $\lambda \to 0$, this approach leads to the constrained maximum likelihood estimator as proposed by McDonald and Diamond (1990).

## 3 Scaling

As with other biased techniques for high dimensional design, GSoft is not invariant to linear transformations of covariates $z_j$. This means, that an estimate $\tilde{\beta}_j$, which corresponds to a covariate $\tilde{z}_j = \alpha z_j$ does not equal $\hat{\beta}_j/\alpha$. Therefore, the estimated predictor depends on the scaling of the covariates. To overcome this problem in Ridge estimation and in the LASSO of Tibshirani (1996), it is required that the design has to be standardized in advance by

$$m(z_j) = \frac{1}{n}\sum_{i=1}^{n} z_{ji} = 0, \qquad \tilde{s}^2(z_j) = \frac{1}{n}\sum_{i=1}^{n}(z_{ji} - m(z_j))^2 = 1. \tag{12}$$

In contrast, GSoft implicitly accounts for the scaling of covariates by adjusting the thresholds $\gamma_j$ appropriately. No modification has to be made to the design matrix and model terms can be forced to enter by using a threshold $\gamma_j = 0$.

### 3·1 *Adjusting the thresholds*

In the following, we assume that an intercept $z_0 = 1$ is included. With the choice $\gamma_0 = 0$ the intercept $\beta_0$ is not penalized and $m(z_j)$ from (12) enters in $\hat{\beta}_0$. Hence, no centering is necessary when $\gamma_0 = 0$. Using the thresholds

$$\gamma_j = \sqrt{\tilde{s}^2(z_j)}, \tag{13}$$

any scaled version $\tilde{z}_j = \alpha z_j$ of $z_j$ yields a threshold $\tilde{\gamma}_j = \{\tilde{s}^2(\tilde{z}_j)\}^{1/2} = |\alpha|\gamma_j$. Inserting $\tilde{\gamma}_j$ into the penalty we have $\tilde{\gamma}_j|\tilde{\beta}_j| = \gamma_j|\beta_j|$ when $\tilde{z}_j\tilde{\beta}_j = z_j\beta_j$. Therefore, the penalty becomes invariant to a scaling factor $\alpha \neq 0$. The special choice $\alpha = 1/\sqrt{\tilde{s}^2(z_j)}$ results in the standardized design (12) and $\tilde{\gamma}_j = 1$. Adjusting of the thresholds $\gamma_j$ by (13) is therefore equivalent to standardization.

### 3·2 *The embedded model*

Alternatively, the thresholds can be adjusted on the probability for selection of a term $\hat{\beta}_j \neq 0$ given $\beta_j = 0$. We start by choosing a set of covariates that have to enter the model, denoted by the subdesign matrix $Z_0$. This model is termed the embedded model $\mathcal{M}_0$, having thresholds $\gamma_j = 0$ for each $j \in \mathcal{M}_0$. For the embedded model, GSoft is equivalent to maximum likelihood estimation of corresponding coefficients $\beta_{(0)}$.

Let the embedded model serve as a true data generating model. Then

$$\{F(\hat{\beta}_{(0)})^{-1}\}_{jj}^{\frac{1}{2}} s_j(y; \hat{\beta}_{(0)}) \xrightarrow{d.} N(0,1), \tag{14}$$

with $\{F^{-1}(\hat{\beta}_{(0)})\}_{jj}$ the last element from the diagonal of the inverse Fisher matrix

$$\{(Z_0, z_j)' F(\hat{\eta}_{(0)})(Z_0, z_j)\}^{-1}, \qquad F(\eta) = -\mathrm{E}\left(\frac{\partial l(y;\eta)}{\partial\eta\partial\eta'}\right)$$

evaluated at the maximum likelihood estimate of the embedded model. Expression (14) is equivalent to a score test on the hypothesis $\beta_j = 0$, which is known to be invariant to any linear transformation of $z_j$, compare e.g. Cox and Hinkley (1974, p. 339).

Using the thresholds

$$\gamma_j = \{F(\hat{\beta}_{(0)})^{-1}\}_{jj}^{-\frac{1}{2}}, \tag{15}$$

the first estimating equation of GSoft (3) has the form

$$|s_j(y; \hat{\beta})| \leq \lambda \{F(\hat{\beta}_{(0)})^{-1}\}_{jj}^{-\frac{1}{2}} \quad \text{for } \hat{\beta} = 0 \tag{16}$$

and is a score test, obeying $\hat{\beta} = \hat{\beta}_{(0)}$. Since generally, the Fisher information is defined as the covariance of the score vector, this strategy adjusts $\gamma_j$ on an estimation of the score function's standard deviation. In the case when $Z_0$ consists of the intercept term only, strategy (15) reduces to the simple standardization by (13).

The form of the score test in (14) is based on the assumption, that the embedded model is correctly specified. Since $Z_0$ usually consists of only a few important terms, we have to account for possible misspecifications and $F_{jj}(\hat{\beta}_{(0)})$ is no more a consistent estimator for the variance of $s_j(y, \hat{\beta}_{(0)})$.

Fahrmeir (1990) discusses different kinds of misspecification in GLM and proposes the estimator $\hat{F}(\hat{\eta}) = R(\hat{\eta})$ with

$$R(\hat{\eta}) = \operatorname{diag}\left\{s_1(y, \hat{\beta})s_1(y, \hat{\beta})', \ldots, s_n(y, \hat{\beta})s_n(y, \hat{\beta})'\right\} \tag{17}$$

as a robust alternative to $\hat{F}(\hat{\eta}) = F(\hat{\eta})$. Therefore, the choice

$$\gamma_j = \{R(\hat{\beta}_{(0)})^{-1}\}_{jj}^{-\frac{1}{2}} \tag{18}$$

yields a more robust adjustment of $\gamma_j$. In numerous simulation studies, the strategy based on (18) has been proven to be superior to (13) and (15).

## 4 Approximation of Variance and Bias

In this section, we give an approach for statistical inference based on linear expansions to GSoft. These approximations are used to construct an estimator for GSoft's covariance matrix. In Subsection 4·1, the total estimation error is decomposed into a deterministic and a stochastic part. The covariance matrix of GSoft is approximated by usual linear expansions of a differentiable approximation to the target function in Subsection 4·3. This approximation indicates, that GSoft has no variance inflation. To obtain a reasonable estimator, discontinuities in the approximate covariance matrix are smoothed in Subsection 4·4.

### 4·1 *Decomposition of the estimation error*

Suppose that the matrix of thresholds, $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_p)$ and the trade–off parameter $\lambda > 0$ is fixed in advance and let $\beta^*$ denote a maximizer of the expected absolute penalized loglikelihood in the GSoft criterion. In the case of uniqueness, this is equivalent to the root of

$$u^*(\beta) = \mathrm{E}\{u(\beta)\}, \qquad u(\beta) = s(y; \beta) - \lambda\Gamma\{g(\beta)\}, \tag{19}$$

and we have

$$u^*(\beta^*) = 0$$

The components of $\Gamma g(\beta)$ in (19) are defined by

$$\gamma_j g(\beta_j) = \begin{cases} -\gamma_j & \text{für} \quad \beta_j < 0 \\ s_j(y;\beta)/\lambda & \text{für} \quad \beta_j = 0 \\ \gamma_j & \text{für} \quad \beta_j > 0 \end{cases} \tag{20}$$

and satisfy $|\gamma_j g(\beta_j)| \leq \gamma_j$.

If the true linear predictor $\eta$ is a linear function of the columns in $Z$, the corresponding coefficients $\beta$ are termed the true model. A true model exists, provided $Z$ has full rank $n$. As with other penalized likelihood estimators, compare e.g. Cox and O'Sullivan (1996), the estimation error of GSoft can be decomposed by

$$\hat{\beta} - \beta = (\beta^* - \beta) + (\hat{\beta} - \beta^*) \tag{21}$$

into two terms where $(\beta^* - \beta)$ is a deterministic error or systematic bias, which is due to the biased estimation and due to selection of model terms. The second term $(\hat{\beta} - \beta^*)$ comprises random errors, but also further bias.

The error decomposition (21) is even correct when the true $\eta$ cannot be expressed by a linear combination of the form $Z\beta$. In this situation, $\beta$ corresponds, in the sense of Kullback–Leibler distances, to an optimal approximation to the true model, compare Fahrmeir (1990). Therefore, properties of the estimator can be characterized even in the case, where the set of regressors or basis functions is not rich enough to ensure the existence of a true model.

### 4·2 *Systematic bias*

Let $\beta$ denote a true model or an optimal approximation to the true model and consider the linear expansion

$$0 = u^*(\beta^*) \approx \mathrm{E}\{s(y;\beta)\} - F(\beta)(\beta^* - \beta) - \lambda\Gamma\mathrm{E}\{g(\beta^*)\} \tag{22}$$

of the score functions expectation around $\beta$. Provided that $F(\beta) > 0$, $\mathrm{E}\{s(y;\beta)\} = 0$ yields

$$b(\beta^*) = (\beta^* - \beta) \approx \lambda F(\beta)^{-1}\Gamma\mathrm{E}\{g(\beta^*)\} \tag{23}$$

as approximation for the deterministic error or systematic bias..

If the embedded model $\beta_{(0)}$ is a true model then, by $\gamma_j = 0$ for all $\beta_j \neq 0$ and $\mathrm{E}\{s_j(y,\beta)\} = \mathrm{E}\{s_j(y,\beta^*)\} = 0$ for all $\beta_j = 0$, we have $\Gamma\mathrm{E}\{g(\beta^*)\} = 0$ in (23) and there is no deterministic error, i.e. $\beta \approx \beta^*$.

By the form of (20) the systematic bias of GSoft is bounded in $\beta_j^* \in \mathbb{R}^p$ and attains its maximum when $\beta_j^* \neq 0$, respectively $g(\beta_j^*) = \pm 1$ for all $j = 1, \ldots, p$. Since $g(\beta^*)$ is bounded, this approximation indicates that GSoft has limited bias in $\beta \in \mathbb{R}$. This property is an important feature of the soft–thresholding studied by Donoho and Johnstone (1994) and Bruce and Gao (1996).

In the limit $n \to \infty$, the systematic bias tends to 0, provided $\lambda \mathrm{E}\{g(\beta^*)\}$ increases at a lower rate than $F(\beta)$. If $\lambda$ is chosen at an appropriate rate, GSoft is asymptotically unbiased under usual maximum likelihood regularity conditions on $F(\beta)$. In the more interesting situation when the design matrix grows with $n$, the limit behavior of $\lambda \mathrm{E}\{g(\beta_j^*)\}$ has to be studied in detail. This is a topic of future research.

### 4·3 *Stochastic error and variance*

Approximations to the variance of GSoft are derived by linear expansions from the stochastic error term $(\hat{\beta} - \beta^*)$ in decomposition (21).

Let

$$a(\beta_j, \delta) = \begin{cases} -\beta_j, & \text{für} \quad \beta_j < -\delta \\ \frac{\beta_j^2 + \delta^2}{2\delta}, & \text{für} \quad -\delta \leq \beta_j \leq \delta \\ \beta_j, & \text{für} \quad \beta_j > \delta \end{cases} \qquad (24)$$

be a continuously differentiable piecewise polynomial approximation to the penalty, satisfying $\lim_{\delta \to 0} a(\beta_j, \delta) = |\beta_j|$, compare Tishler and Zang (1982). An approximation for the variance of $\hat{\beta}(\delta)$, defined as maximizer of the function

$$lp_\delta(y; \beta, \lambda) = l(y; \beta) - \lambda \sum_{j=1}^{p} \gamma_j a(\beta_j, \delta). \qquad (25)$$

is derived first. Let $g(\beta, \delta) = \{g(\beta_1, \delta), \ldots, g(\beta_p, \delta)\}$ with the components

$$g(\beta_j, \delta) = \begin{cases} -1 & \text{für} \quad \beta_j < -\delta \\ \frac{\beta_j}{\delta} & \text{für} \quad -\delta \leq \beta_j \leq \delta \\ +1 & \text{für} \quad \beta_j > \delta \end{cases}$$

denote the first derivative and

$$G(\beta, \delta) = \frac{\partial a(\beta, \delta)}{\partial \beta \partial \beta'} = \mathrm{diag}\left( \frac{I\{|\beta_1| \leq \delta\}}{\delta}, \ldots, \frac{I\{|\beta_p| \leq \delta\}}{\delta} \right)$$

denote the second derivative of the approximation $a(\beta, \delta)$. The maximizer $\hat{\beta}(\delta)$ can be characterized as

$$u_\delta(\beta) = s(y; \beta) - \lambda \Gamma g(\beta, \delta) = 0,$$

which is expanded linearly around $\beta^*$ by

$$0 = u_\delta\{\hat{\beta}(\delta)\} \approx u_\delta(\beta^*) - \{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}\{\hat{\beta}(\delta) - \beta^*\}$$

yielding

$$\{\hat{\beta}(\delta) - \beta^*\} \approx \{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}^{-1} u_\delta(\beta^*) \tag{26}$$

as first approximation to the stochastic error $(\hat{\beta} - \beta^*)$. In the limit $\delta \to 0$, we have $u_\delta(\beta^*) = u(\beta^*)$, as defined in (19), compare Tishler and Zang (1982). For GSoft, the bias in the stochastic error disappears in linear approximation, since $\beta^*$ is defined by $\mathrm{E}\{u(\beta^*)\} = 0$, which results in $\mathrm{E}(\hat{\beta}) \approx \beta^*$.

Using the relation

$$\mathrm{Var}\{u_\delta(\beta^*)\} = \mathrm{Var}\{s(y; \beta^*)\},$$

the approximate variance of $\hat{\beta}(\delta)$ can be calculated by

$$\begin{aligned} V_\delta(\beta^*) &= \mathrm{Var}\{\hat{\beta}(\delta)\} \\ &= \mathrm{Var}\{\hat{\beta}(\delta) - \beta^*\} \\ &\approx \mathrm{Var}\left[\{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}^{-1} s(y, \beta^*)\right], \end{aligned} \tag{27}$$

which leads to the well–known sandwich form

$$V_\delta(\beta^*) = \{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}^{-1} \mathrm{Var}\{s(y, \beta^*)\}\{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}^{-1} \tag{28}$$

of Huber (1967).

First, we consider $\delta > 0$ to be given. In the case where $|\beta_j^*| > \delta$ for all $j = 1, \ldots, p$, we have $G(\beta^*, \delta) = 0$ and $V(\delta)$ results in the sandwich form

$$H(\beta^*)^{-1} \mathrm{Var}\{s(y, \beta^*)\} H(\beta^*)^{-1} \tag{29}$$

as used in possibly misspecified GLMs, compare Fahrmeir (1990). When $|\beta_j^*| \leq \delta$ for some coefficient, we have $G(\beta_j^*, \delta) = 1/\delta$ in the diagonal of $G(\beta^*, \delta)$ and by increasing the diagonal elements of $\{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}$, the eigenvalues of the inverse decrease, resulting in a reduced variance of $\hat{\beta}(\delta)$. Hence, the approximation indicates that GSoft prevents from variance inflation. The trade–off parameter $\lambda$ tunes the variance reduction: It determines the number of $\beta_j^* = 0$ and $G(\beta_j^*, \delta) > 0$ and directly reduces variance by controlling the eigenvalues of $\{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}^{-1}$.

In the limit $\delta \to 0$, leading to GSoft, the diagonal elements of $G(\beta^*, \delta)$ corresponding to $\beta_j^* = 0$ tend to $\infty$. As a consequence, corresponding eigenvalues of $\{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}$ explode and we have to consider a generalized inverse in (28). The covariance matrix $V = \lim_{\delta\to 0} V_\delta(\beta^*)$ becomes singular, approximating the variance of components having $\beta_j^* = 0$ by 0. For the remaining components, (28) leads to an approximate covariance of the form (29). However, by explosion of the eigenvalues of $\{H(\beta) + \lambda\Gamma G(\beta, \delta)\}$, usual regularity conditions of the asymptotic theory are not fulfilled.

### 4·4 *Estimation of the covariance matrix*

The approximation leading to (28) is useful, when $|\beta_j^*|$ is large, but it is insufficient for $\beta_j^* \approx 0$. This is mainly due to violation of the smoothness condition on $\{H(\beta^*) + \lambda\Gamma G(\beta^*, \delta)\}$, which is part of asymptotic theory, compare e.g. Fahrmeir (1990). In the following, we therefore propose an estimator of the covariance of GSoft, which is based on smoothing the jump in $\lim_{\delta\to 0} G(\beta^*, \delta)$.

Consider a continuous random variable $X_j$ having $\mathrm{E}(X_j) = \beta_j^*$ and replace the diagonal elements $G(\beta_j^*, \delta)$ from $G(\beta^*, \delta)$ by $\mathrm{E}\{G(X_j, \delta)\}$. This results in

$$\begin{aligned}\mathrm{E}\{G(X_j, \delta)\} &= \mathrm{E}\left\{\frac{I\{|X_j| \le \delta\}}{\delta}\right\} \\ &= \frac{1}{\delta}\int_{-\delta}^{\delta} f_{X_j}(x)dx \\ &= \frac{1}{\delta}\{F_{X_j}(\delta) - F_{X_j}(-\delta)\}. \end{aligned} \qquad (30)$$

and in the limit $\delta \to 0$ we have the simple form $\mathrm{E}\{G(X_j, 0)\} = 2f_{X_j}(0)$.

Using a normal distribution $\phi$ having mean $\beta_j^*$ and variance $\sigma_j^2$ as a smoothing kernel leads to

$$G^*(\beta^*, \sigma) = \mathrm{diag}\left\{\frac{2}{\sigma_1}\phi(\beta_1^*/\sigma_1), \ldots, \frac{2}{\sigma_p}\phi(\beta_p^*/\sigma_p)\right\} \qquad (31)$$

as smoothed version of $G(\beta^*, 0)$ and substitution of $\beta^*$ by $\hat\beta$ results in

$$\hat V(\hat\beta_j) = \{H(\hat\beta) + \lambda\Gamma G^*(\hat\beta, \hat\sigma)\}^{-1}\hat F(\hat\beta)\{H(\hat\beta) + \lambda\Gamma G^*(\hat\beta, \hat\sigma)\}^{-1}, \qquad (32)$$

as estimator of the covariance.

In (32) we still need a pilot estimate for the smoothers bandwidth $\sigma^2$. The proposed approach suggests to replace $X_j$ by $\hat\beta_j$ and $\sigma_j^2$ by $\mathrm{Var}(\hat\beta_j)$. A conservative pilot estimator for $\hat\sigma_j^2$ can be constructed by

$$(\hat\sigma_1^2, \ldots, \hat\sigma_p^2) = \mathrm{diag}\left[\{H(\hat\beta)\}^{-1}\hat F(\hat\beta)\{H(\hat\beta)\}^{-1}\right], \qquad (33)$$

which is an estimator for the variance of the maximum likelihood estimate. Simulation studies have shown, that the estimator (32) is quite insensitive against the choice of the kernel $f_X$. However, it leads to a systematic overestimation of the true variance, caused by the conservative pilot estimate. A two stage approach, where the pilot estimate of $\sigma_j^2$ is corrected in a second step by (32), overcomes the problem of overestimation.

Tests on the general linear hypothesis can be derived from the formula for the estimators variance, in principle. However, the results depend on the dimension of the model and on the trade off parameter.

### 4·5 *Inequality constraints*

When inequality constraints are employed, as described in section 2·2, a kernel with nonnegative (nonpositive) support has to be used in (31). We propose to use a truncated normal distribution

$$\phi^+(x) = \frac{\phi(x)}{\Phi(x)}, \quad \text{resp.} \quad \phi^-(x) = \frac{\phi(x)}{1-\Phi(x)} \tag{34}$$

for a constraint $\beta_j \geq 0$, respectively $\beta_j \leq 0$. Simulation studies have shown, that together with the two stage pilot estimate, this strategy gives appropriate results.

## 5 Simulation study

In this section, we describe a simulation study, which is part of extensive simulation experiments that have been conducted to investigate into the properties of the methods proposed. We consider a logit model with the linear predictor

$$\begin{aligned} \eta_i = \;& \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\ & x_1 \;=\; \{(i-15.5)/10\}^2 \\ & x_2 \;=\; \log(i/5) \end{aligned} \tag{35}$$

for $i = 1, \ldots, 30$. The coefficient of correlation between $x_1$ and $x_2$ is $-0.3370$ and the response $y_i$ follows a binomial $B(m, \pi_i)$ distribution.

The goodness of fit is measured by the Kullback–Leibler distance or expected deviance, which is adjusted by

$$\inf_{\lambda} \frac{R_{KL}(\hat{\mu}(\lambda), \mu)}{R_{KL}(\mu^{ml}, \mu)} \tag{36}$$

on a corresponding maximum likelihood estimate $\mu^{ml}$.

### 5·1 *Risk reduction*

— Figure 3 about here —

Figure 3 shows the Kullback–Leibler risk ratio (36) of a GSoft estimate $\hat{\eta}$ depending on $x_1$ and $x_2$ with $m = 5$. The thresholds have been adjusted according to (18), with only $\beta_0 = -1/2$ included in the embedded model.

Since GSoft leads to a maximum likelihood estimate for $\lambda \to 0$, the relative risk (36) is bounded by 1. Around the origin at $\beta_1, \beta_2 \in [-0.2, 0.2]$, GSoft has less than half of the risk of a maximum likelihood estimate. The ellipsoidal form of this region is due to the correlation between $x_1$ and $x_2$. We observe a further reduction of risk near the coordinate axes in Fig. 3.

### 5·2 *Covariance estimation*

— Figure 4 about here —

The simulation, shown in Fig. 4, is based on the parameters $\beta_0 = -0.5$, $\beta_1 = 0$, $\beta_2 = 1$, $m = 2$ and $n = 30$. It demonstrates the behavior of variance estimation by selection of $x_1$ and shrinkage of $\beta_2$. With decreasing variance of GSoft, the variance of the estimator for the covariance matrix decreases as well. When $\lambda = 0$, GSoft is equivalent to the maximum likelihood estimate and the variance estimator from (32) reduces to the inverse Fisher matrix. It underestimates the variance in more than 75 % of the runs, compare Fig. 4 (a,b,c). With increasing $\lambda > 0.25$, the median of the variance estimate is near the true variance. The distribution in Fig. 4 (a) becomes very skewed for $\lambda$ closed to 1. This is due to the fact that, with high probability, we have $\hat{\beta}_1 = 0$ and the estimates $\hat{\beta}_1$, $\hat{\beta}_2$ are nearly uncorrelated in Fig. 4 (c). Therefore, the variance estimation for $\hat{\beta}_1 = 0$ often yields similar results.

— Figure 5 about here —

The simulation in Fig. 5 shows the behavior of the estimator under inequality constraints. It is based on small coefficients $\beta_0 = 0.25$, $\beta_1 = 0.25$ and $\beta_2 =$

0.25 and a small sample size $m = 1$, $n = 30$.e The coefficients $\hat{\beta}_1$ und $\hat{\beta}_2$ are estimated under the constraints $\hat{\beta}_1 \geq 0$, $\hat{\beta}_2 \geq 0$ and the intercept is included in the embedded model.

Interpretation of the variance estimates is analogous to Fig. 4. On the left border we observe, that the maximum likelihood estimate corresponding to $\lambda = 0$ has much higher variance. At this small sample size, it is heavily underestimated by the inverse of the Fisher matrix.

## 6 Function estimation by GSoft

An important application of GSoft is in the area of function estimation in structured nonparametric regression models such as generalized additive models (Hastie and Tibshirani, 1990) or varying–coefficient models (Hastie and Tibshirani, 1993). In this situation, the high dimensional design arises from representing each predictor function $f_j(x_j)$ by

$$f_j(x_j) = \sum_{k=1^{n_j}} \phi_{jk}(x_j)\beta_{jk}$$

as sum of basis functions. As has been indicated in section 4, the contribution of basis coefficients $\beta_{jk}$ to the variance of GSoft depends on their magnitude in relation to the threshold. Relatively small coefficients don't essentially contribute to the estimators variance, whereas for large $|\beta_{jk}|$ the variance is bounded by the variance of a maximum likelihood estimate. Hence, provided that the true function $f_j(x_j)$ can parsimoniously be well approximated by a linear combination of the basis functions supplied, GSoft gives efficient estimates.

By specification of an appropriate set of basis–functions, the procedure can therefore easily be tailored to specific purposes. For example, functions which are irregularly smooth can be well approximated by using a library of wavelet basis functions, compare Donoho and Johnstone (1995). Smooth functions are efficiently estimated by the orthogonal basis of Demmler–Reinsch splines (Demmler and Reinsch, 1975) in connection with thresholds depending on the smoothness of basis functions, compare Klinger (1998).

### 6·1 *Spline regression*

In the following, we focus on simple one–sided spline functions, which are specifically easy to interpret and useful in many applications. For convenience, we skip the index $j$ in $f_j(x_j)$ and consider the case of univariate function estimation first. Let

$$\begin{aligned} &\psi_k(x) = x^k, && \text{if} \quad k = 0, 1, \ldots, q \\ &\psi_k(x) = (x - x_{(k-q)})^q_+, && \text{if} \quad k = q+1, q+2, \ldots, n+q-1 \end{aligned} \tag{37}$$

with

$$(x - x_k)^q_+ = \begin{cases} (x - x_k)^q & \text{if} \quad x > x_k \\ 0 & \text{if} \quad x \leq x_k \end{cases}$$

denote a set of truncated power spline basis–functions, where $x_{(k)}$, $k = 1, \ldots, n$ refers to the ordered values of $x$. For $q = 0$, the basis (37) reduces to the set of indicator functions

$$\begin{aligned} &\psi_0(x) = 1, \\ &\psi_k(x) = I\{x > x_{(k)}\}, \quad \text{if} \quad k = 1, 2, \ldots, n-1. \end{aligned} \tag{38}$$

When setting up a predictor function by linear combinations of these basis functions as

$$f(x) = \beta_0 + \sum_{k=1}^{q} \beta_k x^k + \sum_{k=q+1}^{n+q-1} \beta_k (x - x_{(k-q)})^q_+,$$

the $q$–th left sided derivative can be expressed as a simple step function

$$f^{(q-)}(x) = \beta_q + \sum_{k=q+1}^{n+q-1} \beta_k I\{x > x_{(k-q)}\}. \tag{39}$$

It follows, by

$$\beta_{i+q} = f^{(q-)}(x_{(i+1)}) - f^{(q-)}(x_{(i)}),$$

that each coefficient of the spline is identified as a jump in the $q$–th derivative.

Let $Z = (z_0, \ldots, z_p)$ with entries $z_k = \{\psi_k(x_{(1)}), \ldots, \psi_k(x_{(n)})\}'$ for $k = 0, \ldots, p$ and $p = n+q-1$ be a design matrix for the predictor function $f(x)$, then GSoft results in a penalized likelihood estimator $l(y, f) - J(f)$ with the penalty

$$J(f) = \sum_{k=0}^{q} \gamma_k \beta_k + \sum_{i=1}^{n-1} \gamma_{q+i} \left| f^{(q-)}(x_{(i+1)}) - f^{(q-)}(x_{(i)}) \right|. \tag{40}$$

Including all polynomial terms up to order $q$ into the embedded model (i.e. $\gamma_0 = \cdots = \gamma_q = 0$), the penalty reduces to

$$J_{S_q}(f) = \sum_{i=1}^{n-1} |f^{(q-)}(x_{(i+1)}) - f^{(q-)}(x_{(i)})| \tag{41}$$

and judges the variation in the $q$–th derivative. Penalties of this kind have been studied in detail by Mammen and van de Geer (1997). There it is shown, that the form (40) arises from an infinite dimensional total variation penalty on the derivatives. Furthermore, the authors derive asymptotical optimality results for the estimator, provided that the true function has bounded total variation of the $q$–th derivative. For $q = 0$, this class includes even functions with discontinuities, as e.g. step functions.

### 6·2 *Monotonicity constraints*

Since by (39), each basis coefficient corresponds to an increment of the $q$-th derivative, nonnegativity constraints on $\beta_{i+q}$ lead to positive increments of the function $f(x)$. Therefore, monotonicity restrictions on the $q$-th derivative can easily be transformed into nonnegativity constraints on the basis coefficients, as described in Section 2·2.

### 6·3 *Adjusting the thresholds*

In the context of univariate function estimation, the GSoft estimator with thresholds $\gamma_0 = \cdots = \gamma_q = 0$, $\gamma_{q+1} = \cdots = \gamma_{n+q-1} = 1$ corresponds to a penalized likelihood estimator (41) and no adjustment is necessary. If several functions are specified, as in generalized additive models, thresholds have to be adjusted appropriately to account for the scaling of the different covariates.

We proceed as in Section 3 and compute corresponding robustified score statistics for each basis coefficient $\beta_{jk}$. The thresholds $\gamma_{jk}$, $k = 1, \ldots, n_j$ for all basis functions contributing to one term $f_j(x_j)$ are adjusted according to the average estimated variance of the score function under the embedded model:

$$\gamma_{jk} = \left\{ \frac{1}{n_j} \sum_{k}^{n_j} \{R(\hat{\beta}_{(0)})^{-1}\}^{-1}_{jk,jk} \right\}^{\frac{1}{2}}$$

By this strategy, all basis functions describing one single term have the same threshold and the penalized likelihood representation in (40) remains valid also for models with multiple predictor functions.

## 7 Applications

### 7·1 *Credit scoring*

In parametric models, the proposed GSoft methodology is illustrated on the credit scoring data, described in Fahrmeir, Hamerle and Tutz (1996a) and available form the data archive `http://www.stat.uni-muenchen.de/data-sets`. The purpose is to develop a classification rule for solvency prognosis on the basis of a sample of 300 bad and 700 good consumer credits. Besides the response variable

$$y_i = \begin{cases} 1, & \text{client } i \text{ is creditworthy} \\ 0, & \text{client } i \text{ is not creditworthy,} \end{cases}$$

there are 20 covariates, which are summarized in Table 1.

— Table 1 about here —

Except duration ($z_2$), amount of credit ($z_5$) and age ($z_13$), all covariates are on an ordinal scale. This scale is derived from a scoring system from experts, where high values indicate good credit worthiness. The dataset served as basis for a number of investigations in this form. For example, Fahrmeir and Kredler (1984) compare variable selection procedures in the logit model and Klinke and Grassmann (1996) use this data to test neural networks.

— Table 2 about here —

To illustrate the performance of GSoft, the data are randomly divided into a training sample of size $n = 200$ and a validation sample with the remaining 800 observations. We assume a univariate logit model

$$\log \frac{P(y=1|z)}{P(y=0|z)} = \beta_0 + z_1\beta_1 + \cdots + z_{20}\beta_{20}, \tag{42}$$

with all covariates included.

The results, gained by application of the `glm()` function in S–Plus to the total data and to the test data are summarized in Table 2. Furthermore, we applied the function `step.glm()`, targeting on the variable selection criterion

$$AIC = l(y; \beta^{sel}) - 2\sum_{j=1}^{p} I\{\beta_j^{sel} \neq 0\},$$

on the test data. Except for $z_6$ and $z_8$ all significant terms from the total sample are selected by the variable selection algorithm. The effects of the covariates, selected from the learning sample, have about double of the size of the effects estimated from the entire population. This indicates a considerable selection bias of the stepwise variable selection procedure. The classification result, based on the logit model in the test sample and the maximum likelihood classification rule

$$\hat{\eta}_i^{sel} \geq \log(144/56)$$

is shown in Tab. 3.

— Table 3 about here —

— Figure 6 about here —

The same model is analyzed by GSoft. Obviously, the covariate $z_1$ indicating the client's running account is an important indicator. Hence, it is included in the embedded model, which is defined by the design matrix $Z_0 = (1, z_1)$. Results of GSoft based on the learning sample are shown for $\lambda = \{0, 0.01, \ldots, 4\}$ in Fig. 6. A detailed description of the efficient algorithm, which is based on the approximation (24) can be found in Klinger (1998). By decreasing $\lambda$, the number of terms in the model decreases as well. All coefficients $\beta_j$ appear as a smooth function in $\lambda$.

Let $\hat{\eta}_i^{-i}$ denote the linear predictor of an observation $y_i$, estimated by GSoft applied to all data except $(y_i, z_i)$. The global threshold $\lambda = 0.91$ has been chosen as minimizer of the cross–validated deviance criterion

$$CV_D(\lambda) = -2 \sum_{i=1}^{n} \{l(y_i, \hat{\eta}_i^{-i}) - l(y_i, y_i)\}, \tag{43}$$

which is shown in Fig. 6.

The results of GSoft, based on cross–validation only the data from the learning sample are summarized in right two columns of Tab. 2. GSoft identifies the covariate $z_{12}$ having the $p$–value 0.0752 in the total sample in addition to the terms selected by the variable selection procedure. Compared to the maximum likelihood estimation in the selected model, the magnitude of the effects are generally reduced and about the same as in the whole population. As a consequence, missclassification in the validation sample is considerably lower than missclassification rate obtained after the variable selection procedure.

— Table 4 about here —

After we have studied the classification properties of GSoft by dividing the dataset into a learning and a validation sample, we applied the procedure to the total sample of 1000 consumer credits. When the trade–off parameter is chosen as minimizer of the cross–validated deviance, only the covariates $z_{11}$, $z_{17}$ and $z_{18}$ are dropped from the model at $\lambda^{CV_D} = 0.41$. By a cross–validated missclassification rate, we obtain $\lambda^{CV_M} = 0.21$ as optimal trade–off parameter and only $z_{17}$ is dropped from the model.

— Table 5 about here —

The average missclassification rates of GSoft, shown in Tab. 5 are considerably lower than the best rates reported in Fahrmeir et al. (1996b, p. 394), where different classification procedures are compared. There, the best cross–validated missclassifikation rate is obtained by a linear discriminant analysis based on 15 covariates, compare Tab. 5. Also Klinke and Grassmann (1996), who compare different neural networks, couldn't improve the classification rate of linear discriminant analysis.

### 7·2 *Rental guide*

The second example illustrates the GSoft methodology in models with multiple functional terms in the linear predictor. According to the German rental law, owners of apartments or flats can base an increase in the amount that they charge for rent on the "usual rents" for flats comparable in type, size, equipment, quality and location in a community. Commonly, these "usual rents" are calculated from an official rental guide ("Mietspiegel"). Our data are based on a random sample of 1969 flats in Munich, conducted 1993 to construct a rental guide, compare Fahrmeir, Gieger, Mathes and Schneeweiß (1994). The official guide is based on two metrical covariates $F$ (floor space in square meters) and $A$ (year of construction). and a set of indicator variables, shown in Tab. 6

— Table 5 about here —

A main criticism of the official rental guide, calculated from a nonlinear parametric regression on the response variable $R$ ( monthly net rent in DM), from Fahrmeir, Gieger, Mathes and Schneeweiß (1994), is that possible interactions between the indicators and $F$ or $A$ have not been accounted for.

### 7·3 *Model specification*

Fahrmeir et al. (1998), study some general transformation models and conclude, that the response variable $R$ follows approximately a gamma distribution. This model can derived by assuming a constant coefficient of variation in $R$, compare McCullagh and Nelder (1989, Ch. 8). In the following model we account for interactions between an indicator $X_j$ and $F$ or $A$ by assuming nonparametric terms of the form $\{f_1(F) + f_2(A)F\} * X_j$ for each indicator. The identity is used as linkfunction in the gamma model. Hence, the effect of $A$ and each interaction with $A$ can directly be interpreted as a surcharge or discount on the rent per square meter. The assumed predictor

$$\begin{aligned}\mu \;=\;& f_1(F) + f_2(A)F + \{f_3(F) + f_4(A)F\}S^+ + \{f_5(F) + f_6(A)F\}S^- \\ &+\{f_7(F) + f_8(A)F\}A^+ + \{f_9(F) + f_{10}(A)F\}A^- + \{f_{11}(F) + f_{12}(A)F\}Bd^- \\ &+\{f_{13}(F) + f_{14}(A)F\}Zh^- + \{f_{15}(F) + f_{16}(A)F\}W^- + \{f_{17}(F) + f_{18}(A)F\}Bl^+ \\ &+\{f_{19}(F) + f_{20}(A)F\}F^+ + \{f_{21}(F) + f_{22}(A)F\}F^- + \{f_{23}(F) + f_{24}(A)F\}Z^+ \\ &+\{f_{25}(F) + f_{26}(A)F\}Bd^+ + \{f_{27}(F) + f_{28}(A)F\}K^+ + \{f_{29}(F) + f_{30}(A)F\}G^+ \\ &+\{f_{31}(F) + f_{32}(A)F\}R^+ + \{f_{33}(F) + f_{34}(A)F\}Ab^- + \{f_{35}(F) + f_{36}(A)F\}H^- \end{aligned} \tag{44}$$

consists of 36 functional terms and has the form of a varying–coefficient model, compare Hastie and Tibshirani (1993). In the basis configuration of (44), we take into account possible breakpoints in the interactions with $f(A)$, by using the set of indicator functions as basis. This representation of effects allows the communication of the results in common tabular form, in addition. The main effect $f(A)$ is modelled by a quadratic spline–function, whereas the main effect $f(F)$ and all interactions with floor space are described as piecewise linear splines. This is according to the assumption of a continuous effect in $F$, which has a linear component corresponding to the average effect per square meter. Note that a linear interaction of the form $\beta_j F X_j$ acts constantly per square meter. To ensure identifiability, all constant terms are excluded from the set of basis functions, used to describe functions in $A$. Moreover, we excluded the constant terms in the coefficients varying over floor space, to account for no effect for flats having

0 square meters. The embedded model, derived from (44)

$$\begin{aligned}\mu &= \beta_0 + \beta_1 F + \beta_2 AF + \beta_3 A^2 F \\ &\quad + \{\beta_4 S^+ + \beta_5 S^- + \cdots + \beta_{21} H^-\}F,\end{aligned}$$

assumes a constant effect of each indicator on the rent per square meter.

### 7·4 *Model estimation*

The smoothing parameter has been chosen by a randomized 44–fold cross–validated deviance criterion, where each block was buildt around one flat without bathroom. A global minimum of the cross–validation function was found at $\lambda^{CV_D} = 0.79$, where 71 of 1617 possible basis functions enter the model. The covarince estimation of GSoft is only based on selected basis functions and on the moment estimator

$$\hat{\nu}^{-1} = \frac{1}{n-71}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)^2,$$

for the dispersion parameter $\nu$ in the gamma model, compare McCullagh and Nelder (1989, S. 196).

— Figure 7 about here —

Figure 7 shows the main effects from model (44). The linear spline

$$\begin{aligned}\hat{f}_1(F) &= \hat{\beta}_{1,0} + \hat{\beta}_{1,1}F + \hat{\beta}_{1,37}(F-68)_+ \\ &= 320.28 + 5.067F + 0.3170(F-68)_+ \\ &= \{320.28 + 5.067F\}I\{F \leq 68\} + \{659.77 + 5.384F\}I\{F > 68\}\end{aligned} \tag{45}$$

can be interpreted as follows: For the first 68 m$^2$, we have to add 5.067 DM/m$^2$ on the base rent of 320.28 DM. Each additional square meter costs 5.067+0.3170=5.384 DM. Corresponding standard deviations (0.219, 0.269) are based on the quadratic form $e'\hat{V}\{\hat{\beta}\}e$, where $e$ marks corresponding coefficients by 1. The quadratic spline $f_2(A)$ has a single knot in 1940.

Table 8 and Tab. 7 show discounts and surcharges for the different indicators, which depend on $F$ and $A$. The tabular form of the interactions with $F$ is calculated by cummulation of basis coefficients, as described above and corresponds to the derivative of the functions $\hat{f}_j(F)$. With the restriction $\hat{f}_j(0) = 0$, they can be interpreted as effect on the rent per additional square meter. For the interactions with $A$, the reported functions $\hat{f}_j(A)$ directly effect the rent per square meter.

— Table 7 about here —

— Figure 8 about here —

For lack of space, we only give a brief summary of the results. The indicators $Zh^-$, $F^-$ and $Bl^+$ show a nonlinear interaction with floor space. No central heating ($Zh^-$) leads to a larger discount for bigger flats, where a central heating is more benificial. The coefficients for $F^-$ and $Bl^+$ are not monotone in $F$ and show some variaton around a constant. The effect of a balcony in a flat built up before 1970 varies between 69.63 DM at 55 m$^2$, 11.35 DM at 87 m$^2$ and 168.58 DM at 120 m$^2$ and cannot be assumed to be constant on the rent per m$^2$.

The indicators $Bd^+$, $Z^+$ and $K^+$ show a strong interaction with $A$. For example a good equipped kitchen $K^+$ gives a much bigger effect in older buildings. The reason might be, that the equipment of a kitchen is nearly standardized in newer buildings, whereas in old buildings a well equipped kitchen indicates a higher general standard of the flat.

## 8 Discussion

High dimensional generalized linear models combine a number of nonparametric extensions into one model class. All these models can be handeled within the GSoft framework, which allows the user to specify a high dimensional design reflecting his uncertainty about the data generating process. In analogy to robust statistics, the uncertain model formulation has its counterpart in the estimating equations, which allow the score function to be in an interval around 0. This scope is used to select basis functions and to reduce the estimators variance.

Of course, GSoft is not preferable in any situations. For example, the maximum likelihood principle might do better when a parsimonious description of the underlying process can be assumed in advance and linear smoothers might have lower risk in some prespecified smoothnes class. These assumptions are often difficult to state and to verify in real data situations. A careful investigation of a high dimensional model by GSoft can then help to improve the model structure and the estimates substantially. This has been demonstrated by the two examples presented.

## A Proof of the theorem

We follow the characterization of conditions for an optima under inequality constraints, given in Gill, Murray and Wright (1981, Ch. 3). The absolute penalized likelihood criterion can equivalently be expressed by

$$lp(y;\beta,\lambda) = l\{y,(\beta)_+ - (\beta)_-\} - \lambda \sum_{j=1}^{p} \gamma_j \{(\beta_j)_+ + (\beta_j)_-\} \tag{46}$$

under the constraints $(\beta_j)_+ \geq 0, (\beta_j)_- \geq 0$ for $j = 1,\ldots,p$. Let $A^+ = \{j : (\hat{\beta}_j)_+ = 0\}$ $A^- = \{j : (\hat{\beta}_j)_- = 0\}$ denote the set of constraints, active at a maxima, and $C^+$, $C^-$ corresponding complements. Moreover, $e_j$ denotes the $j-$th unit vector and $E_{A^+}$ resp. $E_{A^-}$ is a matrix with rows $e'_j$, $j \in A^+$, resp. $j \in A^-$ and

$$\begin{aligned} \partial lp_1\{y;(\beta)_+ - (\beta)_-,\lambda\}/\partial(\beta_j)_+ &= s_j(y;\beta) - \lambda\gamma_j \\ \partial lp_1\{y;(\beta)_+ - (\beta)_-,\lambda\}/\partial(\beta_j)_- &= -s_j(y,\beta) - \lambda\gamma_j \end{aligned} \tag{47}$$

are partial derivatives of (46). By (47), neccessary conditions for a maximum of (46) can be stated as

$$E'_A h = -\begin{pmatrix} s(y;\hat{\beta}) - \lambda\gamma \\ -s(y;\hat{\beta}) - \lambda\gamma \end{pmatrix}, \quad \text{with} \quad h_1,\ldots,h_{p'} \geq 0, \gamma = (\gamma_1,\ldots,\gamma_p). \tag{48}$$

Due to the form of $E_A$, the conditions (48) simplify to

$$\begin{aligned} s_j(y;\hat{\beta}) &\leq \lambda\gamma_j \quad && \text{für} \quad j \in A^+, \\ s_j(y;\hat{\beta}) &\geq -\lambda\gamma_j && \text{für} \quad j \in A^-, \\ s_j(y;\hat{\beta}) &= \lambda\gamma_j && \text{für} \quad j \in C^+, \\ s_j(y;\hat{\beta}) &= -\lambda\gamma_j && \text{für} \quad j \in C^-. \end{aligned} \tag{49}$$

This leads to $|s_j(y,\hat{\beta})| \leq \lambda\gamma_j$ for a coefficient in $A = \{j : \hat{\beta}_j = 0\} = A^+ \cap A^-$. For positive coefficient in $A^- \cap C^+$ we have $s_j(y;\hat{\beta}) = \lambda\gamma_j$ and for a negative coefficient in $A^+ \cap C^-$ we have $s_j(y;\hat{\beta}) = -\lambda\gamma_j$. This is equivalent to the estimating equations in (3).

To characterize sufficient conditions, let

$$\begin{aligned} \tilde{A}^+ &= \{j : s_j(y;\hat{\beta}) &< \lambda\gamma_j\} \cap A^+, \\ \tilde{A}^- &= \{j : s_j(y;\hat{\beta}) &> -\lambda\gamma_j\} \cap A^-, \end{aligned}$$

From the vectors $\{e_j : j \in \tilde{A}^+\}$, resp. $\{e_j : j \in \tilde{A}^-\}$ we construct a matrix $\tilde{E}_A$ and let $\tilde{E}_C$ denote a matrix, with columns orthogonal to the rows in $\tilde{E}_A$. The condition for a unique maxima is

$$\tilde{E}_C'(Z,-Z)'H(\hat{\eta})(Z,-Z)\tilde{E}_C$$

to be positively definite. Let $Z_\lambda$ denote a matrix, having columns $z_j$ with $j \in \{j : |s_j(y;\hat{\beta})| = \lambda\gamma_j\}$ then $\tilde{E}_C'(Z,-Z)' = Z_\lambda$, leading to the condition $Z_\lambda' H(\hat{\eta}) Z_\lambda > 0$ in (4).

## Acknowledgements

The author thanks Ludwig Fahrmeir for many valuable comments. This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 “Statistische Analyse diskreter Strukturen.”

## REFERENCES

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

Cox, D. D. and O’Sullivan, F. (1996). Penalized likelihood–type Estimators for generalized nonparametric regression, *J. Multivariate Anal.* **56**, 185–206.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

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. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Statist. Assoc.* **90**, 1200–1224.

Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models, *Statistics* **21**, 487–502.

Fahrmeir, L. and Kredler, C. (1984). Verallgemeinerte lineare Modelle, *in* L. Fahrmeir and A. Hamerle (eds), *Multivariate statistische Verfahren, 1. Auflage*, de Gruyter, Berlin, pp. 257–295.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer, New York.

FAHRMEIR, L., GIEGER, C. AND KLINGER, A. (1998). Regression Approaches to Rental Guides, *in* R. Galata and H. Küchenhoff (eds), *Econometrics in Theory and Practice. Festschrift for Hans Schneeweiß*, Physica–Verlag, Heidelberg, pp. 241–254.

FAHRMEIR, L., GIEGER, C., MATHES, H. AND SCHNEEWEISS, H. (1994). Statistische Analyse der Nettomieten, *Gutachten zur Erstellung des Mietspiegels für München 1994*, Sozialreferat – Amt für Wohnungswesen, Landeshauptstadt München.

FAHRMEIR, L., HAMERLE, A. AND TUTZ, G. (1996a). *Multivariate statistische Verfahren, 2. überarbeitete Auflage*, de Gruyter, Berlin.

FAHRMEIR, L., HÄUSSLER, W. AND TUTZ, G. (1996b). Diskriminanzanalyse, *in* L. Fahrmeir, A. Hamerle and G. Tutz (eds), *Multivariate statistische Verfahren, 2. überarbeitete Auflage*, de Gruyter, Berlin, pp. 357–425.

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.

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

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.

HUBER, P. J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions, *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.

KLINGER, A. (1998). *Hochdimensionale Generalisierte Lineare Modelle*, Doktorarbeit, Institut für Statistik, Universität München.

KLINKE, S. AND GRASSMANN, J. (1996). Visualization and implementation of feedforward neural networks, *Discussion Paper 96/92*, SFB 373, Humboldt Universität Berlin.

MAMMEN, E. AND VAN DE GEER, S. (1997). Locally adaptive regression splines, *Ann. Statist.* **25**, 387–413.

MARX, B. D., EILERS, P. H. C. AND SMITH, E. P. (1992). Ridge Likelihood Estimation for Generalized Linear Regression, *in* P. G. M. van der Heijden, W. Jansen, B. Francis and G. U. H. Seeber (eds), *Statistical Modelling*, North–

Holland, Amsterdam, pp. 227–238.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

McDonald, P. and Diamond, I. (1990). On the fitting of generalized linear models with nonnegativity parameter constraints, *Biometrics* **46**, 201–206.

Nyquist, H. (1990). Restricted estimation of generalized linear models, *Appl. Statist.* **40**, 133–141.

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

Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models, *Commun. Statist. – Theory Meth.* **21**, 2227–2246.

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

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp. 197–206.

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

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

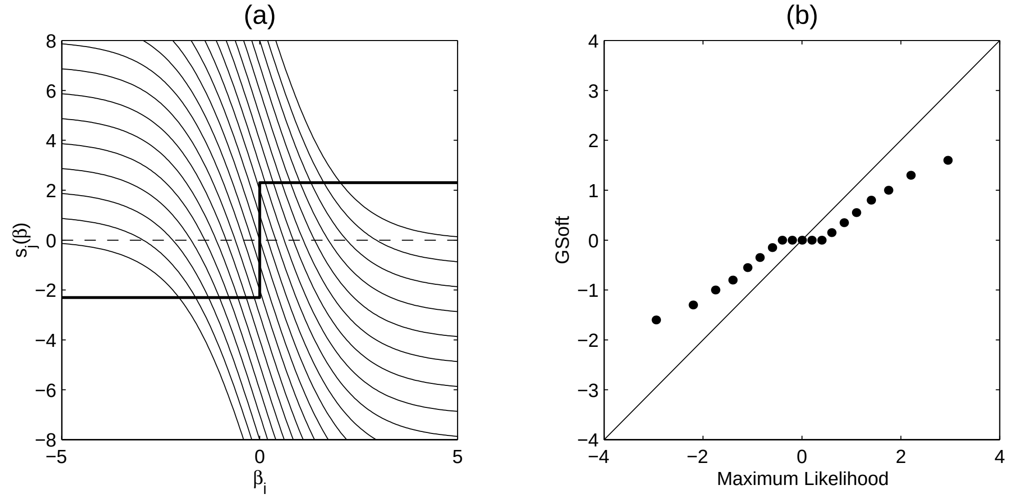

Figure 1: GSoft for a logit model with $y \sim B(20, \pi)$. In (a) the score–functions for $y = 0, \ldots, 20$ are plotted versus $\beta$. In (b), GSoft is plotted against the maximum likelihood estimator in this model. Each single estimation corresponds to the abscissa of the intersection between the score–function and the step function in (a).

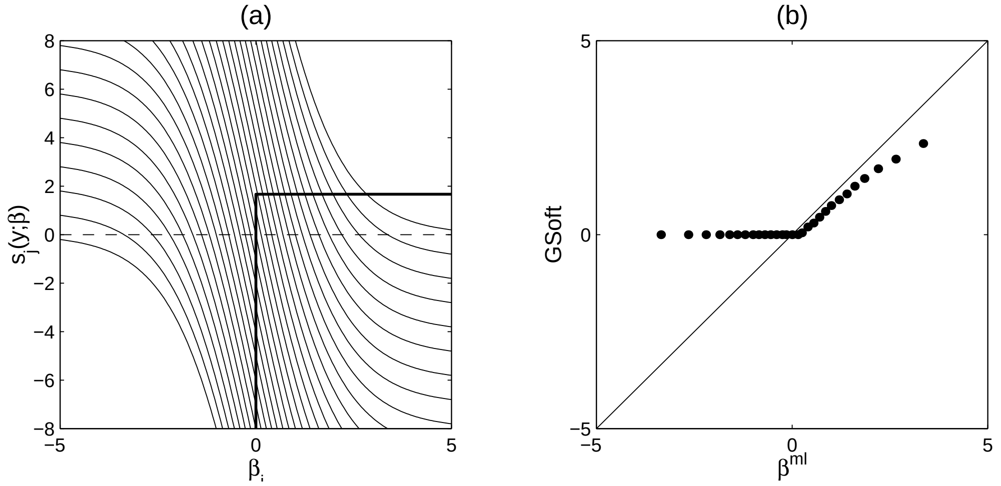

Figure 2: GSoft under the restriction $\beta_j \geq 0$ in a logit–model $y \sim B(30, \pi)$. Score functions for $y = 0, \ldots, 30$ are shown in panel (a). In (b) GSoft is plotted against the maximum likelihood estimate. It corresponds to the abscissa of the intersection between the score–function and the rectangle in (a).

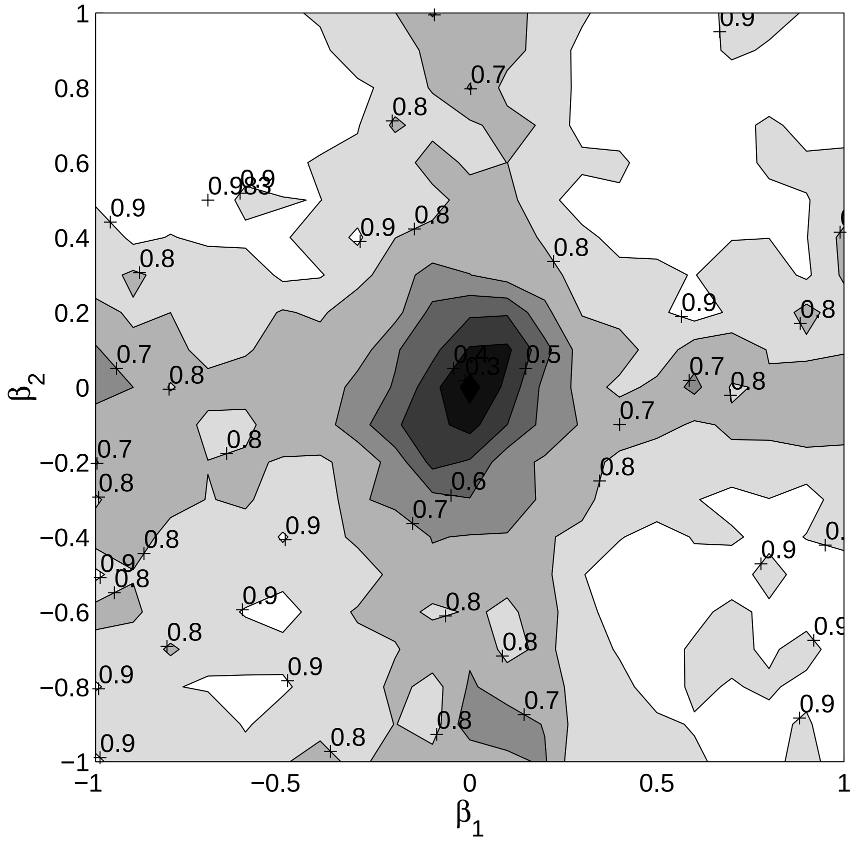

Figure 3: Kullback–Leibler risk ratio based on 100 simulations.

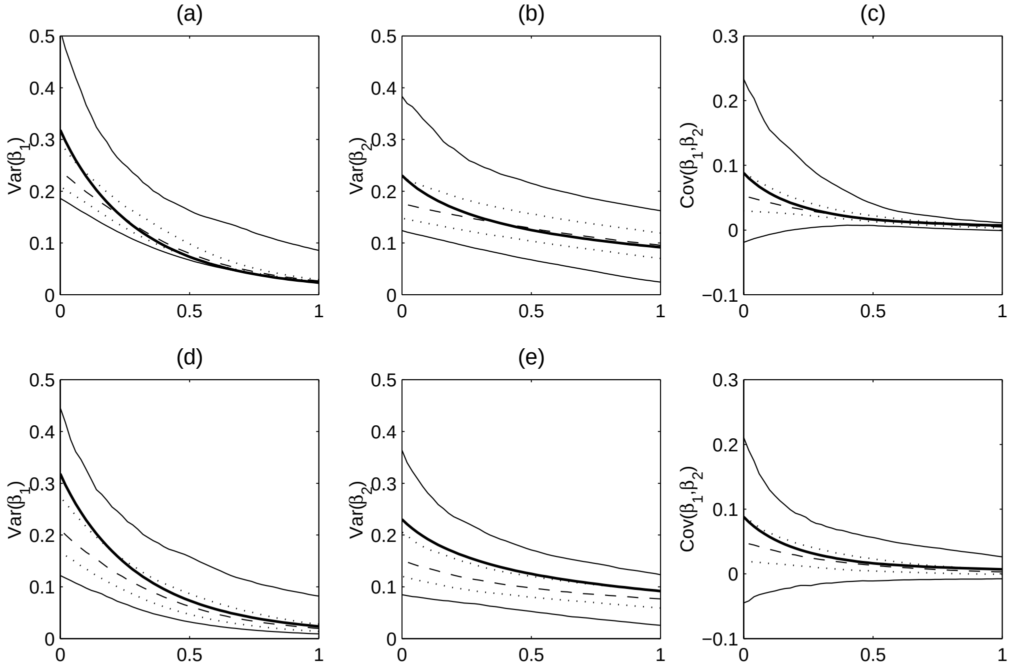

Figure 4: Variance estimation of GSoft as function of $\lambda$. The bold line corresponds to the true variance, computed from 1000 replications. Remaining lines correspond to pointwise 5%, 25%, 50%, 75%, bzw. 95% quantiles of the proposed estimator for the variance.

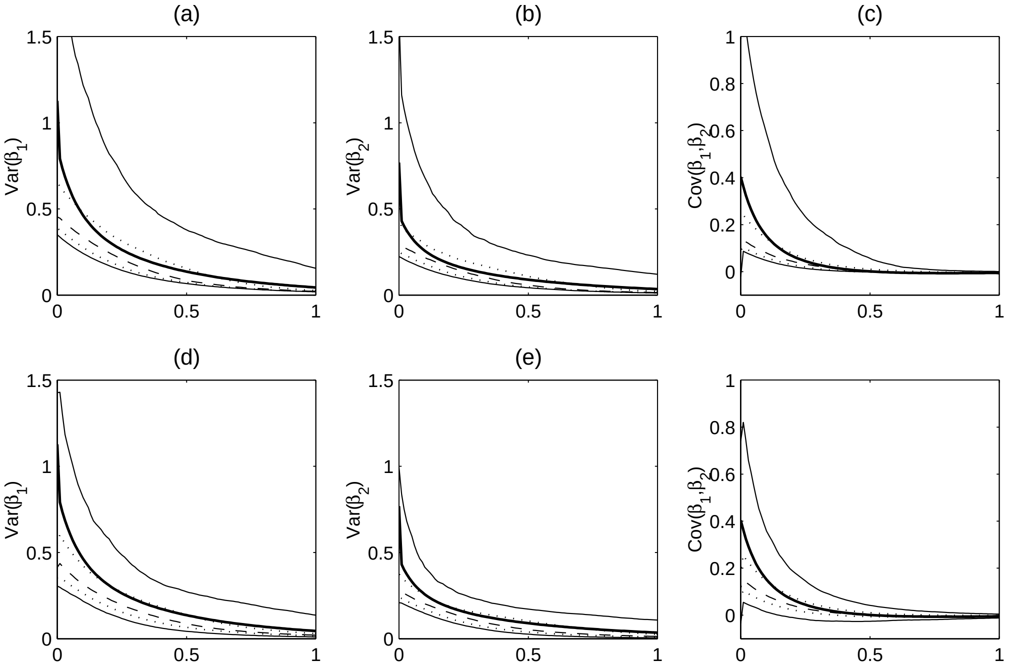

Figure 5: Variance estimation of GSoft under nonnegativity constraints as function of $\lambda$. The bold line corresponds to the true variance, computed from 1000 replications. Remaining lines correspond to pointwise 5%, 25%, 50%, 75%, bzw. 95% quantiles of the proposed estimator for the variance.

| Variable | description | range |
|---|---|---|
| $z_1$ | current account at the bank | $1, 2, 3, 4$ |
| $z_2$ | term of the credit in months | metrical |
| $z_3$ | previous repayment | $1, 2, 3, 4$ |
| $z_4$ | purpose | $0, 1$ |
| $z_5$ | total of the advance | metrical |
| $z_6$ | bankbook or securities | $1, 2, 3, 4, 5$ |
| $z_7$ | working for the present employer since | $1, 2, 3, 4, 5$ |
| $z_8$ | monthly interest in % of the disposable monthly income | $1, 2, 3, 4$ |
| $z_9$ | marital status and sex of the applicant | $1, 2, 3, 4$ |
| $z_{10}$ | other debtors / bails | $1, 2, 3$ |
| $z_{11}$ | lives in the current habitation since | $1, 2, 3, 4$ |
| $z_{12}$ | effects | $1, 2, 3, 4$ |
| $z_{13}$ | age of the applicant in years | metrical |
| $z_{14}$ | other credits | $1, 2, 3$ |
| $z_{15}$ | habitation | $1, 2, 3$ |
| $z_{16}$ | total of former credits at the bank | $1, 2, 3, 4$ |
| $z_{17}$ | job | $1, 2, 3, 4$ |
| $z_{18}$ | number of persons who are entitled to alimonies | $1, 2$ |
| $z_{19}$ | phone | $1, 2$ |
| $z_{20}$ | foreign worker | $1, 2$ |

Table 1: Description of the credit scoring data.

|  | total sample |  | learning sample |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| $j$ | $\beta_j^{ml}$ | $p$–value | $\beta_j^{ml}$ | $p$–value | $\beta_j^{sel}$ | $p$–value | $\hat{\beta}_j$ | $\hat{\sigma}(\hat{\beta}_j)$ |
| 0 | –4.3171 | 0.0001 | –2.8878 | 0.3626 | –2.9189 | 0.0092 | –1.3049 | 1.0885 |
| 1 | 0.5765 | 0.0000 | 0.7611 | 0.0003 | 0.7110 | 0.0001 | 0.6518 | 0.1358 |
| 2 | –0.2780 | 0.0220 | –0.4044 | 0.2841 | –0.4849 | 0.0946 | –0.2351 | 0.1472 |
| 3 | 0.3653 | 0.0001 | 0.5458 | 0.0258 | 0.4168 | 0.0454 | 0.1813 | 0.1223 |
| 4 | 0.5747 | 0.0029 | 1.1072 | 0.0410 | 0.9762 | 0.0484 | 0.4787 | 0.2940 |
| 5 | –0.0926 | 0.0580 | –0.1799 | 0.1659 | –0.1536 | 0.1327 | –0.0984 | 0.0529 |
| 6 | 0.2428 | 0.0002 | –0.0353 | 0.7799 | 0 | — | 0 | 0.0596 |
| 7 | 0.1536 | 0.0794 | 0.6187 | 0.0060 | 0.5669 | 0.0054 | 0.3210 | 0.1425 |
| 8 | –0.3082 | 0.0008 | –0.2314 | 0.4603 | 0 | — | 0 | 0.0608 |
| 9 | 0.2601 | 0.0637 | –0.0421 | 0.7909 | 0 | — | 0 | 0.1165 |
| 10 | 0.3017 | 0.1937 | 0.3399 | 0.5359 | 0 | — | 0 | 0.1214 |
| 11 | –0.0300 | 0.7408 | –0.1156 | 0.6689 | 0 | — | 0 | 0.0580 |
| 12 | –0.1982 | 0.0752 | –0.3775 | 0.2443 | 0 | — | –0.1778 | 0.1254 |
| 13 | 0.0104 | 0.3618 | –0.0027 | 0.7905 | 0 | — | 0 | 0.0059 |
| 14 | 0.2055 | 0.1426 | –0.0460 | 0.7872 | 0 | — | 0 | 0.0998 |
| 15 | 0.2937 | 0.1769 | 0.4242 | 0.4738 | 0 | — | 0 | 0.1212 |
| 16 | –0.1687 | 0.4658 | –0.4544 | 0.3897 | 0 | — | 0 | 0.1275 |
| 17 | –0.0129 | 0.7944 | 0.0664 | 0.7836 | 0 | — | 0 | 0.0923 |
| 18 | –0.1279 | 0.6857 | 0.6707 | 0.3892 | 0 | — | 0 | 0.1671 |
| 19 | 0.3541 | 0.1368 | 0.4746 | 0.4704 | 0 | — | 0 | 0.1344 |
| 20 | 1.2608 | 0.0936 | –0.6072 | 0.6735 | 0 | — | 0 | 0.3202 |

Table 2: Parameter estimates for the credit data with corresponding p–values. The middle columns are based on the S–Plus function `step.glm()` and the learning sample and the two right columns correspond to the GSoft, applied to the learning sample

|  | Learning–sample |  | Validation |  |
|---|---|---|---|---|
|  | $n_{miss}/n$ | error rate | $n_{miss}/n$ | error rate |
| $y = 1$ | 53/144 | 36.81 % | 219/556 | 39.39 % |
| $y = 0$ | 14/56 | 25.00 % | 75/244 | 30.74 % |
| Average | 30.91 % |  | 35.07 % |  |

Table 3: Missclassification based on variable selection in a logit model

|  | $\lambda^{CV_D} = 0.91$ |  |
|---|---|---|
|  | Learning sample | Validation |
|  | $n_{miss}/n$ | $n_{miss}/n$ |
|  | rate | rate |
| $y = 1$ | 40/144 | 186/556 |
|  | 27.78 % | 33.45 % |
| $y = 0$ | 13/56 | 63/244 |
|  | 23.21 % | 25.82 % |

Table 4: Missclassification based on GSoft and a linear logit model.

|  | $\lambda^{CV_D} = 0.41$ |  | $\lambda^{CV_M} = 0.21$ |  | LDA, $p = 15$ |  |
|---|---|---|---|---|---|---|
|  | $n_{miss}/n$ | $CV_M(\lambda)$ | $n_{miss}/n$ | $CV_M(\lambda)$ |  |  |
| $y = 1$ | 27.57 % | 28.29 % | 26.57 % | 27.29 % | 27.43 % | 28.0% |
| $y = 0$ | 23.67 % | 25.33 % | 23.67 % | 24.67 % | 27.00 % | 27.6 % |
| Average | 25.62 % | 26.81 % | 25.12 % | 25.98 % | 27.2% | 27.8 % |

Table 5: Missclassification rates of GSoft and a linear discriminant analysis, based on all 1000 samples.

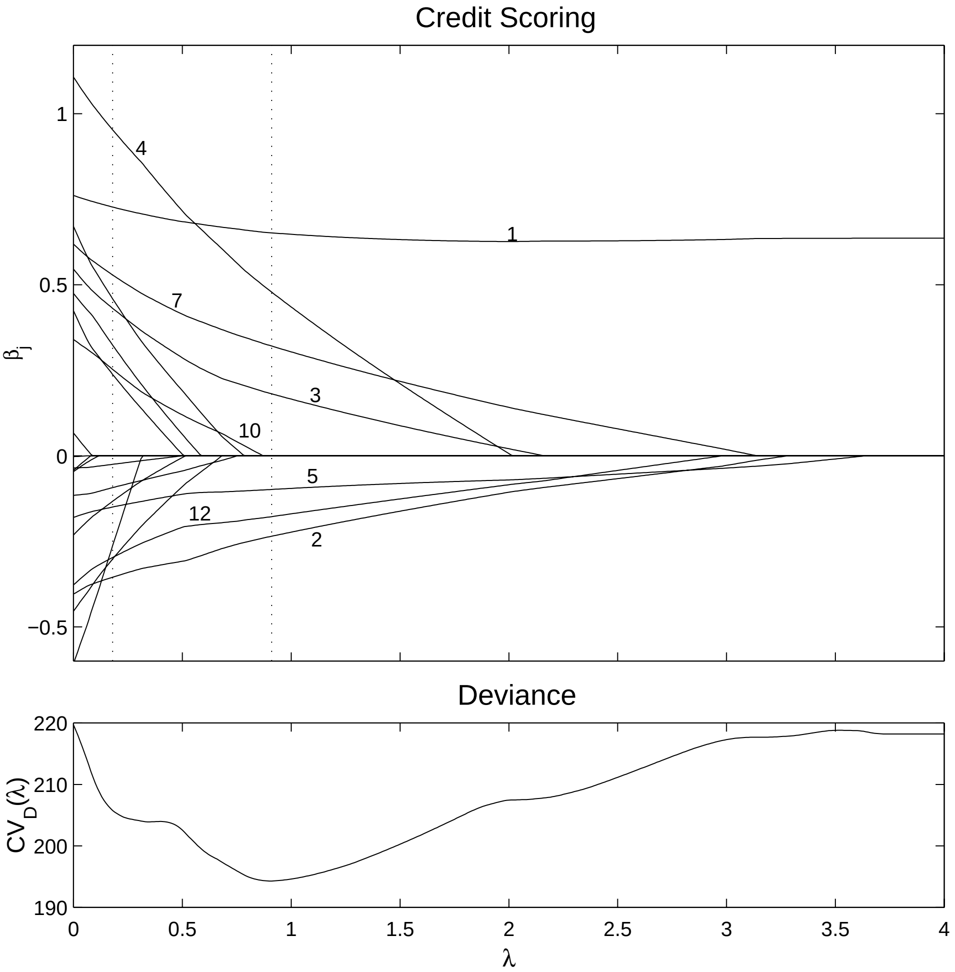

Figure 6: GSoft for the credit data based on a linear logistic regression. In the upper panel $\hat{\beta}_j$ are plotted as a function in $\lambda$. The lower panel shows the cross–validated deviance.

| Indicator | number of flats | description |
|---|---|---|
| $S^+$ | 653 | indicator of location ("site") above average |
| $S^-$ | 173 | indicator of location ("site") below average |
| $A^+$ | 28 | indicator for top location |
| $A^-$ | 43 | indicator for bad location |
| $Bd^-$ | 44 | no bathroom indicator |
| $Zh^-$ | 389 | no central heating indicator |
| $W^-$ | 125 | no central hot water indicator |
| $Bl^+$ | 250 | indicator for a bigger balcony |
| $F^+$ | 49 | indicator for nicely shaped windows |
| $F^-$ | 122 | indicator for isolated glass in windows |
| $Z^+$ | 203 | indicator for general special equipment |
| $Bd^+$ | 1135 | indicator of bathroom equipment above average |
| $K^+$ | 173 | indicator of kitchen equipment above average |
| $G^+$ | 414 | indicator of floor plan above average |
| $R^+$ | 167 | indicator of fundamental renovation |
| $Ab^-$ | 258 | indicator for a simple, old building |
| $H^-$ | 29 | indicator for an old building in the backyard |

Table 6: Indicator variables used in the munich rental guide. The second column indicates the number of flats.

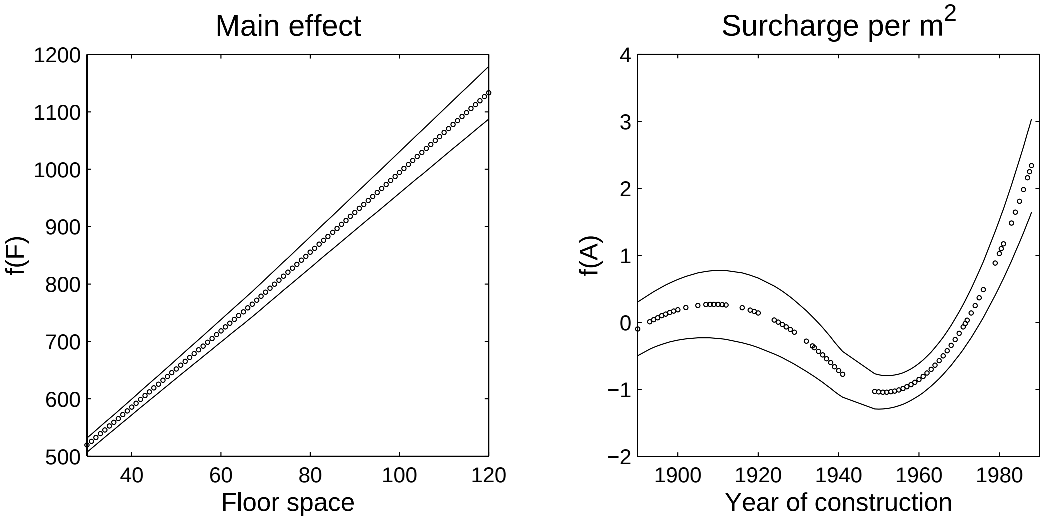

Figure 7: Centered main effects together with pointwise $2\sigma$ bands

| Indicator | m$^2$ | cost per additional $m^2$ | age | $f(A)$ |
|---|---|---|---|---|
| $S^+$ | 0–42 | 0.347 (0.150) | 1958–1971 | 0.422 (0.100) |
| | 43–120 | 1.988 (0.224) | 1972–1989 | 0.037 (0.117) |
| $S^-$ | 0–120 | -1.047 (0.142) | 1928–1938 | -0.037 (0.090) |
| | | | 1939–1989 | -0.927 (0.186) |
| $A^+$ | 0–59 | 1.629 (0.368) | 1958–1989 | 4.063 (0.541) |
| | 60–120 | 1.007 (0.643) | | |
| $A^-$ | 0–50 | 0.312 (0.296) | 1958–1971 | -0.300 (0.112) |
| | 51–120 | 2.862 (0.511) | 1972–1980 | -3.502 (0.417) |
| | | | 1981–1989 | -4.562 (0.278) |
| $Bd^-$ | 0–53 | -1.650 (0.232) | | |
| | 54–120 | -0.958 (0.456) | | |
| $Zh^-$ | 0–41 | -2.710 (0.186) | 1890–1899 | 0.379 (0.110) |
| | 42–89 | -1.574 (0.231) | 1938–1955 | 0.134 (0.121) |
| | 90–120 | -4.470 (0.660) | 1956–1963 | 0.113 (0.141) |
| | | | 1964–1989 | 0.885 (0.240) |
| $W^-$ | 0–42 | -1.883 (0.208) | | |
| | 43–120 | 0.246 (0.314) | | |
| $Bl^+$ | 0–55 | 1.266 (0.172) | 1970–1971 | -0.815 (0.181) |
| | 56–87 | -1.880 (0.501) | 1972–1980 | -0.752 (0.174) |
| | 88–89 | 3.872 (0.967) | 1981–1989 | 0.031 (0.219) |
| | 90–120 | 4.983 (1.096) | | |
| $F^+$ | 0–64 | 2.873 (0.555) | | |
| | 65–120 | -1.702 (0.911) | | |
| $F^-$ | 0–45 | -0.752 (0.317) | 1900–1937 | -0.220 (0.131) |
| | 46–120 | 1.497 (0.437) | 1938–1971 | -2.160 (0.289) |
| | | | 1972–1989 | -0.578 (0.307) |

Table 7: Additional effects for indicators based on the rent per square meter, with corresponding standard deviation in brackets.

| Indicator | m$^2$ | cost per additional $m^2$ | age | $f(A)$ |
|---|---|---:|---|---:|
| $Z^+$ | 0–52 | 2.027 (0.194) | 1919–1967 | -0.005 (0.116) |
| | 53–120 | 0.712 (0.306) | 1968–1989 | -0.895 (0.202) |
| $Bd^+$ | 0–36 | 0.041 (0.155) | 1934–1972 | 0.689 (0.144) |
| | 37–120 | 0.190 (0.172) | 1973–1989 | 1.660 (0.213) |
| $K^+$ | 0–120 | 2.840 (0.273) | 1938–1955 | -0.002 (0.127) |
| | | | 1956–1971 | -0.994 (0.279) |
| | | | 1972–1980 | -1.190 (0.293) |
| | | | 1980–1985 | -1.893 (0.321) |
| | | | 1986–1989 | -3.011 (0.467) |
| $G^+$ | 0–48 | 0.867 (0.174) | 1934–1989 | 0.623 (0.152) |
| | 49–120 | 0.832 (0.214) | | |
| $R^+$ | 0–120 | 2.098 (0.199) | 1919–1989 | -1.266 (0.234) |
| $Ab^-$ | 0–73 | -1.274 (0.126) | 1934–1949 | 0.427 (0.130) |
| | 74–120 | -0.451 (0.285) | | |
| $H^-$ | 0–120 | -1.537 (0.232) | | |

Table 8: Additional effects for indicators based on the rent per square meter, with corresponding standard deviation in brackets.