LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Thomas Kneib, Susanne Konrath, Ludwig Fahrmeir

High-dimensional Structured Additive Regression Models: Bayesian Regularisation, Smoothing and Predictive Performance

Technical Report Number 46, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

High-dimensional Structured Additive Regression Models: Bayesian Regularisation, Smoothing and Predictive Performance

Thomas Kneib, Susanne Konrath, Ludwig Fahrmeir

Department of Statistics

Ludwig-Maximilians-University Munich

Abstract

Data structures in modern applications frequently combine the necessity of flexible regression techniques such as nonlinear and spatial effects with high-dimensional covariate vectors. While estimation of the former is typically achieved by supplementing the likelihood with a suitable smoothness penalty, the latter are usually assigned shrinkage penalties that enforce sparse models. In this paper, we consider a Bayesian unifying perspective, where conditionally Gaussian priors can be assigned to all types of regression effects. Suitable hyperprior assumptions on the variances of the Gaussian distributions then induce the desired smoothness or sparseness properties. As a major advantage, general Markov chain Monte Carlo simulation algorithms can be developed that allow for the joint estimation of smooth and spatial effects and regularised coefficient vectors. Two applications demonstrate the usefulness of the proposed procedure: A geoadditive regression model for data from the Munich rental guide and an additive probit model for the prediction of consumer credit defaults. In both cases, high-dimensional vectors of categorical covariates will be included in the regression models. The predictive ability of the resulting high-dimensional structure additive regression models compared to expert models will be of particular relevance and will be evaluated on cross-validation test data.

Key words: Bayesian Lasso; Laplace prior; Markov random fields; MCMC; penalised splines; Ridge regression; scale mixtures.

1 Introduction

Regression models with high-dimensional vectors of regression coefficients have gained considerable attention within recent years. Much methodological development has been stimulated through biological applications where the high-dimensional covariate vector represents for example gene expression measurements (Bae & Mallick 2004, Goeman 2007, Griffin & Brown 2007). In such applications, the number of regression coefficients may well exceed the sample size, but the predictor is still linear because the effects of gene expressions are assumed to be linear, at least after suitable transformation and normalisation. Regularisation in the resulting high-dimensional linear or generalised linear models is typically based on shrinkage estimators, keeping the parameters identifiable and at the same time aiming at variable selection.

This paper is motivated through a number of consulting cases and applications in social science, economics and public health with the following common feature: A response variable of interest shall be related to a combination of high-dimensional covariate vectors with usual linear effects, several continuous covariates with nonlinear effects of unknown functional form, and possibly additional nonlinear effects of spatial or geographical location. Thus, there is a need for flexible regression models, incorporating these different types of covariates and effects in a structured additive predictor, and combining regularisation with semiparametric smoothing. We will exemplarily focus on analysing regression data from the Munich rental guide and from credit scoring. Other applications with similar data structures include the analysis of childhood undernutrition, morbidity and mortality in developing countries using Demographic and Health Survey (DHS) data (see for example Kandala et al. (2001), Kandala et al. (2007) or Adebayo & Fahrmeir (2005)) or the analysis of claim size and claim frequency in non-life insurance (Fahrmeir et al. 2007).

Previous research has mostly concentrated either on penalisation approaches for semiparametric smoothing (Eilers & Marx 1996, Ruppert, Wand & Carroll 2003, Fahrmeir, Kneib & Lang 2004, Wood 2006) or on regularizing high-dimensional but exclusively linear

predictors. Although regularisation approaches such as Ridge or Lasso regression actually fit in this framework since they also rely on penalised estimation, only very limited work has been done for combining both approaches. A notable exception is Vandenhende et al. (2007) where penalised spline smoothing is supplemented with Ridge regression. Both yield quadratic penalties, allowing to apply penalised least-squares estimation or penalised Fisher scoring.

We propose high-dimensional Bayesian structured additive regression models providing both flexibility in smoothing nonlinear or spatial effects and the possibility to regularise high-dimensional covariate effects. As a unifying concept, conditionally Gaussian priors are assigned to all types of covariate effects. Suitable hyperpriors on the variances of the Gaussian distributions then induce non-Gaussian marginal priors, enforcing the desired smoothness or sparseness properties. Semiparametric smoothing is based on Bayesian penalised splines with Gaussian smoothness priors, corresponding to difference penalties for basis function coefficients, and on Gaussian Markov random field priors for spatial effects as in Fahrmeir, Kneib & Lang (2004), Lang & Brezger (2004) or Brezger & Lang (2006). Regularisation priors can be expressed as scale mixtures of Gaussian distributions, including the Bayesian Lasso (Park & Casella 2008), the broader class of shrinkage priors proposed in Griffin & Brown (2005) and Griffin & Brown (2007), and the spike and slab prior of Ishwaran & Rao (2007), all of them suggested in the context of high-dimensional Gaussian linear models.

The unifying framework of conditionally Gaussian priors facilitates full Bayesian Markov chain Monte Carlo (MCMC) inference from a conceptual and implementational perspective. For Gaussian regression models or models that can be related to latent Gaussian models, such as the probit model, full conditionals for all regression coefficients are Gaussian leading to Gibbs samplers as in Section 2. For non-Gaussian exponential family regression models, block updates for regression coefficients are obtained from Metropolis Hastings steps with Gaussian 'iteratively weighted least squares (IWLS)' proposals as outlined in Section 6.

In many of the consulting cases and applications mentioned in the beginning, predictive performance is at least as important as build-in variable selection properties. We investigate predictive performance exemplarily in two applications where prediction is of high relevance.

Our first application (Section 4) is the Munich rental guide, an ongoing consulting case for the City of Munich. The response variable of interest is the net rent per square meter that shall be related to several hundreds of covariates characterising flats and properties of the neighborhood. The ultimate goal of rental guides is to predict rents for flats with a given covariate combination, providing a reference point both for landlords and tenants. To achieve this goal, the usual strategy of selecting a suitable subset of covariates may be deemed to be suboptimal since information may be lost when ignoring or combining specific covariates. While the large number of observations collected in the Munich rental guide (with a sample size of approximately $n = 3,000$) renders even hundreds of parameters identifiable, a large estimation uncertainty will be associated with estimates that are only weakly identified. Examples may be categorical covariates where some of the categories are only rarely observed, covariates that yield quasi-complete separation, or covariates that are highly correlated. We will therefore supplement the high-dimensional vector of covariates with regularisation priors that mimic Ridge and Lasso regression. In addition, some of the covariate effects included in the data are well-known to follow nonlinear patterns. Examples are the year of construction or the size of the flat. From previous studies it seems difficult to approximate their effects in terms of, for example, polynomials, hinting at the necessity to consider flexible nonlinear effects. Moreover, spatial information on the location of the flats within Munich is available and may provide important information in addition to the available covariates.

In our second example, we will predict failures of consumer credits in a binary regression model from high-dimensional vectors of covariates characterising the credit holder. Again the major aim is prediction and it will turn out to be useful to include all covariates in a regularised fashion. While no spatial data is available in this case, the effects of

age of the credit holder, credit amount and duration of the contract are suspected to be of nonlinear form, yielding the necessity to combine regularisation of high-dimensional regression coefficients with estimation of semiparametric regression models.

2 High-Dimensional Structured Additive Regression

2.1 Observation Model

This section considers Gaussian models $y_i|\eta_i \sim \mathrm{N}(\eta_i, \sigma^2)$ and binary probit models $P(y_i = 1|\eta_i) = \Phi(\eta_i)$ (where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function) with conditionally independent responses y_i and (geo-) additive predictors

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma} + f_1(z_{i1}) + \ldots + f_p(z_{ip}) + f_{\text{spat}}(s_i) \tag{1}$$

where $\boldsymbol{x}$ represents a high-dimensional vector of covariates with regression coefficients $\boldsymbol{\beta}$ that should be subject to regularisation. The additional vector of covariates $\boldsymbol{u}$ with linear effects $\boldsymbol{\gamma}$ contains variables whose effects should be estimated unrestrictedly. This may, for example, be relevant in biostatistical applications where $\boldsymbol{x}$ represents high-dimensional gene expression data while $\boldsymbol{u}$ contains a limited number of clinical covariates. Since the latter are deemed to be more reliable and represent the benchmark information for the model including gene expression data, their effects should not be subject to regularisation (Binder & Schumacher 2008). Nonlinear effects of continuous covariates $z_1, \ldots, z_p$ are included in terms of the functions $f_1, \ldots, f_p$. Each function is approximated through a basis function expansion $f_l(z_l) = \boldsymbol{b}_l(z_l)'\boldsymbol{\alpha}_l$ with a vector of basis function evaluations $\boldsymbol{b}_l(z_l) = (B_1(z_l), \ldots, B_K(z_l))'$ and a vector of basis coefficients $\boldsymbol{\alpha}_l$. We will focus on a B-spline basis and on Bayesian P-splines (Lang & Brezger 2004, Brezger & Lang 2006), but our approach is general and other bases may be employed as well. Spatial information, if included in the regression model, is represented in terms of geographical regions $s \in \{1, \ldots, S\}$ the observations pertain to, leading to a spatial function $f_{\text{spat}}(s)$. The spatial function is expressed in terms of parameters $\alpha_{s,\text{spat}} = f_{\text{spat}}(s)$ and spatial adjacency of the

regions will be taken into account in their prior distribution. Extensions to other types of responses and regression models are outlined in the discussion in Section 6 (see also Fahrmeir, Kneib & Lang (2004)).

For the description of Bayesian inference, it will be helpful to represent model (1) in unified vector-matrix representation as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{Z}_1\boldsymbol{\alpha}_1 + \ldots + \boldsymbol{Z}_p\boldsymbol{\alpha}_p + \boldsymbol{Z}_{\text{spat}}\boldsymbol{\alpha}_{\text{spat}}.$$

where $\boldsymbol{X}$, $\boldsymbol{U}$, $\boldsymbol{Z}_1$, $\ldots$, $\boldsymbol{Z}_p$ are design matrices constructed from covariates and basis functions. To unify notation, we define $\boldsymbol{\alpha}_{\text{spat}} = (\ldots, \alpha_{s,\text{spat}}, \ldots)'$ for the spatial effects and $\boldsymbol{Z}_{\text{spat}}$ is a zero/one incidence matrix linking observations i to the corresponding regions $s_i \in \{1, \ldots, S\}$.

2.2 Bayesian Regularisation and Smoothness Priors

For the vector $\boldsymbol{\gamma}$ of unregularised regression coefficients, we assume a flat prior $p(\boldsymbol{\gamma}) \propto 1$ or a weakly informative Gaussian prior $\boldsymbol{\gamma} \sim \text{N}(\boldsymbol{0}, c\boldsymbol{I})$, c large. Regularisation priors for $\boldsymbol{\beta}$ and smoothness priors for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p, \boldsymbol{\alpha}_{\text{spat}}$ have a common structure: Conditionally upon variances, they are multivariate Gaussian. This facilitates unified formulation and implementation of posterior inference via Gibbs sampling. However, different types of hyperpriors for variances are considered, leading to varying properties of marginal regularisation and smoothness priors.

Regularisation priors and analogue frequentist shrinkage penalties have often been developed primarily for variable selection purposes. A desirable feature for variable selection is to shrink small effects to zero but to shrink important effects only moderately to prevent them from large bias. Our empirical experience suggests that improved variable selection performance does not necessarily improve prediction or classification quality, in particular when taking computational effort into account. Therefore, we restrict presentation here to Bayesian versions of the Ridge and Lasso penalties and mention competing regularisation priors with good variable selection properties as discussed for example in Griffin & Brown

(2005) and Griffin & Brown (2007), Ishwaran & Rao (2007) for Gaussian models, and in Konrath, Kneib & Fahrmeir (2008) for hazard rate models in Section 6.

Bayesian Ridge. The Bayesian version of the Ridge penalty is obtained when assuming (conditional) Gaussian i.i.d. priors

$$\beta_j|\tau_j^2 \overset{\text{i.i.d.}}{\sim} \text{N}(0, \tau_j^2), \quad j = 1, \ldots, q,$$

for the components of $\boldsymbol{\beta}$ together with inverse Gamma hyperpriors $\tau_j^2 \overset{\text{i.i.d.}}{\sim} \text{IG}(a, b)$ for variances (inverse shrinkage parameters). The marginal prior for regression coefficients is a scaled t-distribution with $2a$ degrees of freedom and scale parameter $(b/a)^{0.5}$, see Figure 1 for its log-density.

Bayesian Lasso. From a Bayesian perspective, the Lasso penalty with shrinkage parameter λ (Tibshirani 1996) is the log-prior derived from a Laplace distribution $\beta_j|\lambda \overset{\text{i.i.d.}}{\sim} \text{Laplace}(0, \lambda)$, compare e.g. Park & Casella (2008). Expressing the Laplace density as a scale mixture of normals with an exponential mixing distribution, and specifying a gamma prior for λ^2 leads to the following hierarchical formulation of the Bayesian Lasso:

$$\beta_j|\tau_j^2 \overset{\text{i.i.d.}}{\sim} \text{N}(0, \tau_j^2), \quad \tau_j^2|\lambda \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda^2/2), \quad \lambda^2 \sim \text{Ga}(a, b).$$

Conditionally upon τ_j^2, the prior for β_j is Gaussian, but the marginal prior is non-Gaussian and can be expressed as

$$p(\beta_j|a, b) = \frac{a2^a}{\sqrt{\pi 2b}}\Gamma\left(a + \frac{1}{2}\right) \exp\left(\frac{\beta_j^2}{8b}\right) D_{-2a-1}\left(\sqrt{\frac{\beta_j^2}{2b}}\right). \tag{2}$$

where D_{-2a-1} is the parabolic cylinder function (Griffin & Brown 2005, Griffin & Brown 2007). Comparing the marginal Bayesian Lasso log-prior in Figure 1 with the marginal Ridge log-prior as well as with the frequentist Lasso and Ridge penalties, we see that Bayesian Lasso and Ridge are not so far from each other as in their frequentist versions although the Lasso log-prior still has somewhat heavier tails. Even more importantly, the Bayesian Ridge has much more probability mass centered about zero, with a distinct peak

at zero, compared to the frequentist Ridge (and Lasso) penalty. Therefore, the shrinkage of large coefficients towards zero is only moderate while shrinkage of small coefficients towards zero is encouraged.

Smoothness priors for functions. Priors for function parameters $\boldsymbol{\alpha}_l$, $l = 1, \ldots, p, \text{spat}$ have the same general form. Given the variance (or inverse smoothing) parameter δ_l^2, the priors are conditionally Gaussian

$$p(\boldsymbol{\alpha}_l|\delta_l^2) \propto \left(\frac{1}{\delta_l^2}\right)^{\frac{\text{rank}(\boldsymbol{K}_l)}{2}} \exp\left(-\frac{1}{2\delta_l^2}\boldsymbol{\alpha}_l'\boldsymbol{K}_l\boldsymbol{\alpha}_l\right). \tag{3}$$

The specific form of the penalty matrix $\boldsymbol{K}_l$ depends on the specific assumption made for the function f_l.

For a function $f_l(z_l)$ of a continuous covariate z_l, our preferred option are Bayesian P-splines (Lang & Brezger 2004, Brezger & Lang 2006). In this case, the precision matrix is given as $\boldsymbol{K}_l = \boldsymbol{D}_d'\boldsymbol{D}_d$, where $\boldsymbol{D}_d$ is a matrix of first ($d = 1$) or second ($d = 2$) order differences, and $\boldsymbol{K}_l$ is partially improper with $\text{rank}(\boldsymbol{K}_l) = K - d$. Priors for other basis function representations such as the truncated power basis for regression splines or smoothing spline bases have the same general form.

In our application to the Munich rent data, we will choose a standard Markov random field prior

$$\alpha_{\text{spat},s}|\alpha_{\text{spat},r},\, r \neq s, \delta_{\text{spat}}^2 \sim \text{N}\left(\frac{1}{|N_s|}\sum_{r\in N_s}\alpha_{\text{spat},r}, \frac{\delta_{\text{spat}}^2}{|N_s|}\right)$$

for the spatial effects, where N_s denotes the set of neighbors to region s. The resulting penalty matrix $\boldsymbol{K}_{\text{spat}}$ is an adjacency matrix with entries $|N_s|$ on the diagonal and -1 on the off-diagonals whenever two regions are neighbors. As a consequence, $\text{rank}(\boldsymbol{K}_{\text{spat}}) = S - 1$, i.e. the spatial prior is partially improper again. Other options for modelling spatial effects like proper Gaussian conditional autoregressive (CAR) models and stationary Gaussian random field (Kriging) models are proper, but with the same general form, see Rue & Held (2005) or Banerjee, Carlin & Gelfand (2003).

In all cases, we assume inverse Gamma priors

$$\delta_l^2 \sim \text{IG}(a_l, b_l)$$

with small values $a_l = b_l = \varepsilon > 0$ as a standard option. Note however, that a flat prior $p(\delta_l) \propto 1$, recommended by Gelman (2006), can be written as an improper IG-prior as well with $a_l = -0.5, b_l = 0$.

Summarizing, we recognise the common structure of Bayesian regularisation and smoothness priors: Given the variance parameters τ_j^2, $j = 1, \ldots, q$, in regularisation priors and δ_l^2,, $l = 1, \ldots, p,$spat in smoothness priors, the corresponding prior distributions for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j, \ldots, \beta_q)'$ and $\boldsymbol{\alpha}_l$ are conditionally Gaussian $\text{N}(\mathbf{0}, \boldsymbol{P}_{\boldsymbol{\beta}}^{-1})$ and $\text{N}(\mathbf{0}, \boldsymbol{P}_l^{-1})$, with precision matrices

$$\boldsymbol{P}_{\boldsymbol{\beta}} = \text{diag}(1/\tau_1^2, \ldots, 1/\tau_q^2), \quad \boldsymbol{P}_l = \boldsymbol{K}_l/\delta_l^2.$$

Formally, flat priors $p(\boldsymbol{\gamma}) \propto 1$ for unregularised regression parameters can be expressed in the same form with $\boldsymbol{P}_{\boldsymbol{\gamma}} = \mathbf{0}$.

The difference in *marginal* priors comes from different assumptions on variance parameters at further stages of the hierarchical priors. As a particular advantage, this leads to unified Gibbs or MCMC schemes for posterior inference.

Finally, we assume an inverse Gamma prior $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$ for the variance in Gaussian observation models $y_i \sim \text{N}(\eta_i, \sigma^2)$.

3 Posterior Inference

For Gaussian responses, full conditionals for the vectors β and γ of regularised and unregularised regression parameters as well for function parameters are Gaussian:

- $\boldsymbol{\beta}|\cdot \sim \text{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \frac{1}{\sigma^2} \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\eta}_{-\boldsymbol{\beta}}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\frac{1}{\sigma^2} \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{P}_{\boldsymbol{\beta}} \right)^{-1}$$

  and $\boldsymbol{\eta}_{-\boldsymbol{\beta}} = \boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta}$.

- $\boldsymbol{\gamma}|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ where

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \frac{1}{\sigma^2} \boldsymbol{U}'(\boldsymbol{y} - \boldsymbol{\eta}_{-\boldsymbol{\gamma}}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \sigma^2 (\boldsymbol{U}'\boldsymbol{U})^{-1}$$

  and $\boldsymbol{\eta}_{-\boldsymbol{\gamma}} = \boldsymbol{\eta} - \boldsymbol{U}\boldsymbol{\gamma}$.

- $\boldsymbol{\alpha}_l|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, $l = 1, \ldots, p, \text{spat}$ where

$$\boldsymbol{\mu}_l = \boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \frac{1}{\sigma^2} \boldsymbol{Z}_l'(\boldsymbol{y} - \boldsymbol{\eta}_{-l}), \quad \boldsymbol{\Sigma}_l = \left(\frac{1}{\sigma^2} \boldsymbol{Z}_l'\boldsymbol{Z}_l + \boldsymbol{P}_l \right)^{-1}.$$

  and $\boldsymbol{\eta}_{-l} = \boldsymbol{\eta} - \boldsymbol{Z}_l\boldsymbol{\alpha}_l$.

For variance and shrinkage parameters, full conditionals are as follows

- Bayesian Ridge:

$$\tau_j^2|\cdot \sim \mathrm{IG}\left(a + \frac{q}{2}, b + \frac{1}{2}\beta_j^2 \right).$$

- Bayesian Lasso:

$$\frac{1}{\tau_j^2}\bigg| \cdot \sim \mathrm{InvGauss}\left(\frac{|\lambda|}{|\beta_j|}, \lambda^2 \right), \quad \lambda^2|\cdot \sim \mathrm{Ga}\left(a + q, b + \frac{1}{2}\sum_{j=1}^{q} \tau_j^2 \right).$$

- Smoothing variances:

$$\delta_l^2|\cdot \sim \mathrm{IG}\left(a_l + \frac{\mathrm{rank}(\boldsymbol{K}_l)}{2}, b_l + \frac{1}{2}\boldsymbol{\alpha}_l\boldsymbol{K}_l\boldsymbol{\alpha}_l \right).$$

- Error variance:

$$\sigma^2|\cdot \sim \mathrm{IG}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(y_i - \eta_i)^2 \right)$$

Drawing samples iteratively from these full conditionals results in a Gibbs sampler, without any further need for tuning. Still, some of the updating steps require additional comments. For the high-dimensional vector $\boldsymbol{\beta}$, it will typically be necessary to apply a block-update since a complete update will lead to very long computation times due to the high-dimensional matrix computations involved. We consider blocks of 20 parameters per default. For the smooth and spatial effects, one can take advantage of sparse matrix structures since the precision matrix of the full conditional has band structure (P-splines)

or is at least close to a band matrix (spatial effects). Employing sparse matrix computations (such as the sparse matrix Cholesky decomposition) as described in Rue & Held (2005), the simulation of the high-dimensional Gaussian effects becomes feasible even in case of large spatial effects as in the Munich rental guide example with approximately 400 regions.

For binary probit models $y_i \sim \mathrm{B}(1, \pi_i)$, $\pi_i = \Phi(\eta_i)$, a Gibbs sampler can be constructed using auxiliary Gaussian responses $\tilde{y}_i \sim \mathrm{N}(\eta_i, 1)$ related to y_i via the threshold relation

$$y_i = 1 \iff \tilde{y}_i > 0,$$

following Albert & Chib (1993) for linear predictors and Fahrmeir & Lang (2001b) for categorical response models with geoadditive predictors. This requires additional samples from truncated normals that are arising as the resulting full conditionals for the auxiliary variables $\tilde{y}_i$. The remaining steps for drawing from full conditionals for regression and variance parameters are essentially the same as for Gaussian models, replacing y_i, through $\tilde{y}_i$ and omitting sampling for σ^2.

4 Munich Rental Guide

According to German law, increases in rents for flats can be justified based on "average rents" paid for flats that are comparable in size, equipment, quality and location. As a consequence, most larger cities publish rental guides that provide such average rents, obtained from regression models with net rents or net rents per square meter as dependent variables and characteristics of the flat as explanatory variables. For short, the aim of rental guides is to predict net rents based on a potentially large set of covariates.

In the following, we will analyse data on approximately 3,000 flats in the City of Munich collected by Infratest Sozialforschung for the 2007 rental guide (see `http://www.muenchen.de/mietspiegel` for the official documentation of the 2007 rental guide). The original data (that are available for research purposes upon request) contain approxi-

mately 250 covariates describing characteristics of the flats as diverse as the quality of bathroom equipment, whether the flat is rented for the first time, the presence of a garden or a balcony, etc. While most of the covariates are categorical, there are some continuous covariates of designated interest. These are the size of the flat and the age of the building in which the flat is located. Both are well-known from previous analyses to have nonlinear impact on the net rent. Moreover, a spatial variable is available that represents the subquarter of Munich where the flat is located. Although a variable representing the expert assessment of the location (with the three categories average, high, highest) is available, it has been shown that a more detailed treatment of the spatial effect usually leads to better predictions of the net rent.

Based on the previous considerations, we will predict net rents per square meter from a high-dimensional geoadditive regression model

$$netrent_i = \gamma_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + f_1(size_i) + f_2(year_i) + f_{\text{spat}}(s_i) + \varepsilon_i$$

where γ_0 corresponds to an unpenalised intercept, $\boldsymbol{x}_i$ represents a 238-dimensional vector of mostly categorical covariates with effects summarised in the vector $\boldsymbol{\beta}$ that is assigned an appropriate shrinkage prior. To be more specific, we will consider Bayesian Ridge and Bayesian Lasso priors. The functions f_1 and f_2 are smooth functions of the continuous covariates size of the flat (*size*, in square meters) and year of construction (*year*, in years). Both functions are modelled as Bayesian cubic P-splines with twenty inner knots and second order random walk prior. The spatial effect is assigned a Markov random field prior based on the discrete spatial information on the 411 subquartes of Munich.

To contrast the high-dimensional geoadditive model, we considered the expert model used in the 2007 Munich rental guide (and applied to a data set from previous years in Fahrmeir & Lang (2001a)) as a benchmark for regularised estimation with Ridge or Lasso priors. In the expert model, the number of influential variables is reduced by combining and deleting specific variables in a combination of knowledge from previous rental guides and stepwise variable selection methods. This yields a model specification with 28 explanatory variables

that are assigned flat, noninformative priors. Size of the flat, year of construction and location are treated exactly as in the regularised model, i.e. they are modelled flexibly with penalised spline and Markov random field priors.

Figures 2 and 3 visualise the estimated nonlinear and spatial effects for the three models considered. While the effect of size is rather independent of the model specification, some differences can be observed for the year of construction. Both regularised approaches yield very comparable results while the curve for the expert model shows a higher increase for very old flats and a smaller increase for very new flats. It is also interesting to note that the width of the 95% pointwise confidence intervals does not increase when moving from the expert model with a moderate number of regressors to the high-dimensional models. Obviously, the regularisation priors introduce enough penalisation to yield an effective parameter count that is much smaller than the actual number of regression coefficients. For the spatial effect, point estimates are again close to each other for all three models. Surprisingly, it is possible to identify an even clearer spatial pattern in this case, with more pointwise significant effects when considering the high-dimensional geoadditive models.

When interpreting the estimated nonlinear effects, all three models show the expected results with high net rents per square meter for smaller flats and only very little variation for flats of size larger than about 50 square meters. While the regularised high-dimensional models identify no effect for old buildings since the point estimate is very close to zero, the expert model identifies a positive effect (though zero is mostly included in the confidence bands) for old buildings. All three models pick up the expected decrease in net rents for buildings constructed during the 1950s to 1970s, while rents start to increase again for newly constructed buildings. The spatial effect identifies increased rents for flats in the center of Munich, although several of the categorical covariates already try to account for the higher life quality in the city center and some expensive subquarters in the southern part of Munich. In contrast, significantly negative effects are found for some subquarters at the border of Munich.

Table 1 shows estimated regularised effects for the subset of covariates with associated

posterior probabilities smaller than 0.05 in either the Bayesian Lasso or the Bayesian Ridge model. While the posterior probabilities are in good agreement between the two regularisation approaches, the estimated effects show quite some differences. In particular, the absolute value of the estimated effects is mostly larger for the Lasso whenever differences of a certain magnitude are observed. This confirms the intuition that the Lasso induces less penalisation for larger effects due to the heavier tails of the corresponding prior density. It turns out that the dummy for company housing is a variable with very high impact, leading to a reduction of about one Euro for the net rent per square meter. This of course perfectly makes sense since companies will typically rent flats to their employees for a lower charge. Several other covariates included in the list in Table 1 correspond to specific equipment of the flat such as different types of floor material or the availability of an intercom system.

In summary, the estimated nonlinear and spatial effects are not too different whether a high-dimensional model is employed or not. Still there are of course differences in the parametric part of the predictor and our original motivation was prediction-oriented, anyway. We therefore split the data into ten mutually exclusive data sets and performed a ten-fold cross-validation to assess the predictive ability of the three models. As a measure for predictive quality we considered the mean-squared error of prediction, computed based on the out of bag sample (i.e. the 10% of the data left out of the model fitting process). Figure 4 shows a parallel coordinate plot of the mean squared error of prediction, where each line corresponds to a specific data configuration. By connecting results for the same data it is possible to directly compare the results obtained with the three models for a specific data set. For all except one data set, we identify a strong gain in predictive ability by introducing the high-dimensional covariate vector with any type of regularisation. Furthermore, there seems to be a very minor preference for the Lasso but differences are actually very small.

5 Credit Scoring

In the second example, we are analysing data on the defaults of consumer credits well known under the heading "German credit data" in the classification community (see for example Michie et al. (1994) or Holmes & Adams (2002)) and available from the data archive of the Department of Statistics at the Ludwig-Maximilians-University Munich (`http://www.stat.uni-muenchen.de/service/datenarchiv/`). The data set consists of information on 1000 consumer credits from a German bank and the major aim is to separate credit-worthy and not credit-worthy clients based on covariates. Hence, the response variable of interest is a binary indicator y_i that specifies whether the credit has been paid back ($y_i = 1$, credit-worthy) or not ($y_i = 0$, not credit-worthy). A total of 54 covariates is available to characterise the client, including information on age of the client, credit amount and duration of the credit. These three covariates are continuous and will be inspected for nonlinear effects in the following. The remaining covariates are categorical and will be assigned Ridge and Lasso priors for regularisation, leading to the binary additive probit model

$$P(y_i = 1) = \Phi(\gamma_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + f_1(age) + f_2(amount) + f_3(duration)).$$

Again the only effect that is not regularised is the overall intercept γ_0.

The results from the regularised estimation schemes are compared to results from a literature model that only contains the nonlinear effects and information on three further categorical covariates (payment of previous credits with categories good and bad; intended use with categories professional and private; running account with categories no, medium and good). The comparison is based on results from 10-fold cross-validation and basically investigates the predictive ability of the models. As criteria, we considered the area under the receiver operator characteristic (ROC) curve (AUC, Hosmer & Lemeshow (2000)) as well as a number of scoring rules discussed in Gneiting & Raftery (2007) (hit rate, logarithmic score, Brier score, spherical score). All criteria are evaluated on the out-of-bag sample, i.e. the 10% of the data left out of estimation.

Figure 5 shows a parallel coordinate plot of the predictive performance measures that basically indicates that regularised estimation pays off in terms of predictive ability. Again, both Lasso and Ridge regularisation are quite close to each other with a minor tendency towards Lasso estimation. The larger variability in the results for the hit rate (i.e. the percentage of correctly classified observations) are induced by an undesirable property of this quantity: It does not measure how well a client is predicted in terms of the predicted probability but is only based on a zero/one decision rule. As a consequence, there is a hard break between clients with probabilities slightly larger or smaller than the cut-off value of 0.5 while the other scoring rules induce a smooth transition from clients with small probabilities to those with larger probabilities.

For illustrative purposes, Figure 6 shows the ROC curves for the three models obtained with the full data set. It confirms the improved predictive ability of the regularised estimation scheme since the ROC curves are consistently higher than those of the literature model. Also, the ROC curves for Lasso and Ridge are quite close over the complete domain.

Figure 7 shows the estimated nonlinear effects for the three models. All functions are in very close accordance with mild deviations from linearity. The age effect indicates a somewhat reduced credit-worthiness for young credit-holders up to an age of about 35. For older credit-holders, the effect is mostly constant and approximately equal to zero. The effect of the credit amount is mostly decreasing, indicating a higher probability for larger credits not to be paid back. Similarly, long-term credits have a higher probability of defaulting. Note that it is important to recognise that our analysis only differentiates between complete payment of the credit (credit-worthy) and incomplete payment (not credit-worthy). Especially for the amount and the duration effects, different functional forms may be obtained if the paid back percentage of a credit is considered as the dependent variable.

Table 2 shows the estimated regularised effects for the subset of covariates with posterior probability less than 0.05 in either the Lasso or the Ridge model. These include

information on several variables characterising the running account, confirming the use of analogous variables in the literature example. In addition, some variables characterising the intended use have highly significant impact on the probability of credit defaults.

6 Extensions

Joint regularisation and smoothing in high-dimensional structured additive regression can be extended in several directions. For instance, other regularisation priors that can be expressed as scale mixtures of normals can be employed instead of the Bayesian Ridge and Lasso. Ishwaran & Rao (2007) considered the so-called spike and slab prior constructed as a bimodal mixture of inverse Gamma priors. To be more specific, the prior variances τ_j^2 of high-dimensional covariate vectors are assigned the prior

$$\tau_j^2 \sim (1-\omega)\,\mathrm{IG}(a, \nu_0 b) + \omega\,\mathrm{IG}(a, \nu_1 b)$$

i.e. a mixture of two inverse gamma priors with mixture probabilities ω and $(1-\omega)$. The hyperparameters of the inverse gamma components share the parameters a and b but are different in the parameters ν_0 and ν_1. These are chosen to produce a mixture where one component is concentrated close to zero (ν_0 close to zero, the spike) and one component is clearly separated away from zero ($\nu_1 = 1$, the slab). Ishwaran & Rao (2007) have demonstrated that the spike and slab prior has very good variable selection and oracle properties and can also be used to select influential covariates based on the posterior probabilities for the mixture components. The spike and slab prior is also included in our software and is investigated in combination with smoothness priors in (Konrath, Kneib & Fahrmeir 2008).

Additional semiparametric components such as varying coefficient and interaction terms can be included in the predictor, specifying appropriate Gaussian priors of the same general form (3), see Fahrmeir, Kneib & Lang (2004). Furthermore, the entire approach can be extended to generalised regression models with the same high-dimensional predictor structure (1) but with different types of responses. In particular this includes

high-dimensional exponential family regression for discrete and other non-Gaussian responses and Cox-type hazard rate models (Konrath, Kneib & Fahrmeir 2008) that would for example allow to analyse the credit data with a logit model. For such intrinsically non-Gaussian models, the general structure of MCMC algorithms remains the same, but the full conditionals for regression and basis function coefficients are no longer Gaussian. Instead, block updates for $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p, \boldsymbol{\alpha}_{\text{spat}}$ are obtained from an Metropolis Hastings step with a Gaussian 'iteratively weighted least squares (IWLS)' proposal, where the full conditional is approximated by a Gaussian distribution. This leads to Gaussian proposal densities with expectation and covariance vector

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{X}' \boldsymbol{W} (\tilde{\boldsymbol{y}} - \boldsymbol{\eta}_{-\boldsymbol{\beta}}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\boldsymbol{X}' \boldsymbol{W} \boldsymbol{X} + \boldsymbol{P}_{\boldsymbol{\beta}})^{-1}$$

where $\boldsymbol{W}$ and $\tilde{\boldsymbol{y}}$ are the usual generalised linear model weights and working responses. Analogous expressions result for $\boldsymbol{\gamma}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_{\text{spat}}$. IWLS proposals are computationally efficient and require no manual fine tuning, see Brezger & Lang (2006) in the context of structured additive exponential family regression, and Konrath, Kneib & Fahrmeir (2008) for hazard rate models with high-dimensional predictor.

Acknowledgement: We thank Felix Heinzl for assistance in data analyses and gratefully acknowledge financial support from the German Science Foundation, grant FA 128/5-1. This paper has been written while the first author was visiting the Georg-August-University Göttingen in the winter term 2008/09.

References

Adebayo, S. and Fahrmeir, L. (2005). Analysing Child Mortality in Nigeria with Geoadditive Survival Models. *Statistics in Medicine*, **24**, 709–728.

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Bae, K. & Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, **20**, 3423–3430.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2003). *Hierarchical Modelling and Analysis of Spatial Data.* Chapman & Hall / CRC.

Binder, H. & Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* **9** 14.

Brezger, A. & Lang, S. (2006). Generalized additive regression based on Bayesian P-splines *Computational Statistics & Data Analysis*, **50**, 967–991.

Efron, B., Hastie, T., Johnstone, I. M. & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**, 407–499.

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

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statist. Sinica*, **14**, 731–761.

Fahrmeir, L. & Lang, S. (2001a). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Applied Statistics*, **50**, 201–220.

Fahrmeir, L. & Lang, S. (2001b). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics*, **53**, 10–30

Fahrmeir, L., Sagerer, F. & Sussmann, G. (2007). Geoadditive regression for analyzing small-scale gegraphical variability in car insurance. *Blätter der Deutschen Gesellschaft für Versicherungsmathematik*, **28**, 47–65.

Gamerman, D. (1997). Efficient Sampling from the Posterior distribution in Generalized Linear Models. *Statistics and Computing*, **7**, 57–68.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–534.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Goeman, J. J. (2007). An Efficient Algorithm for L1-penalized Estimation. Leiden, University Medical Center.

Griffin, J. E. & Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. University of Warwick, Deptartment of Statistics, Technical report.

Griffin, J. E. & Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization. University of Warwick, Deptartment of Statistics, Technical report.

Hans, C. (2008). Bayesian Lasso regression. Technical Report, Department of Statistics, Ohio State University.

Holmes, C. C. & Adams, N. M. (2002). A probabilistic nearest neighbour method for statistical pattern recognition. *Journal of the Royal Statistical Society Series B*, **64**, 295-306.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*, Wiley.

Ishwaran, H., & Rao, S. J. (2005). Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection. *Journal of the American Statistical Association*, **462**, 438–455.

Konrath, S., Kneib, T. & Fahrmeir, L. (2008). Bayesian Regularization and Smoothing for Hazard Regression. Technical Report No. 35, Department of Statistics, Ludwig-Maximilians-University Munich.

Lang, S. & Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Michie, D., Spiegelhalter, D. J. & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. Horwood, London.

Kandala, N.-B., Lang, S., Klasen, S. & Fahrmeir, L. (2001). Semi-parametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries. *Research in Official Statistics*, **1**, 81–100.

Kandala, N.-B., Fahrmeir, L., Klasen S. & Priebe J. (2007). Geo-additive models of childhood undernutrition in three Sub-Saharan African countries. *Population, Space and Place*, **14**.

Park, M. Y. & Hastie, T. (2006). L1 Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society, Series B*, **69**, 659–677.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **103**, 681–686.

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

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields. Theory and Applications*. CRC / Chapman & Hall, London.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Vandenhende, F., Eilers, P. H. C., Ledent, E., Renard, D. & Tibaldi, F. (2007). Joint detection of important biomarkers and optimal dose-response model using penalties *Statistics in Medicine*, **26**, 4876–4888

Wood, S. N. (2006). *Generalized Additive Models.* Chapman & Hall / CRC.

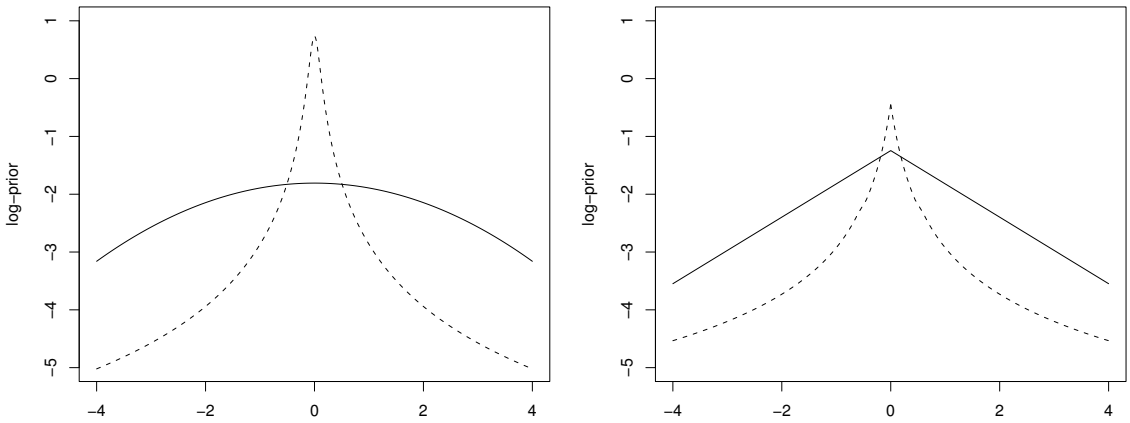

Figure 1: Log-densities of regularisation priors for fixed hyperparameters (solid line) and marginalised over the hyperparameters (dashed line). The left panel shows the Ridge prior and the marginal scaled t-distribution, the right panel shows the Bayesian Lasso with the marginal log-density (2).

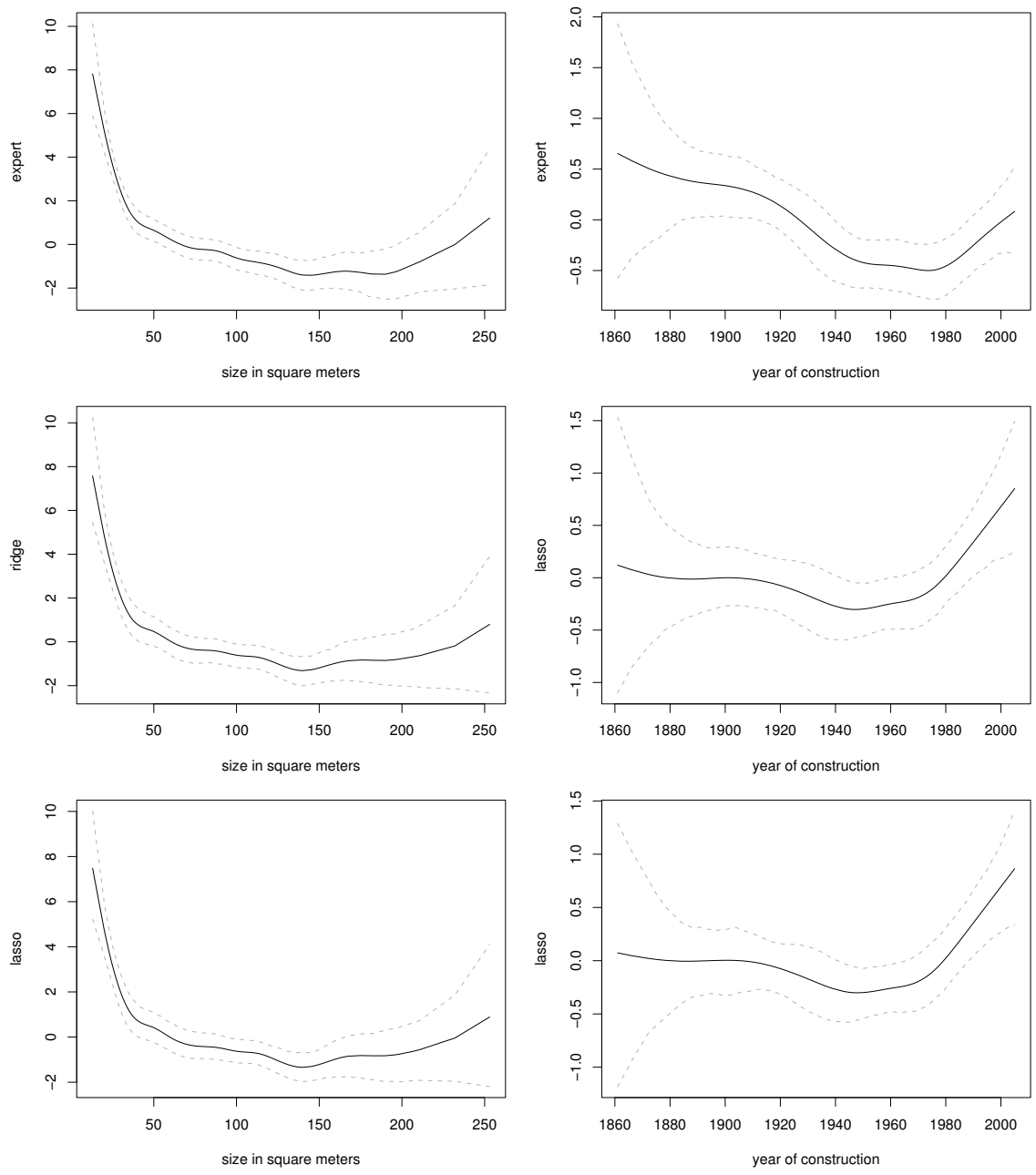

Figure 2: Rental guide: Estimated nonparametric effects and 95% pointwise credible bands in the expert model (top row), the Bayesian Ridge model (middle row) and the Bayesian Lasso model (bottom row).

Figure 3: Rental guide: Estimated spatial effects and 80% pointwise posterior probabilities in the expert model (top row), the Bayesian Ridge model (middle row) and the Bayesian Lasso model (bottom row).

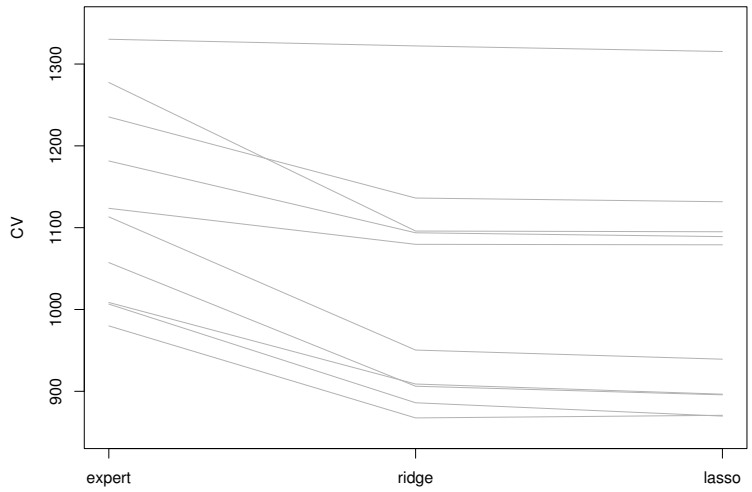

Figure 4: Rental guide: Parallel coordinate plot of cross-validated prediction error.

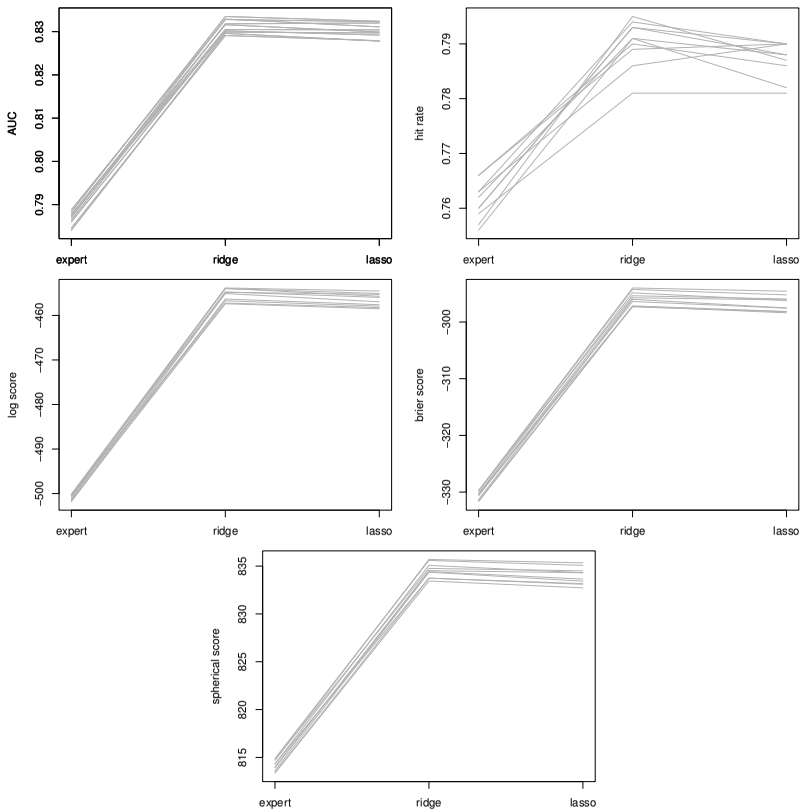

Figure 5: Credit scoring: Parallel coordinate plots of cross-validated predictive performance measures.

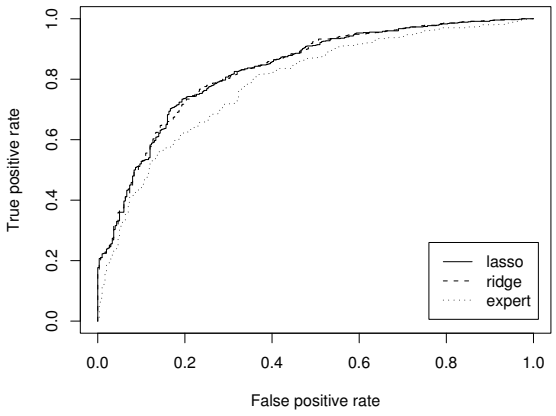

Figure 6: Credit scoring: Receiver operating characteristic curve.

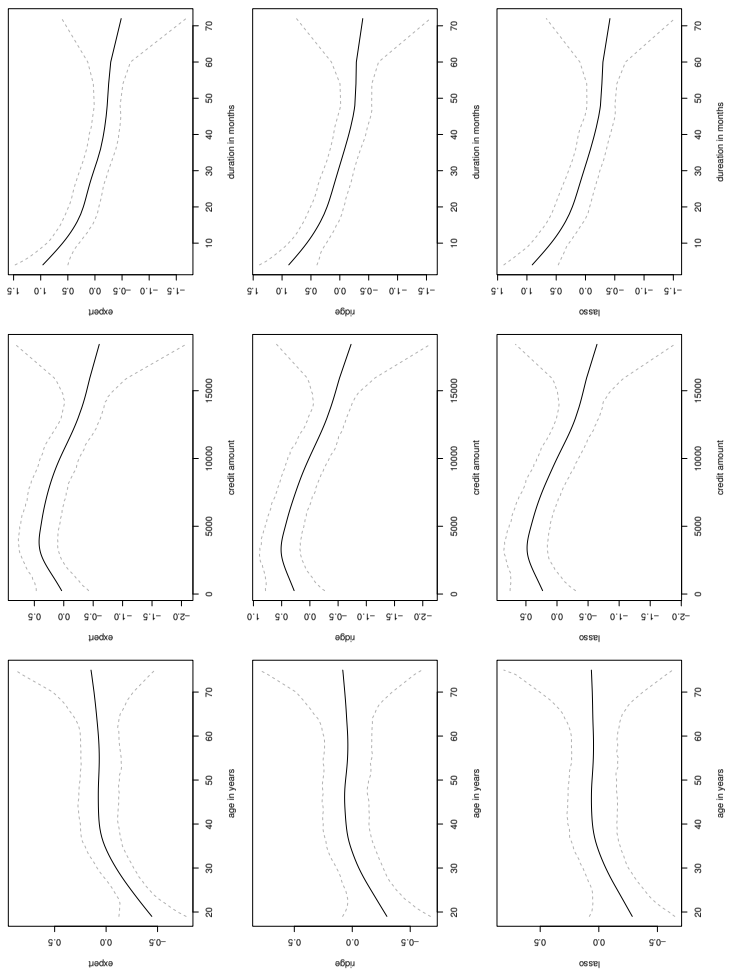

Figure 7: Credit scoring: Estimated nonparametric effects and 95% pointwise credible bands in the expert model (top row), the Bayesian Ridge model (middle row) and the Bayesian Lasso model (bottom row).

	Lasso		Ridge	
Variable	Post. Mean	Post. Prob.	Post. Mean	Post. Prob.
company housing	-1.318	0.000	-0.977	0.000
fridge	0.436	0.000	0.414	0.000
intercom	0.375	0.000	0.362	0.000
drying area	-0.429	0.000	-0.427	0.000
playground	-0.474	0.000	-0.455	0.000
attic	-0.565	0.000	-0.540	0.000
no. of occupants	0.119	0.000	0.125	0.006
side costs	-0.005	0.000	-0.005	0.000
floor covering (PVC)	0.545	0.000	0.489	0.000
floor covering (felt files)	0.441	0.000	0.389	0.004
rent start date before 2002	-0.916	0.000	-0.847	0.000
tiled bathroom	0.239	0.002	0.250	0.002
average residential area	-0.459	0.002	-0.429	0.004
no waterheating	-1.052	0.004	-0.532	0.004
no additional heating	-0.375	0.014	-0.400	0.002
special windows	0.268	0.014	0.272	0.010
grassy area	-0.260	0.016	-0.294	0.004
kitchen size $> 12m^2$	-0.207	0.022	-0.233	0.006
drying room	-0.182	0.022	-0.202	0.006
bath width	0.139	0.024	0.160	0.018
no. of roof terraces	0.506	0.026	0.406	0.024
floor covering (simple wooden floor)	0.337	0.026	0.312	0.010
modernised windows	-0.161	0.028	-0.189	0.024
modernised floor	0.183	0.030	0.197	0.024
bathtub	0.260	0.032	0.297	0.018
room 3 facing garden	0.219	0.040	0.247	0.054
built-in kitchen	0.215	0.044	0.236	0.030
2nd bathroom	0.327	0.048	0.327	0.030
shower	0.216	0.054	0.254	0.006
no balcony	-0.207	0.056	-0.242	0.036
satellite dish	-0.253	0.060	-0.273	0.022
room 4 facing street	-0.290	0.060	-0.300	0.032
cooking hobs	0.231	0.064	0.262	0.026
room 1 facing south	0.173	0.064	0.219	0.016
stairwell renovation	-0.175	0.066	-0.205	0.020
penthouse apartment	0.279	0.076	0.297	0.024
floor covering (linoleum)	-0.248	0.088	-0.313	0.010
2nd toilet	0.187	0.100	0.199	0.048

Table 1: Rental guide: Regularised effects with posterior probability lower than 0.05 in either the Bayesian Lasso or the Bayesian Ridge model.

	Lasso		Ridge	
Variable	Post. Mean	Post. Prob.	Post. Mean	Post. Prob.
no current account	-0.891	0.000	-0.820	0.000
no positive balance	-0.637	0.000	-0.593	0.000
no savings account	-0.390	0.004	-0.422	0.002
for new vehicle	0.491	0.012	0.445	0.010
poor paying habits	-0.509	0.018	-0.479	0.012
risky account	-0.554	0.018	-0.530	0.006
alternative purpose	-0.309	0.036	-0.353	0.008
medium term employment	0.270	0.040	0.269	0.070
no previous loans	-0.233	0.092	-0.269	0.024
no assets	0.206	0.130	0.277	0.046

Table 2: Credit scoring: Regularised effects with posterior probability lower than 0.05 in either the Bayesian Lasso or the Bayesian Ridge model.