Fahrmeir, Lang:

Bayesian Inference for Generalized Additive Regression based on Dynamic Models

Sonderforschungsbereich 386, Paper 134 (1998)

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

Projektpartner

Bayesian Inference for Generalized Additive Regression based on Dynamic Models

Ludwig Fahrmeir and Stefan Lang
Institut für Statistik, Universität München
Ludwigstr. 33
80539 München
e-mail: fahrmeir@stat.uni-muenchen.de and lang@stat.uni-muenchen.de

Abstract

We present a general approach for Bayesian inference via Markov chain Monte Carlo (MCMC) simulation in generalized additive, semiparametric and mixed models. It is particularly appropriate for discrete and other fundamentally non-Gaussian responses, where Gibbs sampling techniques developed for Gaussian models cannot be applied. We use the close relation between nonparametric regression and dynamic or state space models to develop posterior sampling procedures that are based on recent Metropolis-Hasting algorithms for dynamic generalized linear models. We illustrate the approach with applications to credit scoring and unemployment duration.

Keywords: generalized additive models, Markov chain Monte Carlo, mixed models, semiparametric Bayesian inference, state space models, varying coefficients

1 Introduction

In this paper we propose a general Bayesian approach via Markov chain Monte Carlo (MCMC) for inference in generalized additive and varying coefficients models, including extensions to models with random effects. Although additive models with Gaussian responses are also covered by the framework, our main interest lies in models for fundamentally non-Gaussian responses, such as binary or other discrete-valued responses. For Gaussian models, Gibbs sampling can be used for full Bayesian analyses, see for example Wong and Kohn (1996), who use state space or dynamic model representations of splines, or Hastie and Tibshirani (1998), who derive the Gibbs sampler as a Bayesian version of backfitting. For non-Gaussian responses, Gibbs sampling is no longer appropriate, and more general MCMC techniques are needed. Hastie and Tibshirani (1998) make a corresponding suggestion for a type of Metropolis-Hastings algorithm.

Our approach is based on the close relationship between dynamic generalized linear models (see, e.g., Fahrmeir and Tutz, 1997, ch.8) and generalized additive or varying coefficient models (Hastie and Tibshirani, 1990, 1993). To be more specific, consider the classical smoothing problem, where observations $y = (y(1), \ldots, y(n))$ are assumed to be the sum

$$y(t) = f(t) + \varepsilon(t), \qquad \varepsilon(t) \sim N(0, \sigma^2) \tag{1}$$

of a smooth regression function f, evaluated at equidistant design points $t = 1, \ldots, n$, and independent Gaussian noise variables. Within a dynamic or state space framework, observations are usually considered as time series data, observed at time

points $t = 1, \ldots, n$. The observation model (1) is supplemented by a linear Gaussian Markov model for the parameters or states $f = (f(1), \ldots, f(n))$. A common choice as so-called smoothness prior is a second order random walk model (RW(2))

$$f(t) = 2f(t-1) - f(t-2) + u(t), \qquad u(t) \sim N(0, \tau^2), \tag{2}$$

where i.i.d. errors $u(t)$ are independent from the noise variables in (1). For given variances σ^2 and τ^2, posterior means $\hat{f}(t)$ and variances can be efficiently computed by the Kalman filter and smoother. Assuming diffuse initial priors for $f(1)$, $f(2)$ and using the linear Gaussian models (1) and (2), the posterior means $\hat{f} = (\hat{f}(1), \ldots, \hat{f}(n))$ can also be derived as posterior mode estimators, that is, as minimizers of the negative (log-) posterior

$$\sum_{t=1}^{n}(y(t) - f(t))^2 + \frac{\sigma^2}{\tau^2}\sum_{t=3}^{n}(f(t) - 2f(t-1) - f(t-2))^2. \tag{3}$$

Obviously, the penalized least squares criterion (3), with smoothing parameter $\lambda = \sigma^2/\tau^2$, also has a non-Bayesian interpretation and is a discretized version of the corresponding criterion leading to cubic smoothing splines (e.g. Green and Silverman, 1994). Already for a moderate number of equidistant design points, cubic smoothing splines and the discrete version $\hat{f}$ from (3) are often difficult to distinguish visually. Basically, this equivalence extends to additive Gaussian models as well as to non-equally spaced design points or covariate observations.
For a full Bayesian analysis with hyperpriors for the variances σ^2 and τ^2, the Kalman filter and smoother can be exploited for efficient, blockwise Gibbs sampling (Carter and Kohn, 1994; Fruehwirth-Schnatter, 1994). Alternatively, Gibbs sampling could also be carried out as in Hastie and Tibshirani (1998), using the presentation $f'Kf$ of the penalty term with a symmetric, block diagonal penalty matrix K.
For fundamentally non-Gaussian responses as considered in this paper, the observation model (1) has to be replaced by a non-Gaussian model, and, as a consequence, the equivalence between posterior mean and posterior mode estimation by (3) is lost. The linear Kalman filter and smoother is no longer applicable and Gibbs sampling techniques cannot be reasonably applied. Therefore, we base posterior sampling on recent Metropolis-Hastings algorithms with so-called conditional prior proposals, developed by Knorr-Held (1998) in the context of dynamic generalized linear models. Other recently proposed procedures for MCMC inference in these models (Gamerman, 1998; Shephard and Pitt, 1997) might also be useful. The MCMC procedure gives rich output and permits estimation of posterior means, medians, quantiles and other functionals of regression functions. No approximations based on conjectures of asymptotic normality have to be made, and data-driven choice of smoothing parameters is automatically incorporated.
Bayesian generalized additive models are described in Section 2, while Section 3 contains details about the chosen MCMC techniques. In Section 4 we illustrate our approach by reanalyzing a semiparametric additive model for the credit scoring data given in Fahrmeir and Tutz (1997) and for data on unemployment durations from the German Federal Employment Office. The concluding section makes some suggestions for future research.

2 Bayesian generalized additive and varying coefficient models

Let us now turn to regression situations where observations $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, on a response y and a vector $(x_1, \ldots, x_p)$ of metrical covariates are given. In longitudinal studies, as in our application to duration of unemployment in Section 4, the covariate vector will typically include one or more time scales, such as duration and calendar time. Generalized additive models (Hastie and Tibshirani, 1990) assume that, given $x_i = (x_{i1}, \ldots, x_{ip})$, the distribution of y_i belongs to an exponential family, with mean $\mu_i = E(y_i|x_i)$ linked to an additive predictor η_i by

$$\mu_i = h(\eta_i), \ \eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}). \tag{4}$$

Here h is a known link or response function, and $f_1, \ldots, f_p$ are unknown smooth functions of the covariates. For identifiability reasons, unknown functions are centered appropriately. A slightly more general predictor is

$$\eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + w_i'\beta, \tag{5}$$

where $w_i = (w_{i1}, \ldots, w_{ir})$ is a vector of further covariates whose effect is assumed to be linear. Models with predictor (5) are sometimes termed generalized partially linear or semiparametric additive models. For example, w_i may contain binary indicators for categorical covariates as in our application to credit scoring in Section 4, for smoothing is not sensible for such covariates. Observation models of the form (4) or (5) may be appropriate if heterogeneity among units is sufficiently described by covariates. A common way to deal with this problem is the inclusion of additive random effects into the predictor. This leads to mixed models with predictor of the form

$$\eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + w_i'\beta + b_{g_i}, \tag{6}$$

where b_{g_i} is a unit- or group- specific random effect, with $b_{g_i} = b_g$ if unit i is in group g, $g = 1, \ldots, G$. For example, in our application to unemployment durations, b_g is an additional effect for unemployed from county g. Due to the large number of countries a fixed effect approach will not be feasible, and a random effects model is chosen instead. A further extension leads to varying coefficient models (Hastie and Tibshirani, 1993), possibly incorporating random effects,

$$\eta_i = f_1(x_{i1})z_{i1} + \ldots + f_p(x_{ip})z_{ip} + w_i'\beta + b_{g_i}. \tag{7}$$

The design vector $z = (z_1, \ldots, z_p)$ may contain components of x as well as some additional covariates. If a design variable is identical to 1, e. g. $z_j = 1$, then the corresponding function f_j is the main effect of x_j, while terms like $f_p(x_{ip})z_{ip}$ model an effect of z_p that varies over x_p or, in other words, interaction between x_p and z_p. For Bayesian semiparametric inference, the unknown functions $f_1, \ldots, f_p$, more exactly corresponding vectors of function evaluations, and the parameters $\beta = (\beta_1, \ldots, \beta_r)$ are considered as random variables. The observation models (4),(5),(6) or (7) are understood to be conditional upon these random variables, and have to be supplemented by appropriate prior distributions.

Priors for the unknown functions $f_1, \ldots, f_p$ are based on Gaussian smoothness

priors that are common in dynamic generalized linear models, see, for example, Fahrmeir and Tutz (1994, ch. 8). Let us first consider the case of a single covariate x with *equally-spaced observations* x_i, $i = 1, \ldots, n$. Then the ordered sequence $x_{(1)} < \ldots < x_{(t)} < \ldots < x_{(n)}$ defines an equidistant grid on the x-axis. The typical case for this situation arises if the covariate x corresponds to time t, and the grid points correspond to time units such as weeks, months, or years. Define $f(t) := f(x_{(t)})$ and let

$$f = (f(1), \ldots, f(t), \ldots, f(n))'$$

denote the vector of function evaluations. Then, just as for the time trends example in Section 1, common priors for smooth functions are, respectively, first or second order random walk models

$$f(t) = f(t-1) + u(t) \quad \text{or} \quad f(t) = 2f(t-1) - f(t) + u(t) \tag{8}$$

with Gaussian errors $u(t) \sim N(0; \tau^2)$ and diffuse priors $f(1) \propto \text{const}$, and $f(1)$ and $f(2) \propto \text{const}$, for initial values, respectively. Of course, higher order difference priors are also possible. For example if the covariate x is time t, measured in months, then a common smoothness prior for a seasonal component $f(t)$ is

$$f(t) + f(t-1) + \ldots + f(t-11) = u(t) \sim N(0, \tau^2). \tag{9}$$

Generally, we might specify Gaussian autoregressive priors of order k

$$f(t) = \sum_{l=1}^{k} \varphi_l f(t-l) + u(t), \;\; u(t) \sim N(0, \tau^2), \tag{10}$$

with diffuse priors assigned to initial values $f(1), \ldots, f(k)$. The prior (10) is equivalent to

$$f(t)|f(t-1), \ldots f(t-k), \ldots, f(1), \tau^2 \sim N(\sum_{l=1}^{k} \varphi_l f(t-l), \tau^2). \tag{11}$$

In shortened notation, write this as

$$f(t)|\cdot \quad \sim AR(k; \tau^2)$$

for (10) or (11), assuming diffuse priors for initial values $f(1), \ldots, f(k)$.
Due to the chronological ordering in (10) or (11), priors for $f = (f(1), \ldots, f(n))$ are seemingly defined in an *asymmetric* way. However it is important to note that these priors can always be rewritten in a *symmetric* form that is invariant to chronological ordering. Generally, this follows from the fact that any discrete Markov process like (10) or (11), can also be formulated in a symmetric way by conditioning not on previous variables $f(t-1), f(t-2), \ldots$, but also on future variables $f(t+1), f(t+2), \ldots$. For example, a first order random walk prior can be rewritten as

$$f(t)|f(s), s \neq t, \tau^2 \quad \sim \quad \begin{cases} N(f(2), \tau^2) & \text{for } t = 1 \\ N(\frac{1}{2}f(t-1) + \frac{1}{2}f(t+1), \frac{\tau^2}{2}) & \text{for } 2 \leq t \leq n-1 \\ N(f(n-1), \tau^2) & \text{for } t = n. \end{cases}$$

Symmetry is also evident from the multivariate Gaussian prior for the entire vector $f = (f(1), \ldots f(n))$ of function evaluations: From (8) it is easy to show that f has a partially improper prior

$$f \sim N(0; \tau^2 K^-),$$

where K^- is a generalized inverse of the precision matrix

$$K = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

Symmetric definitions of (10) or (11) can be derived in a similar way. Thus, the priors are reasonable not only for time scales with natural chronological order, but for any other metrical covariate.
Next we consider the general case with non-equally spaced observations. Let

$$x_{(1)j} < \ldots < x_{(t)j} < \ldots < x_{(n_j)j}, \qquad j = 1, \ldots, p$$

denote the $n_j \leq n$ strictly ordered, different observations of the covariate x_j, and

$$f_j = (f_j(1), \ldots, f_j(t), \ldots, f_j(n_j))'$$

with $f_j(t) := f_j(x_{(t)j})$, the vector of function evaluations.
Random walk or autoregressive priors have to be modified to account for nonequal distances $\delta_{tj} = x_{t(j)} - x_{t-1,(j)}$ between observations. Random walks of first order are now specified by

$$f_j(t) = f_j(t-1) + u_j(t), \quad u_j(t) \sim N(0; \delta_{tj}\tau_j^2), \tag{12}$$

i. e. , by adjusting error variances from τ_j^2 to $\delta_{tj}\tau_j^2$. Random walks of second order are

$$f_j(t) = \left(1 + \frac{\delta_{tj}}{\delta_{t-1,j}}\right) f_j(t-1) - \frac{\delta_{tj}}{\delta_{t-1,j}} f_j(t-2) + u_j(t), \tag{13}$$

$u(t) \sim N(0; w_{tj}\tau_j^2)$, where w_{tj} is an approppriate weight. The simplest weight is $w_{tj} = \delta_{tj}$ as in (12). More complex weights can be derived from corresponding continuous-time random walk models, i. e. stochastic differential equation priors, or with other arguments. Based on experience with simulated and real data, we recommend $w_{tj} = \delta_{tj}$ as a standard option. A related, yet different proposal for a second order autoregressive prior is given by Berzuini and Larissa (1996). Another possibility would be to use state space representations of stochastic differential equation priors based on the work by Kohn and Ansley (1988). Biller and Fahrmeir (1997) follow this idea, but there are significant problems associated with convergence and mixing behaviour of posterior samples, than in the case with the simpler priors chosen here.
The general formulation are autoregressive priors

$$f_j(t)|f_j(t-1), \ldots f_j(t-k) \sim N(\sum_{l=1}^{k_j} \varphi_{lj}(t) f_j(t-l); w_{tj}\tau_j^2) \tag{14}$$

with diffuse priors for initial values. In shortened notation, we write this as

$$f_j(t)|\cdot \quad \sim AR(k_j; w_{tj}\tau_j^2),\ j = 1, \ldots, p. \tag{15}$$

The variance parameters τ_j^2 in (14) act as smoothness parameters in analogy to penalized likelihood estimation. Smaller values of the variance τ_j^2 impose more smoothness on the unknown function f_j. For a fully Bayesian analysis, hyperpriors for variances are introduced in a further stage of the hierarchy. This allows for simultaneous estimation of the unknown function and the amount of smoothness. A common choice are highly dispersed inverse gamma priors

$$p(\tau_j^2) \sim IG(a_j; b_j).$$

A common choice for a and b is very small $a = b$, for example $a = b = 0.0001$, leading to almost diffuse priors for the variance parameters. An alternative proposed, for example, in Besag et al. (1995) is $a = 1$ and a small value for b, such as $b = 0.005$. However since estimation results tend to be sensitive to the choice of hyperpriors, especially in situations when data is sparse, some kind of sensitivity analysis should always be performed.

For the fixed effect parameters $\beta_1, \ldots, \beta_r$, we will usually assume independent diffuse priors $\beta_j \propto \text{const}$, $j = 1, \ldots, r$. Another choice would be highly dispersed Gaussian priors.

For random effects, we make the usual assumption that the b_g's are i.i.d. Gaussian,

$$b_g|v^2 \sim N(0, v^2), \quad g = 1, \ldots, G$$

and use again a highly dispersed hyperprior for v^2.

In the following, let

$$f = (f_1, \ldots, f_p), \quad \tau^2 = (\tau_1^2, \ldots, \tau_p^2), \quad \beta = (\beta_1, \ldots, \beta_r), \quad b = (b_1, \ldots, b_G)$$

denote parameter vectors for function evaluations, variances, fixed, and random effects. Then the Bayesian model specification is completed by the following *conditional independence assumptions*:

i) For given covariates and parameters f, β and b observations y_i are conditionally independent.

ii) Priors $p(f_j \mid \tau_j^2)$, $j = 0, \ldots, p$, are conditionally independent.

iii) Priors for fixed and random effects, and hyperpriors $\tau_j^2, j = 1, \ldots, p$, are mutually independent.

3 MCMC inference

Full Bayesian inference is based on the entire posterior distribution

$$p(f, \tau^2, \beta, b|y) \propto p(y|f, \tau^2, \beta, b)p(f, \tau^2, \beta, b).$$

By assumption (i), the conditional distribution of observed data y is the product of individual likelihoods:

$$p(y|f, \tau^2, \beta, b) = \prod_{i=1}^{n} L_i(y_i; \eta_i), \tag{16}$$

with $L_i(y_i; \eta_i)$ determined by the specific exponential family distribution and the form chosen for the predictor η.
Together with the conditional independence assumptions (ii) and (iii), we have

$$p(f, \tau^2, \beta, b|y) \propto \prod_{i=1}^{n} L_i(y_i; \eta_i) \prod_{j=1}^{p} \{p(f_j|\tau_j^2)p(\tau_j^2)\} \prod_{k=1}^{r} p(\beta_k) \prod_{g=1}^{G} p(b_g|v^2)p(v^2)$$

for the posterior.
Bayesian inference via MCMC simulation is based on drawings from full conditionals of single parameters or blocks of parameters, given the rest and the data. For Gaussian models, Gibbs sampling can be applied, and posterior samples for the unknown functions can be obtained by updating the entire vector $f_j = (f_j(1), \ldots, f_j(n_j))$ in a so-called multimove step, see, for example, Carter and Kohn (1994) and Wong and Kohn (1996), who use dynamic model representations of cubic splines, or Hastie and Tibshirani (1998), who derive the Gibbs sampler as a stochastic generalization of the backfitting algorithm. For fundamentally non-Gaussian responses as considered in this paper, Gibbs sampling is no longer feasible and more general Metropolis-Hastings algorithms are needed. Single-move steps, as in Carlin, Polson and Stoffer (1992), which update each parameter $f_j(t)$ separately, suffer from problems with convergence and mixing. Hastie and Tibshirani (1998) suggest Metropolis-Hastings multi-move steps. We adopt a computationally very efficient M-H-algorithm with conditional prior proposals developed recently by Knorr-Held for dynamic generalized linear models. Convergence and mixing is considerably improved by block moves, where blocks $f_j[r, s] = (f_j(r), \ldots, f_j(s))$ of parameters are updated instead of single parameters $\beta_j(s)$. Suppressing conditioning parameters and data notationally, the full conditionals for the blocks $f_j[r, s]$ are

$$p(f_j[r, s] \mid \cdot) \propto L(f_j[r, s])\, \mathrm{p}(f_j[r, s] \mid f_j(l), l \notin [r, s], \tau_j^2)$$

The first factor $L(f_j[r, s])$ is the product of all likelihood contributions in (16) that depend on $f_j[r, s]$. The second factor, the conditional distribution of $f_j[r, s]$ given the rest $f(l)$, $l \notin [r, s]$, is a multivariate Gaussian distribution. Its (conditional) mean $\mu_j(r, s)$ and covariance matrix $\Sigma_j(r, s)$ are obtained from the joint Gaussian prior for f_j by the usual formulae for conditional Gaussian distributions. M-H-block-move updates for $f_j[r, s]$ are obtained by drawing a conditional prior proposal $f_j^*[r, s]$ from the conditional Gaussian $N(\mu_j[r, s], \Sigma_j[r, s])$ and accepting it with probability

$$min\{1, \frac{L(f_j^*[r, s])}{L(f_j[r, s])}\},$$

see Knorr-Held (1998) for proofs and details. From a computational point of view, the main advantage is the simple form of the acceptance probability. Only the likelihood has to be computed, no first or second derivates etc. are involved, thus considerably reducing the number of calculations.
The full conditionals for the variance parameters τ_j^2, $j = 1, \ldots, p$ are inverse gamma densities

$$p(\tau_j^2 \mid \cdot) \propto IG(a_j', b_j') \tag{17}$$

with parameters

$$a_j' = a_j + \frac{n_j - k_j}{2}$$

and

$$b_j' = b + \frac{1}{2} \sum_{t=k_j+1}^{n_j} \frac{1}{w_{tj}} \sum_{l=0}^{k_j} \varphi_{lj}(t) f_j(t-l),$$

respectively. Thus updating of variance parameters can be done by simple Gibbs steps, drawing directly from the inverse gamma densities (17).
With a diffuse prior $\mathrm{p}(\beta_j) = const$ for the fixed effects parameters, the full conditional for β is

$$p(\beta \mid \cdot) \propto \prod_{i=1}^{n} L_i(y_i; \eta_i).$$

Updating of β can in principle be done by MH steps with a random walk proposal $q(\beta, \beta^*)$, but a serious problem is tuning, i.e. specifying a suitable covariance matrix for the proposal that guarantees high acceptance rates and good mixing. Especially when the dimension of β is high, with significant correlations among components, tuning "by hand" is no longer feasible. An alternative is the weighted least squares proposal suggested by Gamerman (1997). Here a Gaussian proposal is used with mean $m(\beta)$ and covariance matrix $C(\beta)$, where β is the current state of the chain. The mean $m(\beta)$ is obtained by making one Fisher scoring step to maximize the full conditional $\mathrm{p}(\beta \mid \cdot)$ and $C(\beta)$ is the inverse of the expected Fisher information, evaluated at the current state β of the chain. In this case the acceptance probability of a proposed new vector β^* is

$$min\{1, \frac{p(\beta^* \mid \cdot) q(\beta^*, \beta)}{p(\beta \mid \cdot) q(\beta, \beta^*)}\} \tag{18}$$

Note that q is not symmetric because the covariance matrix from C of q depends on β. Thus in principle the fraction $q(\beta^*, \beta)/q(\beta, \beta^*)$ can not be omitted from (18). In practice however experience shows that this fraction is almost always near one, so omitting the fraction does not affect significantly the efficiency of the algorithm, but rather leads to a considerable saving in computation. Further computer time, can be saved by omitting the Fisher scoring step when computing the mean of Gamermans

proposal, and simply taking the current state of the chain as the mean. Compared to Gamermans original proposal our slightly modified updating scheme for fixed effects parameters is more efficient and avoids tuning "by hand".

For an additional random intercept, the full conditional for parameter b_g is given by

$$p(b_g \mid \cdot) \propto \prod_{i \in \{j: g_j = g\}} L_i(y_i; \eta_i) p(b_g | v^2)$$

Here a simple Gaussian random walk proposal with mean b_g and variance v^2 works well in most cases. To improve mixing, tuning is sometimes required by multiplying the prior variance v^2 in the proposal with a constant factor, e.g. 2. An alternative is, again, Gamerman's weighted least squares proposal or a slight modification. This becomes especially attractive when the observation model contains one or more random slope parameters in addition to the random intercept. By analogy to the variance parameters τ_j of nonparametric terms the full conditional of v^2 is again an inverse gamma distribution, so updating is straightforward.

4 Applications

4.1 Credit-Scoring

In our first application, we reanalyze the credit–scoring problem described in Fahrmeir and Tutz (1994, ch. 2.1). The aim of credit–scoring is to model or predict the probability that a client with certain covariates ("risk factors") is to be considered as a potential risk, and therefore will probably not pay back his credit as agreed upon by contract. The data set consists of 1000 consumers' credits from a South German bank. The response variable is "creditability", which is given in dichotomous form ($y = 0$ for creditworthy, $y = 1$ for not creditworthy). In addition, 20 covariates assumed to influence creditability were collected. As in Fahrmeir and Tutz (1997), we will use a subset of these data, containing only the following covariates, which are partly metrical and partly categorical:

- x_1 running account, trichotomous with categories "no running account" (= 1), "good running account" (= 2), "medium running account" ("less than 200 DM" = 3 = reference category)
- x_3 duration of credit in months, metrical
- x_4 amount of credit in DM, metrical
- x_5 payment of previous credits, dichotomous with categories "good", "bad" (=reference category)
- x_6 intended use, dichotomous with categories "private" or "professional" (=reference category)
- x_8 marital status, with reference category "living alone".

Effect coding is used for all categorical covariates. A parametric logit model for the probability $\text{pr}(y = 1|x)$ of being not creditworthy, leads to the conclusion that the covariate "amount of credit" has no significant influence on the risk. Here, we

reanalyze the data with a partial linear logit model

$$\log \frac{\text{pr}(y=1|x)}{1-\text{pr}(y=1|x)} = \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + f_3(x_3) + f_4(x_4) + \beta_5 x_5 + \beta_6 x_6 + \beta_8 x_8,$$

where x_1^1 and x_1^2 are dummies for the categories "good" and "medium" running account. The predictor has semiparametric additive form: The smooth functions $f_3(x_3)$, $f_4(x_4)$ of the metrical covariates "duration of credit" and "amount of credit", are estimated nonparametrically using second order random walk models for non–equally spaced observations. The constant β_0 and the effects $\beta_1, \ldots, \beta_8$ of the remaining categorical covariates are considered as fixed and estimated jointly with the curves f_3 and f_4, assuming diffuse priors.

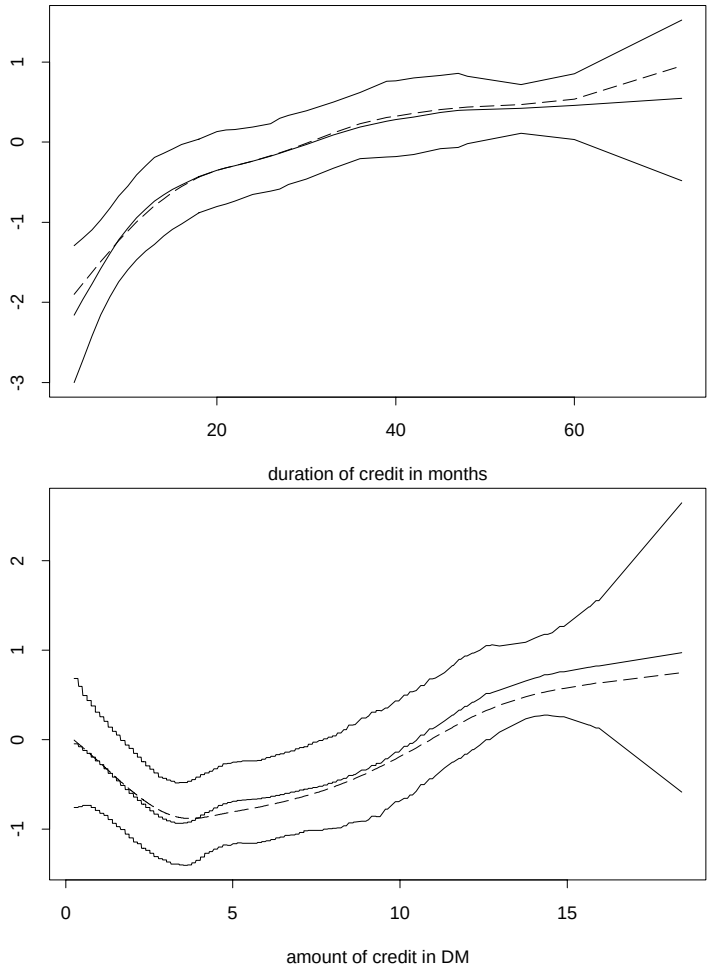

Figure 1: Estimated effects of duration and amount of credit. Shown is the posterior mean within 80 % credible regions and for comparison cubic smoothing splines (dotted lines)

.

Figure 1 shows estimates for the curves f_3 and f_4. For comparison, cubic smoothing splines are included in addition to posterior mean estimates. Although cubic splines

are posterior mode estimators and the penalty terms are not exactly the same, both estimates are close. While the effect of the variable "duration of credit" is almost linear, the effect of "amount of credit" is clearly nonlinear. The curve has a bathtub shape, and indicates that not only high credits but also low credits increase the risk, compared to "medium" credits between 3000–6000 DM. Apparently, if the influence is misspecified by assuming a linear function $\beta_4 x_4$ instead of $f_4(x_4)$, the estimated effect $\widehat{\beta}_4$ will be near zero, corresponding to an almost horizontal line $\widehat{\beta}_4 x_4$ near zero, and falsely considered as nonsignificant.

Table 1 gives the posterior means together with 80% credible intervals and, for comparison, maximum likelihood estimates of the remaining effects. Both estimates are in close agreement. They also have the same signs and are quite near to the estimates for a parametric logit model given in Fahrmeir and Tutz (1997), so that interpretation remains qualitatively the same for these constant effects.

covariate	mean	10 % quantile	90 % quantile	ML estimator
x_1^1	0.86	0.63	1.07	0.86
x_1^2	-1.09	-1.32	-0.85	-1.09
x_5	-0.49	-0.74	-0.25	-0.50
x_6	-0.22	-0.37	-0.07	-0.22
x_8	-0.26	-0.42	-0.11	-0.26

Table 1: Estimates of constant parameters for the credit–scoring data.

4.2 Duration of unemployment

In this second application, we analyze unemployment data from the German Federal Employment Office ("Bundesanstalt für Arbeit"). Typical questions that arise in studies on duration of unemployment are: How can the baseline effect (duration dependence) be modelled? How can trend and seasonal effects of calendar time be flexibly incorporated? What effect has age? Are there regional differences for the probability of leaving unemployment and seeking a new job? An important problem in connection with persistant unemployment in the 90′s in Europe, is the effect of unemployment compensation and social welfare. Are there negative side-effects of public unemployment compensation?

Our analysis is based on the following covariates:

- D calendar time measured in months
- A age (in years) at the beginning of unemployment
- S sex, dichotomous with categories "male" and "female" (= reference category)
- N nationality, dichotomous with categories "german" and "foreigner" (= reference category)
- U unemployment compensation, trichotomous with categories "unemployment benefit" (=reference category), "unemployment assistance" (U_1) and "subsistence allowance" (U_2).
- C county in which the unemployed have their domicil

Note that calendar time D and unemployment compensation U are both duration time dependent covariates. As in our first application effect coding is used for all categorical covariates. Since duration of unemployment is measured in months, we use a discrete time duration model as described in Fahrmeir and Tutz (1997, ch. 9). Let $T = t \in \{1, \ldots, q+1\}$ denote end of duration in month t after beginning of unemployment, and $x_t^* = (x_1, \ldots, x_t)$ the history of covariates up to month t. Then the discrete hazard function is given by

$$\lambda(t; x_t^*) = \text{pr}(T = t \mid T \geq t, x_t^*), \quad t = 1, \ldots, q.$$

We assume that censoring is noninformative and occurs at the end of the interval, so that the risk set R_t includes all individuals who are censored in interval t. We define binary event indicators y_{it}, $i \in R_t$, $t = 1, \ldots, t_i$, by

$$y_{it} = \begin{cases} 1 & \text{if } t = t_i \text{ and } \delta_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then the duration process of individual i can be considered as a sequence of binary decisions between remaining unemployed $y_{it} = 0$ or leaving for the absorbing state $y_{it} = 1$, i.e. end of unemployment at t. For $i \in R_t$, the hazard function for individual i can be modelled by binary response models

$$\text{pr}(y_{it} = 1 \mid x_{it}^*) = h(\eta_{it}), \tag{19}$$

with appropriate predictor η_{it} and response function $h : \mathbf{R} \to (0, 1)$. We choose a logit model with semiparametric predictor

$$\eta = f_0(t) + f_1^T(D) + f_2^S(D) + f_3(A) + \beta_1 S + \beta_2 N + \beta_2 U^1 + \beta_2 U^2 + b_C.$$

The baseline effect $f_0(t)$, the calendar time trend $f_1^T(D)$, and the effect of age $f_3(A)$ are estimated nonparametrically using second order random walks. For the seasonal effect $f_2^S(D)$ we choose the smoothness prior (9). The influences of the categorical covariates sex, nationality, and unemployment compensation, are modelled as fixed effects. To cope with regional heterogeneity, a county specific random effect b_C is incorporated into the linear predictor. The estimation results of the nonparametric terms and the seasonal component are shown in Figure 2 a)-f). The baseline effect (Figure a)) is downward sloping. Therefore, the possibility of finding a job is a decreasing function of the duration of unemployment. The effect of age in figure b) is slowly declining until age 53, dramatically declining for people older than 53. Figure c) displays the calendar time trend. For comparison with the estimated trend, the absolute number of unemployed people in Germany from 1980 to 1995 is shown in Figure d). Not surprisingly, a declining calendar time trend corresponds to an increase in the unemployment rate, and vice versa. So the estimated calendar time trend accurately reflects the economic trend of the labor market in Germany.

The estimated seasonal pattern (Figure e)) is relatively stable over the observation period. To gain a better insight, a section of the seasonal pattern for 1988 is displayed in Figure f). It shows typical peaks in spring and autumn, a global minimum in winter, and a local minimum in July. Low rates of hirings in summer can be explained by the distribution of holidays and vacations. In Figure 3 the estimated posterior mean of the county specific random effect b_C is displayed, showing a strong spatial pattern, with better chances of getting a new job in the southern part of West Germany, and lower chances in the middle and in the north.
Table 2 gives results of the remaining effects.

covariate	mean	10 % quantile	90 % quantile
S	0.19	0.17	0.20
N	0.08	0.04	0.12
U^1	0.11	0.07	0.16
U^2	-0.49	-0.57	-0.42

Table 2: Estimates of constant parameters in the unemployment data.

Males and Germans have improved job chances compared to females and foreigners, but the effects are not overwhelmingly large. The estimate of -0.49 for the subsistance allowance is significantly negative, while the effect of unemployment is slightly positive. Due to effect coding, the effect of insurance based unemployment benefits is $0.38 = 0.49 - 0.11$ and is therefore clearly positive. At first sight, this result seems to contradict the widely-held conjecture about the negative side-effects of unemployment benefits. However, it may be that the variable "unemployment benefit" also acts as a surrogate variable for those who have worked, and therefore contributed regularly to the insurance system in the past. Further substantive research will be necessary to give definite answers.

5 Conclusions

Non- and semiparametric Bayesian regression is a useful tool for practical data analysis. It provides posterior mean or median estimates, confidence bands, and estimates of other functionals, without approximate normality of estimators. Data-driven choice of smoothing parameters is also incorporated as part of the model. Many recent approaches based on smoothness priors or basis functions considered the case of Gaussian or related responses, our method is particularly useful for nonparametric regression with fundamentally non-Gaussian responses. The main advantage of hierarchical Bayesian models for nonparametric regression is their modular structure and flexibility. By appropriate modifications of observation models or priors, generalizations and extensions to other settings are conceptually simple.
For example, inclusions of interactions between metrical covariates in the observation model can be based on the suggestion of Clayton(1996) for the interaction effects between two time scales. Let $x_{(t)j}$, $t = 1, \ldots, n_j$, and $x_{(s)k}$, $s = 1, \ldots, n_k$ denote the strictly ordered, different observations of two covariates x_j and x_k, and $f_{jk}(t, s) := f_{jk}(x_{(t)j}, x_{(s)k})$ the interaction effects. If the smoothness priors for the

main effects f_j and f_k are, for example, first order random walks as in (8) or (12), the smoothness priors for f_{jk} are defined by "first differences of first differences". This leads to the interaction smoothness prior

$$f_{jk}(t,s) - f_{jk}(s-1,t) - f_{jk}(t-1,s) + f_{jk}(s-1,t-1) = u_{jk}(t,s) \sim N(0, \delta_{ts,jk}\tau^2_{jk})$$

where $\delta_{ts,jk}$ is a measure of the distance between the observation pairs $(x_{(t-1)j}, x_{(s-1)k}$ and $(x_{(t)j}, x_{(s)k})$. It can be shown that this defines a global prior $f_{jk} \sim N(0, \tau^2_{jk}K^-_{jk})$, where the precision matrix is obtained as the Kronecker product $K_{jk} = K_j \oplus K_k$ of corresponding precicison matrics K_j and K_k of the main effects. The same idea remains valid for other main effect priors like second order random walk models, and can be considered as the Bayesian analogue of modelling interactions by tensor product splines in a penalized log-likelihood framework.
For regression data with spatial labels on them, as in our second application, the i.i.d. prior for the random effects in the predictor β_g could be replaced by a Markov random field prior

$$\beta_g|\beta_{g'\neq g}, v^2 \sim N(\sum_{g'\sim g} \frac{\beta'_g}{a_g}, \frac{v^2}{a_g}),$$

where a_g is the number of neighboring regions.
To fit unsmooth functions $f(x)$, i.e. functions with discontinuities, edges or rather volatile curvature, the Gaussian prior for the errors in random walk or autoregressive models might be replaced by heavy-tail distributions, or by Gaussian distributions with locally varying variances

$$u_j(t) \sim N(0, \tau^2_{tj}), \qquad \tau^2_{tj} = exp(h_{tj}),$$

with h_{tj} obeying a random walk model in a further stage of the hierarchy. We intend to investigate these possibilities in future research.

Software:
We have implemented most of the ideas in this paper as a Windows NT based application. The program will soon be available for public use under **http://www.stat.uni-muenchen.de/ lang/**.

Acknowledgement:
This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen". We thank Leo Knorr-Held for helpful discussions and Stefan Bender for providing the unemployment data, and Murray Smith for help with the final preparation of the manuscript.

References

Berzuini, C. Larizza, Ch. (1996): *A unified approach for modeling longitudinal*

and failure time data, with application in medical monitoring. Transactions on Pattern Analysis and Machine Intelligence, **18**, 2, 109-122.

Biller, C., Fahrmeir, L. (1997): *Bayesian spline-type smoothing in additive generalized regression.* Computational Statistics,

Clayton, D. (1996): *Generalized linear mixed models.* In: Gilks, W., Richardson S. and Spiegelhalter D. (eds), Markov Chain Monte Carlo in Practice. London: Chapman and Hall, 275-301.

Carter, C.K., Kohn, R. (1996): *Markov chain Monte Carlo in conditionally Gaussian state space models.* Biometrika, **83**, 3, 589-601.

Fahrmeir, L., Knorr-Held, L. (1997): *Dynamic discrete time duration models.* Sociological Methodology, **27**, 417-452.

Fahrmeir, L., Tutz, G. (1997): *Multivariate Statistical Modelling based on Generalized Linear Models.* New York: Springer–Verlag.

Fruehwirth-Schnatter, (1994): *Data augmentation and dynamic linear models.* Journal of Time Series Analysis, **15**, 2, 183-202.

Gamerman, (1997): Efficient Sampling from the posterior distribution in generalized linear models. Statistics and Computing, **7**, 57-68.

Gamerman, D. (1998): Markov Chain Monte Carlo for dynamic generalized linear models. Biometrika **85**, 215-227.

Green, P.J., Silverman, B. (1994): Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

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

Hastie, T., Tibshirani, R. (1993): *Varying-coefficient Models.* Journal of the Royal Statistical Society, **B 55**, 757-796.

Hastie, T., Tibshirani, R. (1993): *Bayesian Backfitting.* Preprint, Department of Statistics, Stanford.

Knorr-Held, L. (1998): *Conditional Prior Proposals in Dynamic Models.* Scandinavian Journal of Statistics, to appear.

Shephard, N., Pitt, M.K. (1997): *Likelihood analysis of non-Gaussian measurement time series.* Biometrika, **84**, 653-667.

Wong, C., Kohn, R. (1996): *A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models.* Journal of Time Series Analysis, **17**, 203-220.

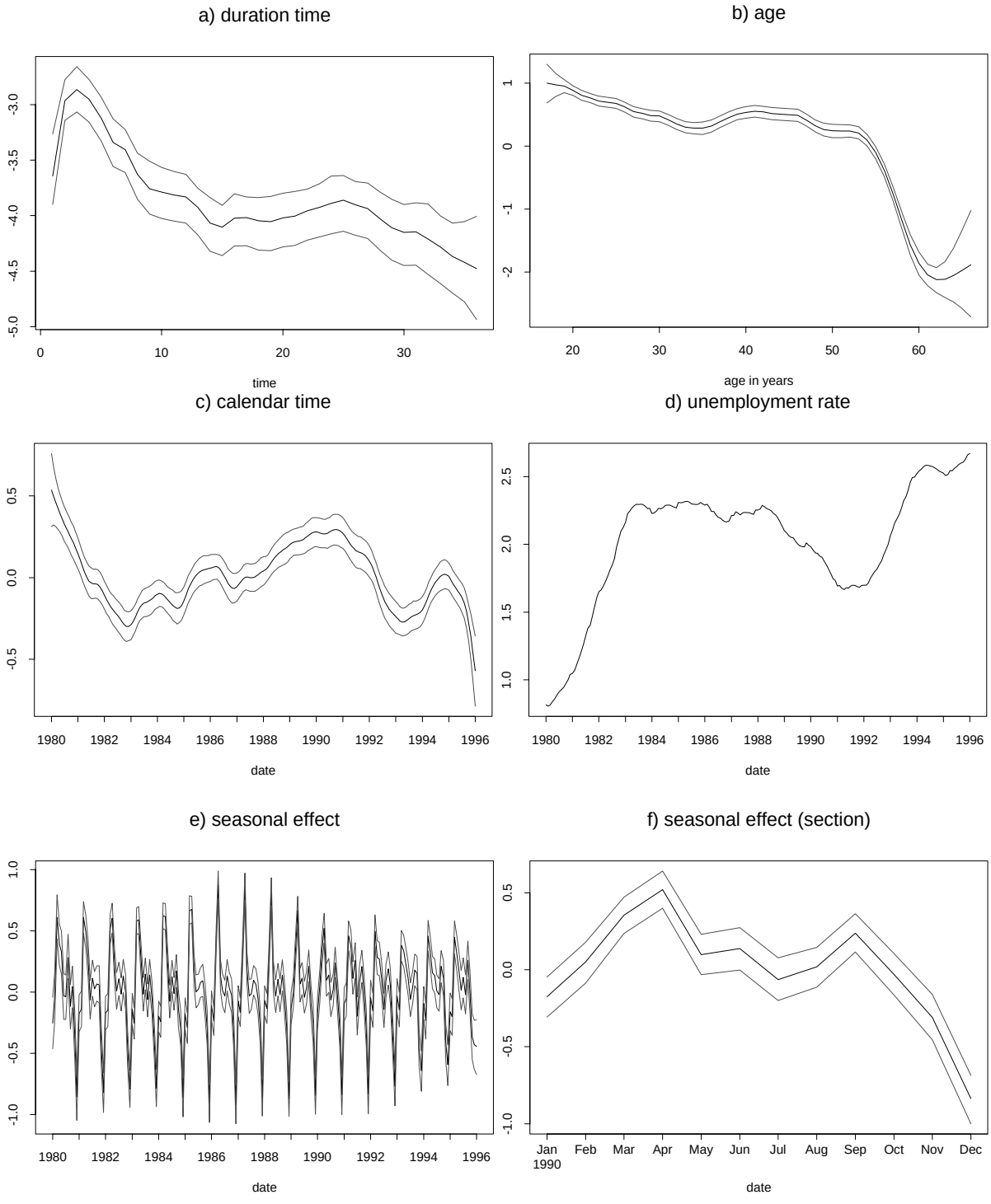

Figure 2: Estimated nonparametric functions and seasonal effect. Shown is the posterior mean within 80 % credible regions

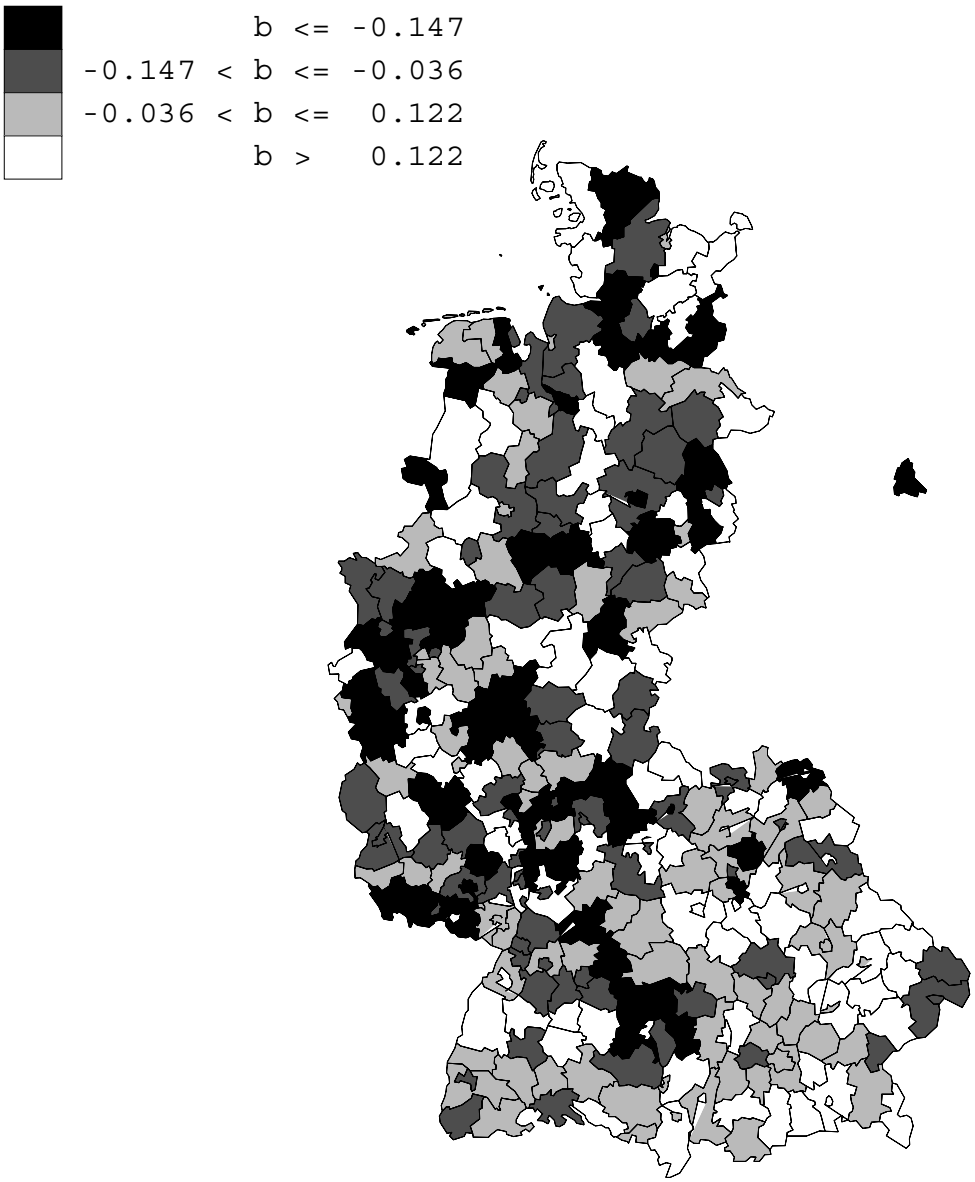

Figure 3: posterior mean of the county specific random effect