Generalized Semiparametric Regression with Covariates Measured with Error

Thomas Kneib, Andreas Brezger and Ciprian M. Crainiceanu

Abstract We develop generalized semiparametric regression models for exponential family and hazard regression where multiple covariates are measured with error and the functional form of their effects remains unspecified. The main building blocks in our approach are Bayesian penalized splines and Markov chain Monte Carlo simulation techniques. These enable a modular and numerically efficient implementation of Bayesian measurement error correction based on the imputation of true, unobserved covariate values. We investigate the performance of the proposed correction in simulations and an epidemiological study where the duration time to detection of heart failure is related to kidney function and systolic blood pressure.

Key words: additive hazard regression; generalized additive models; MCMC; measurement error correction; penalized splines

1 Introduction

The presence of covariates measured with error in regression models can have severe impact on inferential conclusions drawn from naive estimates. This is particularly true for semiparametric regression models where the relation between responses and covariates is specified flexibly and therefore also more prone to disturbances induced by measurement error. A common phenomenon in naive analyses are estimates that

Thomas Kneib
Institut für Mathematik, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany, e-mail: thomas.kneib@uni-oldenburg.de

Andreas Brezger
HypoVereinsbank Munich, e-mail: andreas.brezger@hvb.de

Ciprian M. Crainiceanu
Department of Biostatistics, Johns-Hopkins-University Baltimore, e-mail: ccrainic@jhsph.edu

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_8,

are biased towards zero and therefore underestimate effects. In particular, in semiparametric regression models it will be more difficult to detect local extrema of a functional relationship and curvature will be underestimated. In general, the effect of measurement error is insidious and leads to biased estimates, misspecified variability and feature masking (Carroll et al. 2006). Hence, it is likely to erroneously conclude that covariates are not associated with the response variable or to obtain false conclusions about the precise functional form of relationships.

Based on work by Berry et al. (2002) for Gaussian scatterplot smoothing, we develop a flexible Bayesian correction procedure based on Markov chain Monte Carlo (MCMC) simulations for general semiparametric exponential family and hazard regression models. The key ingredient is the imputation of the unobserved, true covariate values in an additional sampling step, an idea dating back to Stephens & Dellaportas (1992) and Richardson & Gilks (1993), see also Gustafson (2004). The Bayesian approach considered in this paper combines a number of distinct advantages:

Flexibility in terms of the response type: A wide range of response types is supported, including exponential family regression (e.g. Binomial or Poisson responses) as well as right-censored continuous-time survival times. This is made possibly by the consideration of a iteratively weighted least squares proposals for the regression coefficients (Gamerman 1997, Brezger & Lang 2006), a proposal scheme that relies on Gaussian approximations of the full conditionals.

Flexibility in terms of the model equation: All nonparametric model components are specified flexibly in terms of Bayesian penalized splines (Brezger & Lang 2006, Jullion & Lambert 2007). The modular structure of Bayesian computations based on MCMC enables the consideration of models where several covariates are measured with error in combination with further nonparametric effects of covariates observed exactly. Spatial effects, varying coefficient terms, or random effects are readily available as additional model components and are also included in our software.

Flexibility in terms of the measurement error equation: Based on the classical model of uncorrelated additive Gaussian measurement error, longitudinally correlated repeated observations on the measurement error equation or other extended measurement error models could easily be included.

Numerically efficient implementation: Sparse matrix computations and efficient storage schemes in combination with data compression based on rounding provide a rather fast estimation procedure. This, in particular, allows us to consider more complex applications with large sample size and extensive simulation setups.

In the application that motivated our research, the duration until detection of heart failure is analyzed in a hazard regression model that includes nonlinear effects of kidney function measured by the glomerular filtration rate (GFR) and systolic pressure (SP). Both covariates are inherently subject to measurement error due to different reasons: while SP is measured with error due to the relatively imprecise instruments involved in standard hemodynamometry, GFR can only be obtained accurately based on a time-consuming, awkward procedure; thus, in practice, this procedure is replaced with an estimate (eGFR) predicted from creatinine, gender and age (Hsu et al. 2005).

The prediction equation has been derived from a regression model and an estimate of the measurement error variance is also available from a replication study. In case of SP, the measurement error variance is available from previous studies. The sample size of 15,000 observations and two covariates measured with error make this application challenging, since we are faced with the imputation of 30,000 true covariate values and the re-evaluation of the corresponding parts of the design matrix in each iteration.

2 Semiparametric Regression Models with Measurement Error

2.1 Observation Model

In a generalized semiparametric regression models (see for example Ruppert et al. (2003), Fahrmeir et al. (2004) or Wood (2006)), the expectation of (conditionally) independent responses y_i, $i,\ldots,n$, from univariate exponential families is related to an additive predictor

$$\eta_i = f_1(x_{i1}) + \ldots + f_r(x_{ir}) + v_i'\gamma \tag{1}$$

based on a response function h, i.e. $\mu_i = E(y_i|\eta_i) = h(\eta_i)$. The predictor is additively composed of smooth functions $f_1,\ldots,f_r$ of continuous covariates $x_1,\ldots,x_r$ in combination with parametric effects γ of further, typically categorical covariates v. For hazard regression models employed in survival analysis, data are given in the form of (conditionally) independent survival data (t_i,δ_i), $i=1,\ldots,n$ where t_i is the (right-censored) observed survival time and δ_i is the censoring indicator. Extending the classical Cox model, semiparametric hazard regression models (Hennerfeind et al. 2006, Kneib & Fahrmeir 2007) can then be specified as $\lambda_i(t) = \exp(\eta_i(t))$ where

$$\eta_i(t) = g_0(t) + f_1(x_{i1}) + \ldots + f_r(x_{ir}) + v_i'\gamma$$

is a semiparametric predictor consisting of the log-baseline hazard rate $g_0(t)$, r smooth functions of continuous covariates, and linear effects summarized in $v'\gamma$. The time-dependent function $g_0(t)$ relates to the baseline hazard rate $\lambda_0(t)$ in the Cox model via $\lambda_0(t) = \exp(g_0(t))$. In contrast to usual partial likelihood estimation, determination of the baseline hazard rate will be an integral part of model estimation in our framework. In particular, estimation will be based on the full instead of the partial likelihood.

Estimation of the nonlinear functions $f_j(x_j)$ is frequently complicated by the fact that in applications the corresponding covariates x_j are not observed exactly so that only contaminated surrogate variables are available. Naive estimates based on these surrogate variables will then be oversmoothed leading to estimates that are biased towards "no effect" models. In the following, we assume that the first r_1 covariates $x_1,\ldots,x_{r_1}$ are subject to measurement error while the remaining $r_2 = r - r_1$ covariates

$x_{r_1+1},\ldots,x_r$ are observed exactly. In particular, we allow for several covariates x_j measured with error.

2.2 Measurement Error Model

In the classical measurement error model (Carroll et al. 2006), the true measurements of the covariates are contaminated by i.i.d. Gaussian noise, leading to the measurement of proxy variables

$$w_{ij}^{(m)} = x_{ij} + u_{ij}^{(m)}, \quad m = 1,\ldots,M$$

where $u_{ij}^{(m)} \sim \mathrm{N}(0,\tau_{u,j}^2)$. In our modeling framework, we allow for the possibility of repeated measurements (indexed by $m = 1,\ldots,M$) on a covariate. For simplicity, we assume that the measurement error contaminations are mean zero and independent, i.e. $u_{ij} = (u_{ij}^{(1)},\ldots,u_{ij}^{(M)})' \sim \mathrm{N}(\mathbf{0},\tau_{u,j}^2 I_M)$. However, the MCMC sampling mechanism presented in Section 3 can straightforwardly be extended to more general situations where $u_{ij} \sim \mathrm{N}(\mu,\Sigma)$. Inclusion of covariances in Σ could for example be useful in combination with a longitudinal collection of the repeated measurements where Σ contains an equicorrelation or autoregressive correlation structure (see Wang & Pepe (2000) for such an example). Non-zero expectations μ can, for example, be employed to adjust for measurement bias in the repeated observations.

2.3 Prior Distributions

To complete the Bayesian specification, suitable priors have to be assigned to all model parameters. In the Bayesian perspective on the model, the unknown true covariate values x_{ij} are treated as additional unknowns and imputation becomes a part of the MCMC algorithm. For the fixed effects γ, we assume standard noninformative priors, i.e. $p(\gamma) \propto \text{const}$. In contrast, we assign informative priors to the smooth function to enforce smoothness of the corresponding estimates.

2.3.1 P-spline Priors

A parsimonious yet flexible modelling possibility for nonparametric function estimation are penalized splines as popularized by Eilers & Marx (1996) and extensively discussed in Ruppert et al. (2003). In our Bayesian framework, we employ the Bayesian analogue developed by Brezger & Lang (2006). For the sake of simplicity, we drop the function index j in the following description. To represent $f(x)$ (or $g_0(t)$ in case of hazard regression models) in terms of a flexible but finite dimensional class of functions, we assume that it can be expanded in B-splines of leading to the basis

function representation

$$f(x) = \sum_{k=1}^{K} \beta_k B_k^l(x)$$

where $B_k^l(x)$ are B-spline basis functions of degree l defined upon a set of knots $\kappa_1 < \ldots < \kappa_K$, and β_k are the corresponding regression coefficients. In the classical frequentist formulation of P-splines, smoothness of the functions $f(x)$ is enforced by adding a squared difference penalty of order d to the likelihood that essentially penalizes large variation in terms of the d-th derivative. In a Bayesian formulation, d-th order differences are replaced by d-th order random walks, e.g.

$$\beta_k - \beta_{k-1} \sim \mathrm{N}(0, \tau_\beta^2 \omega_k)$$

for first order random walks in the most simple case. This prior specification corresponds to local increments in the coefficient sequence with expectation zero and deviations controlled by the variance τ_β^2 and the distance between the corresponding knots $\omega_k = \kappa_k - \kappa_{k-1}$. The underlying rationale of the latter choice is that larger steps between two knots should also be reflected in the prior in allowing for larger variation. In contrast, the variance parameter τ_β^2 controls the overall variability of the function estimate with small values corresponding to very flat estimates whereas large values yield very flexible estimates. The weighted first order random walk can also be interpreted as a discrete approximation to continuous Brownian motion that yields a similar structure of the variance. Weighted second order random walks are also available (see Fahrmeir & Lang (2001)) but are less suitable in the context of measurement error correction since they enforce too smooth function estimates.

In combination with flat priors on the initial parameters, the joint distribution of the vector of regression coefficients $\beta = (\beta_1, \ldots, \beta_K)'$ can be deduced from the random walk specifications as the multivariate Gaussian distribution

$$p(\beta|\tau_\beta^2) \propto \exp\left(-\frac{1}{2\tau_\beta^2}\beta' K \beta\right).$$

The precision matrix K is also derived from the univariate random walk priors. For a first order random walk it can be represented as $K = D'\Omega D$, where D is a first order difference matrix and $\Omega = \mathrm{diag}(\omega_2, \ldots, \omega_K)$ contains the knot distances as weights.

In case of smoothing without measurement error, cubic P-splines (i.e. splines of degree $l = 3$) with approximately 20 equidistant knots and second order random walk prior have proven to be a useful standard choice (Brezger & Lang 2006). However, exploratory simulations showed that this claim no longer holds when measurement error is present. In particular, the high degree of the spline basis and the second order random walk enforce smoothness of the function estimates. Since measurement error in general leads to an attenuation of functional relationships, i.e. functions appear smoother than under the true relationship, a suitable prior in measurement error correction has to allow for more flexibility. In addition, choosing equidistant knots has the disadvantage that the prior variance of the random walk remains constant over

the whole domain of the covariate. When correcting for measurement error, adaptive priors with more variability in areas where a lot of observations have been collected mostly showed a better performance. In summary, we found linear splines with 20 quantile-based knots and (weighted) first order random walk prior to be a suitable default choice for nonparametric smoothing of covariates with measurement error. Cheng & Crainiceanu (2009) also support the choice of linear splines in showing that the full conditionals both for the regression coefficients and the true covariate values are then log-concave.

On a further stage of the hierarchy, a hyperprior is assigned to the variance parameter τ^2_β to allow for a data-driven amount of smoothness. Since the random walk prior is multivariate Gaussian, a computationally attractive choice is the conjugate inverse gamma prior $\tau^2_\beta \sim \text{IG}(a,b)$ that leads to a simple Gibbs update for the variance parameter.

A further generalisation of the model can be achieved by allowing for prior uncertainty in the knot positions as in the adaptive spline smoothing approaches by Denison et al. (1998) or Biller (2000). However, in most situations it will be sufficient to either assign a smoothness prior to the regression coefficients (provided that the basis is sufficiently rich) or to allow for data-driven determination of the knot placements (see also the supporting simulation results in Brezger & Lang (2006)). In combination with measurement error correction we found it advantageous to fix the knot positions since this avoids additional re-evaluations when imputing the unobserved covariate values.

2.3.2 Measurement Error Priors

For the covariates with measurement error, a prior for the true covariate values has to be specified, since they will be treated as additional unknowns in the Bayesian inferential procedure. A flexible default choice is given by the Gaussian distribution

$$x_{ij} \sim \text{N}(\mu_{x,j}, \tau^2_{x,j}).$$

Assigning hyperpriors to the parameters such as $\mu_{x,j} \sim \text{N}(0, \tau^2_\mu)$ with τ^2_μ fixed at a large value and $\tau^2_{x,j} \sim \text{IG}(a,b)$ allows the prior to accommodate to a variety of data-generating processes. In particular, the prior for the expectation is essentially noninformative when assuming a large value for the hypervariance τ^2_μ.

Note that treating the true covariate values as unknown parameters is not only a computational trick to obtain a fully specified model within the MCMC sampler, but allows inferences to be drawn about the true covariate values. In particular, we obtain a sample from the posterior of the true covariate values allowing to investigate for example the precision of the correction or whether the true covariate value exceeds a certain threshold.

Finally, a prior may be assigned to the measurement error variances, if uncertainty about the $\tau^2_{u,j}$ has to be incorporated. In combination with the Gaussian contamination error, again an inverse gamma prior $\tau^2_{u,j} \sim \text{IG}(a,b)$ is a suitable default choice.

Note, however, that reliable estimation of $\tau^2_{u,j}$ will typically require a larger number of repeated measurements on the covariates, in particular in non-Gaussian observation models where the likelihood carries less information on the variability in measurement error than in Gaussian models. In our application, the measurement error variances are available from replication experiments. Therefore we will also restrict our attention to the case of known measurement error variances in our simulations.

3 Bayesian Inference

3.1 Posterior & Full Conditionals

Summarizing all unknown quantities in the vector θ and assuming conditional independence of the prior distributions, the joint posterior in our class of semiparametric models can be summarized as

$$\begin{aligned}
p(\theta|\text{data}) \propto\ & p(\text{data}|\beta_1,\ldots,\beta_r,\gamma,x_1,\ldots,x_r) && \text{observation model likelihood}\\
& \prod_{j=1}^{r_1} p(w_j|x_j,\tau^2_{u,j}) && \text{measurement error likelihood}\\
& \prod_{j=1}^{r_1} p(\tau^2_{u,j}) && \text{measurement error variance priors}\\
& \prod_{j=1}^{r_1} p(x_j|\mu_{x,j},\tau^2_{x,j})p(\mu_{x,j})p(\tau^2_{x,j}) && \text{true covariate value priors}\\
& \prod_{j=1}^{r} p(\beta_j|\tau^2_{\beta,j})p(\tau^2_{\beta,j}) && \text{nonparametric effect priors.}
\end{aligned}$$

The likelihood is derived under the assumption of conditional independence such that the complete data likelihood factorises to individual likelihood contributions. In case of exponential family regression, the likelihood contributions equal the corresponding exponential family densities evaluated at the predictor η_i. Assuming non-informative, random right censored survival times, the complete data likelihood contributions in hazard regression models with individual hazard rates $\lambda_i(t)$ are given by

$$L_i(\eta_i) = \lambda_i(t_i)^{\delta_i} \exp\left(-\int_0^{t_i} \lambda_i(u)du\right),$$

see Hennerfeind et al. (2006).

From the posterior, we can now derive the full conditional distributions for all unknowns to construct a Markov Chain Monte Carlo simulation algorithm. While Gibbs updates can be derived for several parameters, Metropolis-Hastings steps are necessary for the regression coefficients and the true covariate values. Since some of the priors involved in the model specification are (partially) improper, it is not obvious that the joint posterior will be proper (see for example Hobert & Casella (1996)). Fahrmeir & Kneib (2009) provide conditions for the propriety of the posterior in

semiparametric Bayesian regression models without measurement error that will be fulfilled in most practically situations and we expect these results to carry over to the case with measurement error.

The full conditional for a true covariate value x_{ij} depends only on the i-th likelihood contribution $L_i(\eta_i)$ to the observation model and the i-th contribution to the measurement error model. Combining this likelihood information with the relevant priors yields (up to an additive constant) the log-full conditional

$$\log(p(x_{ij}|\cdot)) = l_i(\eta_i) - \frac{1}{2\tau^2_{u,j}} \sum_{m=1}^{M} (w_{ij}^{(m)} - x_{ij})^2 - \frac{1}{2\tau^2_{x,j}} (x_{ij} - \mu_{x,j})^2$$

where $l_i(\eta_i) = \log(L_i(\eta_i))$ is the i-th log-likelihood contribution. Obviously this full conditional does not correspond to a known distribution since both the log-likelihood contributions and the B-spline basis functions are non-linear in the covariate values. Following Berry et al. (2002) we consider a random walk proposal for imputing the covariate values where, based on the current value x_{ij}^{curr}, a new value is proposed as

$$x_{ij}^{prop} = x_{ij}^{curr} + \varepsilon, \qquad \varepsilon \sim \mathrm{N}\left(0, \frac{4\tau^2_{u,j}}{M}\right).$$

The choice of the random walk variance as being proportional to the measurement error variance but inverse proportional to the number of replicated measurements balances between the more precise knowledge about the true value that can be gathered from repeated measurements on the one hand and uncertainty introduced by large measurement error variance. The constant factor 4 has proven to work well in practice, according to our experience, but can be adjusted by the user to adapt the acceptance probabilities if needed.

The Gaussian measurement error model in combination with the conjugate inverse gamma priors for the measurement error variances, yields full conditionals that are also inverse gamma, i.e.

$$\tau^2_{u,j}|\cdot \sim \mathrm{IG}\left(a + \frac{nM}{2}, b + \frac{1}{2}\sum_{i=1}^{n}\sum_{m=1}^{M}(x_{ij}^{(m)} - x_{ij})^2\right).$$

Similarly, we obtain closed form full conditionals for the true covariate value hyperparameters:

$$\mu_{x,j}|\cdot \sim \mathrm{N}\left(\frac{n\bar{x}_j\tau^2_\mu}{n\tau^2_\mu + \tau^2_{x,j}}, \frac{\tau^2_{x,j}\tau^2_\mu}{n\tau^2_\mu + \tau^2_{x,j}}\right)$$

$$\tau^2_{x,j}|\cdot \sim \mathrm{IG}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_{ij} - \mu_{x,j})^2\right)$$

where $\bar{x}_j$ is the empirical mean of the currently imputed true covariate values.

Finally, the full conditionals for the regression coefficients have to be derived. Again, these are not available in closed form since the likelihood is non-linear in the parameters (for non-Gaussian responses). Based on work by Gamerman (1997) in the context of random effects, Brezger & Lang (2006) propose to construct a Gaussian approximation to the full conditional by performing one-step of a Fisher scoring algorithm based on the current sample for β_j. More precisely, this leads to an iteratively weighted least squares (IWLS) proposal for β_j based on a Gaussian distribution with precision matrix and mean

$$P_j = X_j' W X_j + \frac{1}{\tau^2_{\beta,j}} K_j \qquad \text{and} \qquad m_j = P_j^{-1} X_j' W (\tilde{y} - \eta_{-j}),$$

where the diagonal matrix W and the vector of working observations $\tilde{y}$ are constructed in complete analogy to the usual GLM case (compare Fahrmeir & Tutz (2001)) and $\eta_{-j} = \eta - X_j \beta_j$ is the j-th partial residual. Similar expressions are obtained for the vector of fixed effects, compare Brezger & Lang (2006) for details. The rationale for the IWLS proposal mechanism is that it automatically adapts to the location and the curvature of the corresponding full conditional thereby avoiding the necessity of manually tuning the MCMC sampler. Hennerfeind et al. (2006) describe similar proposal schemes for hazard regression models. The full conditional of the smoothing parameters $\tau^2_{\beta,j}$ is again inverse Gamma with updated parameters, i.e.

$$\tau^2_{\beta,j} | \cdot \sim \text{IG}\left(a + \frac{1}{2}\text{rank}(K_j), b + \frac{1}{2}\beta_j' K_j \beta_j\right).$$

3.2 Implementational Details & Software

Though a Metropolis-Hastings sampler can immediately be set up based on the full conditional distributions and proposals described in the previous section, an efficient implementation requires careful fine-tuning at several places. This is particularly the case for nonparametric function estimation involving a large number of regression coefficients and the measurement error correction problem, where the data, and therefore also the design matrices, change in each iteration. A naive implementation in a general specification language for Bayesian modeling such as WinBUGS or in a high-level interpreted programming language such as R would therefore be inefficient. As a consequence, we implemented our methodology as a part of the software package BayesX (`http://www.stat.uni-muenchen.de/~bayesx`, Brezger et al. (2005)), which has been specifically designed for the estimation of semiparametric regression models. The computational kernel is implemented in C++, allowing for an efficient treatment of loop-intensive MCMC simulations. A graphical user interface provides convenient access to the methodology and allows for a flexible model specification.

Table 1 Impact of rounding on computing times (in minutes) in the different simulation scenarios.

Digits	1	2	3	4	5
scenario (a)	3:11	7:11	9:56	10:03	10:28
scenario (b)	3:25	7:47	10:22	10:41	10:53
scenario (c)	6:12	14:24	19:41	20:14	20:21
scenario (c')	5:24	13:50	19:12	20:17	20:25

The computational bottleneck is simulating the regression coefficients of the penalized splines, in particular for the covariates measured with error. The first difficulty arises from the fact that for simulating from a K-dimensional multivariate Gaussian distribution, a K-dimensional system of equations has to be solved in each iteration. Replacing the simultaneous update with a single move algorithm would speed up computation but comes at the price of deteriorated mixing and convergence due to the ignored correlation of the elements in β_j. We therefore make use of sparse matrix computations, since the precision matrix P_j is a band matrix, see Rue (2001) and Brezger & Lang (2006) for details. This approach has the advantage to provide fast computations while keeping the correlation information included in the proposals.

The second difficulty is specific to the imputation of true covariate values: In each iteration new values are sampled, requiring the re-evaluation of the design matrix X_j. To shorten computation times, we consider two tricks: Firstly, instead of storing the complete design matrix, we only store the relevant part of it. Note that B-splines form a local basis such that in each row of X_j there are only $l+2$ non-zero entries (where l denotes the degree of the spline). Since we chose $l = 1$ as the standard in measurement error correction, there are actually only three values to be stored instead of K which is typically in the range of 20 to 40. Furthermore, only rows of the design matrix corresponding to distinct observed values of x_j have to be stored in combination with an index vector associating the observations with the different values for x_j. This storage scheme allows for a further reduction of computing times in a second step: Instead of storing the exact covariate values in double precision, we round them to a user-specified number of decimal places. As a consequence, several formerly distinct covariate values now coincide so that only a smaller number of rows of X_j has to be stored and re-computed in each iteration. In our simulations and applications we used two decimal places, a choice that lead to only negligible changes in the results while making a significant change in computing times in exploratory analyses. Table 1 provides some exemplary results for different decimal places and the simulation scenarios considered in the following section. There obviously is a tremendous gain in computing times for small decimal places, while computing times level off when using a large precision corresponding to almost no rounding.

Note also that due to the modular structure of MCMC algorithms, computing time only grows linearly when, for example, increasing the number of covariates subject to measurement error. Hence, computations with two covariates measured with error take approximately twice as long as computations with one covariate, which is in contrast to approaches where a decomposition of the correction problem in separate sub-problems is not feasible.

4 Simulations

4.1 Simulation Setup

To assess the properties of the proposed measurement error correction scheme and the validity of our implementation, we performed an extensive simulation study investigating model scenarios of increasing complexity:

(a) One covariate with measurement error:

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_i) + v_i\gamma, \\ \text{Measurement error model:} & w_i|x_i \sim \mathrm{N}(x_i, 1), \\ \text{Further settings:} & x_i \sim \mathrm{N}(0,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

(b) One covariate measured with error in combination with a further nonparametric effect:

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_{i1}) + x_{i2}^2 + v_i\gamma, \\ \text{Measurement error model:} & w_i|x_i \sim \mathrm{N}(x_i, 1), \\ \text{Further settings:} & x_{i1} \sim \mathrm{N}(0,1),\, x_{i2} \sim \mathrm{U}(-1,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

(c) Two covariates measured with error

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_{i1}) + 0.2x_{i2}^2 + v_i\gamma \\ \text{Measurement error model:} & w_{i1}|x_{i1} \sim \mathrm{N}(x_{i1}, 1),\, w_{i2}|x_{i2} \sim \mathrm{N}(x_{i2}, 0.64), \\ \text{Further settings:} & x_{i1} \sim \mathrm{N}(0,1),\, x_{i2} \sim \mathrm{N}(0,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

Model (a) is the most simple one, where only one covariate is measured with error and the predictor contains only one single additional parametric covariate. In model (b), a second nonparametric effect is added to the predictor, but the corresponding covariate is observed exactly. Finally, in scenario (c), the covariate associated with the second nonparametric effect is also measured with error. Since scenario (c) is the most demanding one, we re-ran it with two replicated measurements on each of the covariates x_1 and x_2 to get an idea of the performance improvement by repeated observations on the measurement equation:

(c') Two covariates measured with error in two replications

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_{i1}) + 0.2x_{i2}^2 + v_i\gamma \\ \text{Measurement error model:} & w_{i1}^{(m)} \sim \mathrm{N}(x_{i1}, 1),\, w_{i2}^{(m)} \sim \mathrm{N}(x_{i2}, 0.64),\, m = 1,2, \\ \text{Further settings:} & x_{i1} \sim \mathrm{N}(0,1),\, x_{i2} \sim \mathrm{N}(0,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

For each of the scenarios, we simulated data sets with responses from the following four types of responses:

(a) Binomial distribution with three replicated binary observations, i.e. $y_i \sim \mathrm{B}(3, \pi_i)$, $\pi_i = \exp(\eta_i)/(1+\exp(\eta_i))$.

(b) Binomial distribution with ten replicated binary observations, i.e. $y_i \sim \mathrm{B}(10, \pi_i)$, $\pi_i = \exp(\eta_i)/(1+\exp(\eta_i))$.
(c) Poisson distribution, i.e. $y_i \sim \mathrm{Po}(\lambda_i)$, $\lambda_i = \exp(\eta_i)$.
(d) Exponentially distributed duration times $T_i \sim \mathrm{Exp}(\lambda_i)$, $\lambda_i = \exp(\eta_i)$ subject to independent uniform censoring $C_i \sim \mathrm{U}(0, 50)$ resulting in an average censoring rate of 10%. The observed data is given by $t_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}(T_i \leq C_i)$.

For each response and each scenario, the sample size was fixed at $n = 500$ and the number of simulation replications was given by 100.

To benchmark the performance of the correction method, we did not only consider estimates from the imputation scheme, but also estimates based on the true covariate values and naive estimation based on the average of the measurements with error:

(a) Exact estimation: Use the true covariate values x_{ij} in the estimation procedure.
(b) Naive estimation: Use the average of repeated measurements $\bar{w}_{ij} = \sum w_{ij}^{(m)}/M$ as covariate.
(c) Corrected estimation: Impute the estimated true covariate values with MCMC.

The results from the exact and the naive estimation approach can serve as an upper and a lower bound for the performance of the corrected results.

4.2 Simulation Results

Figures 1 visualizes average estimates for the sine curve in scenario (a). As expected, the estimated curve in the naive approach is far too flat and almost equals a linear fit. In contrast, using the true covariate values leads to a satisfactory reproduction of the curve over a large part of the covariate domain. Note that only a very small number of observations is located outside the interval $[-2, 2]$ and therefore the deterioration of the average estimates in this area is simply due to a lack of data. MCMC-based measurement error correction falls in between the naive and the exact estimation results but indeed shows considerable correction. This becomes even more obvious from considering the MSEs (Figure 2), where the corrected results clearly outperform results from naive estimation. The improvement is smallest in the case of a binomial response with only three replications, where not too much information from the likelihood is available. For all other types of responses with increased likelihood information, the correction improves and the MSEs are closer to exact information than when using naive estimation.

When including an additional nonparametric effect to the model, results for the sine curve actually remain practically the same and are therefore not presented. To assess the impact on the effects without measurement error, Figure 3 shows boxplots of the MSE for binomial responses with ten replications and for survival times. Obviously there is some impact of measurement error also on the effects of covariates observed exactly but the change is much smaller compared to the effect on the sine curve. The most significant change is observed for survival times, and in this case MCMC-based imputation also yields more correction than for Binomial responses.

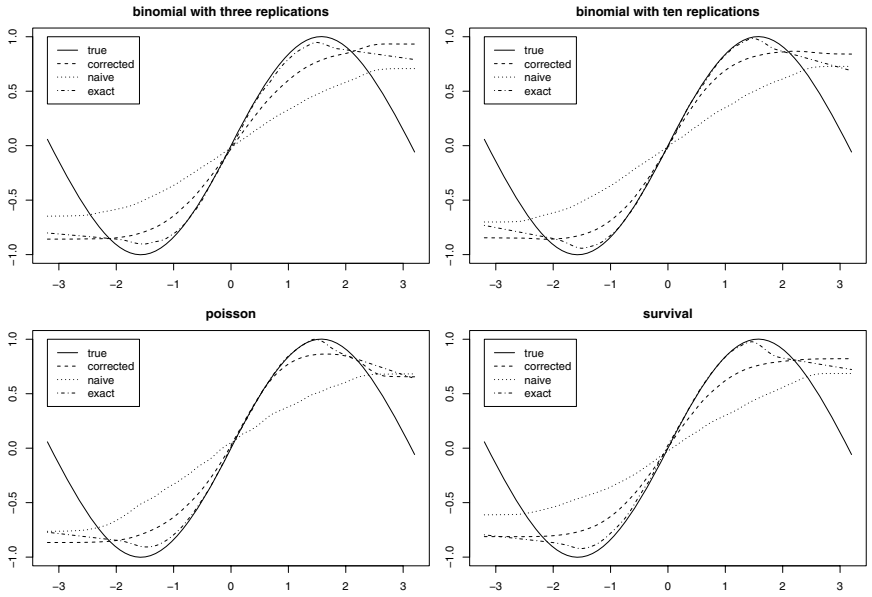

Fig. 1 Average function estimates for $\sin(x)$ for all four response types in scenario (a).

When considering two covariates with measurement error (Figure 4), results remain qualitatively the same as with one covariate: Quality of the estimates considerably increases when applying the proposed correction scheme with larger impact in case of response types with more information. Note, that the signal to noise ratio is smaller for the quadratic functions than for the sine curve and therefore correction is generally smaller for x_2 in terms of the bias although comparable improvements are achieved in terms of MSE. When including a second replication on the covariates measured with error, results improve even further (although of course also the results from the naive approach improve). In this case (Figure 5), the corrected estimates even start to indicate the local minimum and maximum of the sine curve, although the data in this area already get quite sparse. Similarly, the reproduction of the square function is now very close to the true function. In addition, the boxplots indicate that the corrected estimates perform almost as well as the estimates obtained with the true covariate values.

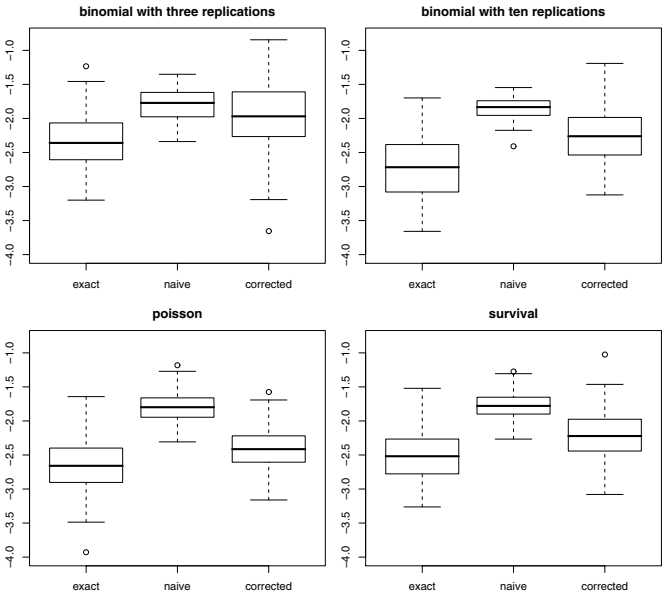

Fig. 2 Boxplots of log(MSE) for $\sin(x)$ for all four response types in scenario (a)

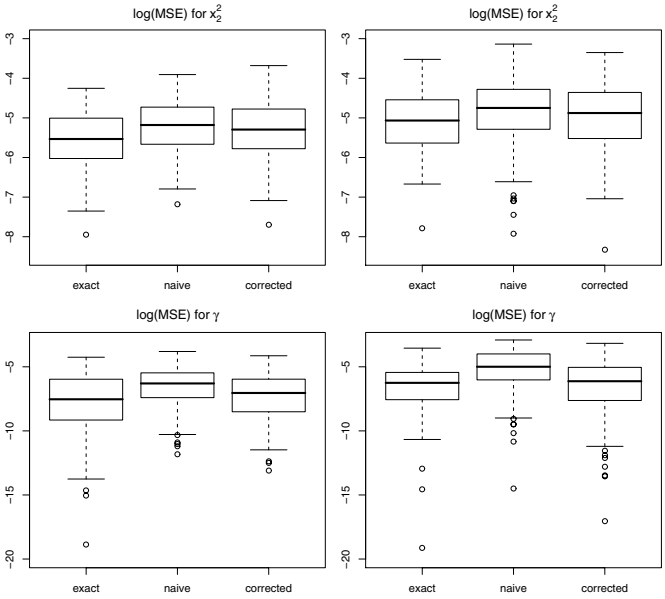

Fig. 3 Boxplots of log(MSE) for x_2^2 and γ for two response types in scenario (b) (binomial with ten replications in the left panel, survival in the right panel).

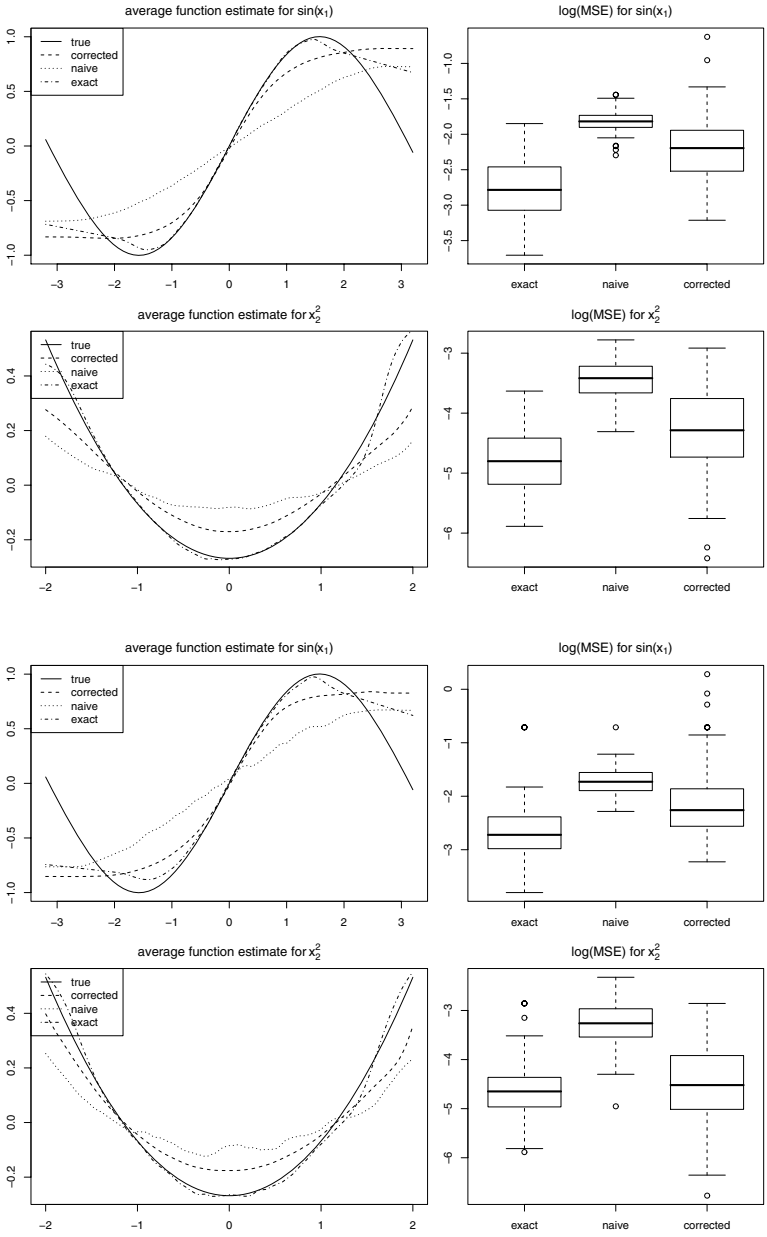

Fig. 4 Average estimates and boxplots for two response types in scenario (c) (binomial with ten replications in the upper two rows, poisson in the lower two rows).

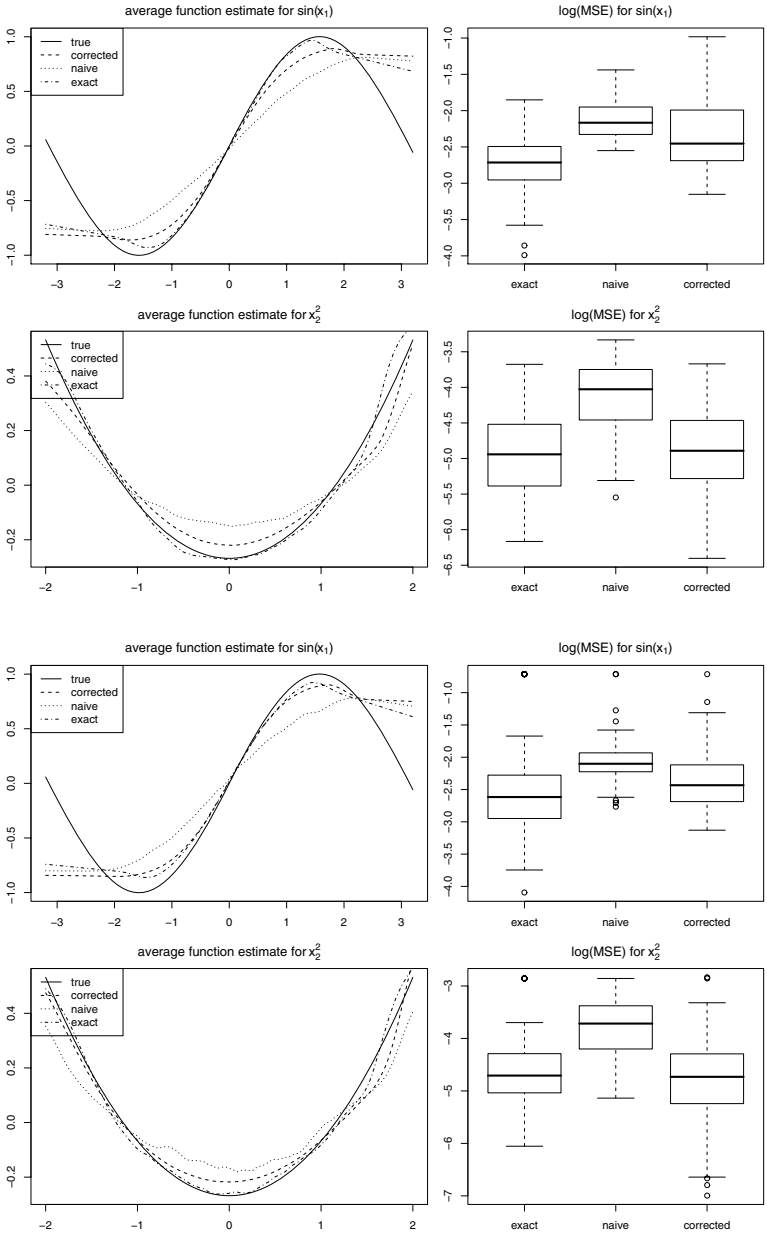

Fig. 5 Average estimates and boxplots for two response types in scenario (c') (binomial with ten replications in the upper two rows, poisson in the lower two rows).

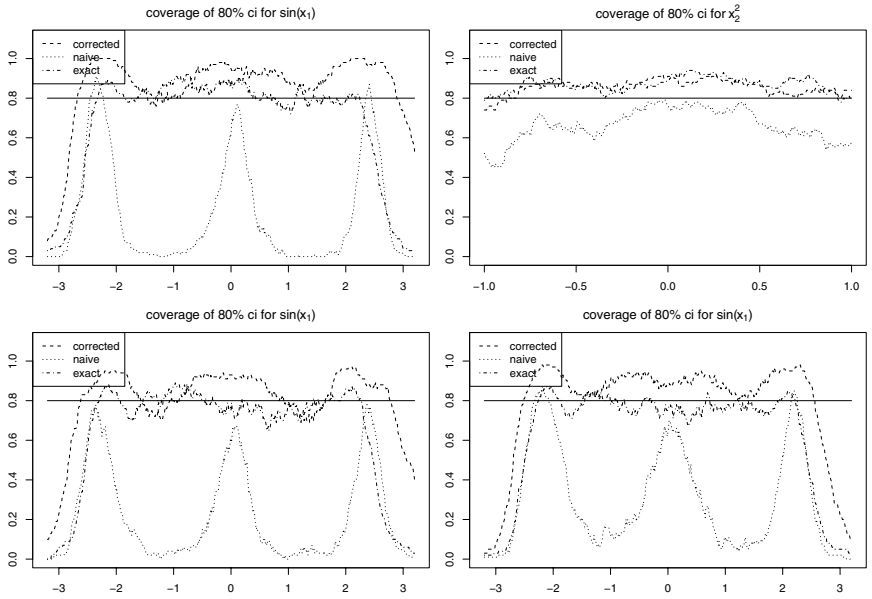

Fig. 6 Average coverage probabilities of 80% credible intervals for different effects in scenario (a) (upper left), scenario (b) (upper right), scenario (c) (lower left) and scenario (c') (lower right).

Finally, Figure 6 visualizes average coverage probabilities for different effects in the four scenarios. Again we find the impact of flattened estimation when using the naive approach: The empirical coverages are far too low not only for the effects of covariates measured with error but also for the square function in scenario (b). In contrast, the coverages of the corrected estimates are on average close to the nominal value all over the relevant covariate domain. Only at the boundaries, where data become sparse, the empirical coverage decreases. Note also, that using the true covariate values actually leads to somewhat too conservative credible intervals – an artefact that is found frequently in the context of Bayesian credible intervals.

In summary, our simulations allow the following conclusions to be drawn:

- MCMC-based imputation of the true covariate values allows to correct for the adverse effects of covariates measured with error. The correction effect is particularly expressed for the nonparametric effects of the covariates with measurement error while the amount of correction varies for effects of covariates observed exactly.
- In our simulation, measurement error had the expected impact on naive nonparametric regression results, i.e. nonparametric effects are underestimated and far too smooth.
- Ignoring measurement error also has dramatic impact on the coverage properties of the credible intervals.

We confirmed our findings in a second simulation study with smaller measurement error variances with practically the same results (not shown).

5 Incident Heart Failure in the ARIC Study

Our proposed methodology was motivated by the analysis of time to event data from the Atherosclerosis Risk in Communities (ARIC) study. ARIC is a large multipurpose epidemiological study conducted in four US communities (Forsyth County, NC; suburban Minneapolis, MN; Washington County, MD; and Jackson, MS). From 1987 through 1989, 15,792 male and female volunteers aged 45 through 64 were recruited from these communities for a baseline and three subsequent visits. The baseline visit (visit 1) included at-home interviews, laboratory measurements, and clinic examinations. The study participants returned for additional visits in 1990-92 (visit 2), 1993-95 (visit 3), and 1996-98 (visit 4). Details of the enrollment process and the study procedures are fully described by The ARIC INVESTIGATORS (1989).

Time to event data is observed continuously for multiple end points, but we focus here on the event *detection of heart failure* (HF), the inability of the heart to pump blood with normal efficiency. After exclusion of 752 participants with prevalent heart failure, 14,857 ARIC study participants were followed for incident heart failure hospitalization or death from 1987 to 2002. During a mean follow-up of 13.2 years, 1,193 participants developed HF (Kottgen et al. 2007).

The relationship between various risk factors, such as race, age or sex, and progression time to heart failure may be confounded by a series of baseline covariates. Two such important confounders are the baseline systolic blood pressure (SBP) and the baseline kidney function as measured by the glomerular filtration rate (GFR). Both SBP and GFR are measured with moderate error and their corresponding dose/response functions are expected to be non-linear. Taking into account these features of the data is necessary for satisfactory inference and can be handled using the methodology and software introduced in this paper. A reasonable approach to statistical modeling of the present data is to consider a survival model for time to heart failure with the following log-hazard function

$$\log\{\lambda_0(t)\} + f_1\{\log(\text{SBP} - 50)\} + f_2\{\log(\text{GFR})\} + \gamma_1 \text{sex} + \gamma_2 \text{AA} + \gamma_3 \text{age}, \quad (2)$$

where $\lambda_0(t)$ is the baseline hazard, $f_1(\cdot)$ and $f_2(\cdot)$ are unspecified smooth functions modeled as penalized splines, sex is a $0/1$ variable with 1 corresponding to males, AA is a $0/1$ variable with 1 corresponding to African Americans, and age being the baseline age. For $f_1(\cdot)$ and $f_2(\cdot)$ we used degree 1 penalized B-splines with 30 equidistant knots. We also employed quantile based knots but found that they produce very wiggly estimates both with and without measurement error correction in this example. This is probably due to the concentration of observations in a smaller part of the domain, that is more prevalent in the large data set of the application than in the comparable small simulation data sets.

Table 2 Corrected and naive posterior mean estimates, and 80% credible intervals for the parametric effects

	corrected			naive		
	$\hat{\gamma}$	80% ci		$\hat{\gamma}$	80% ci	
intercept	-8.577	-9.264	-7.938	-8.861	-9.402	-8.314
male	0.419	0.341	0.495	0.421	0.340	0.506
african american	0.355	0.254	0.451	0.350	0.262	0.440
age at first visit	0.083	0.075	0.091	0.081	0.074	0.089

In model (2), SBP represents the true long term average SBP and GFR represents the true filtration rate of the kidney at the time it was measured. Both variables are measured with error and replication studies are used to estimate the variance of the error process. To obtain the measurement error variance of $\log(\text{SBP} - 50)$ we use a replication study from the Framingham Heart Study described in Carroll et al. (2006), pages 112-114. In short, the Framingham study consists of a series of exams taken two years apart. The estimated measurement error using exams 2 and 3 is $\hat{\tau}^2_{\text{SBP}} = 0.01259$, which in the ARIC study corresponds to a reliability of 81%. Thus, in our model $\log(\text{SBP} - 50)$ is the true long term average $\log(\text{SBP} - 50)$ over a 2 year period.

There are important technical differences between measuring blood pressure with a sphygmomanometer and measuring the filtration rate of the kidney. Indeed, GFR can only be obtained through a long and awkward procedure that is impractical for routine analyses, as required by medical practice and large epidemiological studies. Instead, the estimated GFR (eGFR) is used in practice and is obtained from a prediction equation based on creatinine, gender and age (Hsu et al. (2005), Kottgen et al. (2007), Cheng & Crainiceanu (2009)). More precisely, the eGFR is predicted from the following equation:

$$\text{eGFR} = 186.3 * (\text{Serum Creatinine})^{-1.154} * (\text{Age})^{-0.203} * (0.742)^{(1-sex)} * (1.21)^{(\text{AA})}.$$

Thus, the eGFR measurement contains at least two non-ignorable sources of error: 1) the biological variability unaccounted for by the prediction equation; and 2) the laboratory variability associated with urine serum creatinine. To assess the variability of eGFR, a replication study was conducted in the Third National Health and Nutrition Examination Survey (NHANES III). Duplicate eGFR measurements were obtained for each of 513 participants aged 45 to 64 with eGFR ≥ 60 from two visits at a median of 17 days apart (Coresh et al. 2002). We assumed a classical measurement error model for log(eGFR) and calculated the measurement error variance as $\hat{\tau}^2_u = \frac{1}{2}\sum_{i=1}^{513}(w_{i1} - w_{i2})^2$, where w_{im} is the observed log(eGFR) for subject i at visit m. The estimated measurement error variance was $\hat{\tau}^2_u = 0.009743$ corresponding to a reliability of 0.80 in the ARIC data set and will be treated as a constant in our subsequent analyses.

Figure 7 and Table 2 summarize the results of both a naive and a measurement error corrected analysis. While the estimated baseline hazard rate remains practically

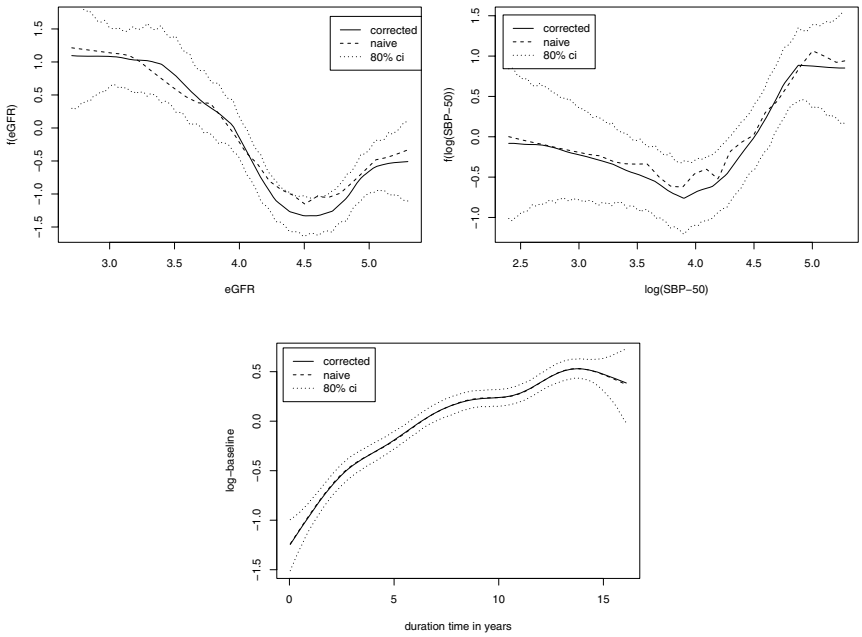

Fig. 7 Corrected and naive posterior mean estimates for the nonparametric effects of eGFR and log(SBP-50), and the log-baseline hazard rate with 80% pointwise credible intervals.

unchanged when correcting for measurement error, there are obvious changes in the results for SBP and eGRF. In particular, the local minima at 4.5 (eGFR) and 4.0 (SBP) are underestimated due to oversmoothing in the naive analysis. This effect is expressed more clearly for eGFR where the reliability is smaller and therefore the (relative) measurement error is larger.

Since the data set of the application is much larger than the data sets employed in the simulation, it is also worthwhile to consider the impact of rounding on the computing times again. With two valid decimal places, the corrected analysis (28,000 MCMC iterations on a dual core processor PC with 3Ghz CPU) including the imputation for two covariates took about 99 minutes, which is very competitive taking the complexity of the model and size of the data set into account. When increasing the number of valid decimal places, computing times increase to 215 minutes for 4 decimal places with visually indistinguishable results.

6 Summary

We have introduced a flexible Bayesian imputation scheme for correcting for measurement error in a large class of semiparametric regression models including models for the expectation in exponential family regression and models for the hazard rate in the case of survival data. The model specification permits quite flexible structures involving several nonparametric effects and several covariates measured with error. A variety of situations has been studied in a simulation study, indicating that the proposed algorithm works well even in complicated settings. The approach has been implemented in a user-friendly and efficient software package, allowing for easy access to the new methods. Moreover, the software supports a number of extended modeling possibilities not considered in this paper. To be more specific, varying-coefficient terms, interaction surfaces, spatial effects, or time-varying effects in survival can be augmented to the model specification if needed. This large flexibility of the model class is available due to the modular structure of MCMC simulations that makes all modeling components introduced previously to Bayesian semiparametric regression readily available as components in the measurement error correction approach.

A frequent drawback of approaches based on MCMC simulations are long computation times and difficulties in mixing and convergence. We circumvent both by considering a specialized implementation that relies on numerically fast sparse matrix computations in combination with efficient storage and rounding schemes. In addition, MCMC makes model combinations accessible that would require quite involved methodological treatment and computations in a frequentist approach.

Acknowledgements The authors thank Thomas Augustin and Ludwig Fahrmeir for valuable discussions at various stages of preparing this paper. The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, and N01-HC-55022. The authors thank the staff and participants of the ARIC study for their important contributions. The support for Ciprian Crainiceanu was provided by contracts N01-HC-55020 and R01-DK-076770-01.

References

The ARIC INVESTIGATORS (1989). The Atherosclerosis Risk in Communities (ARIC) study: design and objectives, *American Journal of Epidemiology* **129**: 687—702.

Berry, S. M., Carroll, R. J. & Ruppert, D. (2002). Bayesian Smoothing and Regression Splines for Measurement Error Problems, *Journal of the American Statistical Association* **97**: 160–169.

Biller, C. (2000). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models, *Journal of Computational and Graphical Statistics* **9**: 122–140.

Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: Analysing Bayesian structured additive regression models, *Journal of Statistical Software* **14**: (11).

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

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models - a modern perspective (2nd edition)*, Chapman & Hall / CRC, New York.

Cheng, Y.-J. & Crainiceanu, C. M. (2009). Cox models with smooth functional effects of covariates measured with error, *Journal of the American Statistical Association*, to appear.

Coresh, J., Astor, B., McQuillan, G., Kusek, J., Greene, T., Van Lente, F. & Levey, A. (2002). Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate, *American Journal of Kidney Diseases* **39**: 920—929.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). Automatic Bayesian curve fitting, *Journal of the Royal Statistical Society Series B* **60**: 333–350.

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. (2009). Propriety of Posteriors in Structured Additive Regression Models: Theory and Empirical Evidence, *Journal of Statistical Planning and Inference* **139**: 843–859.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer, New York.

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

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

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

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*, Chapman & Hall / CRC, Boca Raton.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models, *Journal of the American Statistical Association* **101**: 1065–1075.

Hobert, J. P. & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, *Journal of the American Statistical Association* **91**: 1461–1473.

Hsu, C. C., Kao, W. H., Coresh, J., Pankow, J. S., Marsh-Manzi, J., Boerwinkle, E. & Bray, M. S. (2005). Apoliprotein E and progression of chronic kidney disease, *Journal of the American Medical Association* **293**: 2892–2899.

Jullion, A. & Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models, *Computational Statistics & Data Analysis* **51**: 2542–2558

Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics* **34**: 207–228.

Kottgen, A., Russell, S. D., Loehr, L. R., Crainiceanu, C. M., Rosamond, W. D., Chang, P. P., Chambless, L. E. & Coresh, J. (2007). Reduced kidney function as a risk factor for incident heart failure: The Atherosclerosis Risk in Communities (ARIC) study, *Journal of the American Society of Nephrology* **18**: 1307–1315.

Richardson, S. & Gilks, W. R. (1993). Conditional independence models for epidemiolgoical studies with covariate measurement error, *Statistics in Medicine* **12**: 1703–1722.

Rue, H. (2001). Fast sampling of Gaussian Markov Random Fields with Applications, *Journal of the Royal Statistical Society B* **63**: 325–338.

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

Stephens, D. A. & Dellaportas, P. (1992). Bayesian analysis of generalised linear models with covariate measurement error, *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. & Smith, A. F. M. (eds), *Bayesian Statistics 4*, Oxford University Press.

Wang, C.-Y. & Pepe, M. S. (2001). Expected estimating equations to accomodate covariate measurement error, *Journal of the Royal Statistical Society B* **62**: 509–524.

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