Correction for covariate measurement error in nonparametric regression

David Rummel

Institut für Statistik München

GSF – National Research Center for Environment and Health
member of the Helmholtz Association

München, 2006

Correction for covariate measurement error in nonparametric regression

David Rummel

Dissertation
zur Erlangung des Grades Doctor oeconomiae publicae
(Dr. oec. publ.)
an der Ludwig–Maximilians–Universität, München

vorgelegt von
David Rummel
2006

Institut für Statistik, Fakultät für Mathematik, Informatik und Statistik
In Zusammenarbeit mit der GSF - Forschungszentrum für Umwelt und Gesundheit, Leiter Prof. Dr. Dr. H.Erich Wichmann

Referent: Prof. Dr. Thomas Augustin

Korreferent: Prof. Dr. Helmut Küchenhoff

Promotionsabschlussberatung: Mittwoch, den 26. Juli 2006

Vorwort

Die vorliegende Arbeit umfasst meine wissenschaftliche Tätigkeit am Department für Statistik an der Ludwig-Maximilians-Universität München. Dort bin ich seit Mai 2003 als Mitarbeiter am Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen" im Teilprojekt C2 angestellt und werde somit durch Mittel der Deutschen Forschungsgemeinschaft (DFG) gefördert. Sowohl für diese finanzielle Unterstützung, als auch für die Gelegenheit in diesem besonderen Klima forschen zu können, bin ich dankbar.

In erster Linie bedanke ich mich bei meinen Doktoreltern, oder vielmehr Doktorvätern, Prof. Dr. Thomas Augustin und Prof. Dr. Helmut Küchenhoff. Sie haben sowohl durch ihre fachliche Unterstützung, als auch mit ihrem Respekt vor meinen Forschungsinteressen, einen wesentlichen Anteil an dieser Arbeit. Als professorale Ideengeber möchte ich hier außerdem Prof. Dr. Ludwig Fahrmeir und Prof. Dr. Leonhard Held nennen.
Dr. Angela Döring und Prof. Dr. Dr. H.Erich Wichman von der GSF - Forschungszentrum für Umwelt und Gesundheit (Neuherberg) danke ich für die freundliche Genehmigung zur Analyse der MONICA-Daten (Kapitel 4.3).

Besonders hilfreich war die entspannte und offene Atmosphäre im Institut, wofür vor allem meine Kollegen verantwortlich sind. Für technischen, organisatorischen und sonstigen Support sage ich im besonderen Brigitte Maxa und Christa Jürgensonn ein herzliches 'Dankeschön'.

An letzter, aber prominenter Stelle denke ich an meine Familie – meinen Bruder, meine Eltern und meine Großeltern. Seit kurzem habe ich eine eigene Familie, die mich liebevoll und rücksichtsvoll unterstützt: süße Stefanie und kleine Antonia Leila, ich hab' euch lieb!

München, den 7.November 2006 *David Rummel*

Zusammenfassung

In vielen Anwendungsgebiete der Statistik wird man zunehmend aufmerksam auf das Problem messfehlebehafteter Variablen und die Problematik einer angemessenen Analyse. Das schlichte Ignorieren dieser Fehler führt in vielen Fällen, wie zum Beispiel in der Regression mit fehlerbehafteten Kovariablen, zu verzerrten Schätzungen. Während das Problem für die parametrische Regression ausführlich diskutiert wurde, gibt es nur wenige Vorschläge zur Korrektur der nonparametrischen Regression. Die vorhandenen Ansätze sind leider oft computerintensiv oder wenig effektiv.

In Rahmen dieser Arbeit werden verschiedene neue Methoden entwickelt, die zum Teil die Effektivität von bestehenden 'state-of-the-art' Korrekturverfahren besitzen, gleichzeitig aber nur einen Bruchteil deren Rechenzeit beanspruchen. Diese neuen Methoden verwenden hauptsächlich die sogenannte Relevance Vector Machine (RVM) zur nichtparametrischen Regression - allerdings nun erweitert um Messfehlerkorrekturideen aus der Regressionskalibrierung, dem sogenannten SIMEX und dem Markov Chain Monte Carlo (MCMC) Korrekturansatz. Ausführliche Simulationsstudien mit gausschen, binären und poissonverteilten Responsevariablen vergleichen die Methoden untereinander. Es wird weiterhin der Fall mehrerer messfehlerbehafteter Kovariablen berücksichtigt.

Für den in der Epidemiologie besonders relevanten Fall von binären Longitudinaldaten wird außerdem ein MCMC-Ansatz zur nichtparametrischen Modellierung dieser Daten unter Berücksichtigung von Kovariablenmessfehlern vorgestellt.

Abstract

Many areas of applied statistics have become aware of the problem of measurement error-prone variables and their appropriate analysis. Simply ignoring the error in the analysis usually leads to biased estimates, like e.g. in the regression with error-prone covariates. While this problem has been discussed at length for parametric regression, only few methods exist to handle nonparametric regression under error, which are usually either computer intensive or little effective.

This thesis develops new methods achieving the correction quality of state of the art methods while demanding only a trickle of their computing time. These new methods use the so-called relevance vector machine (RVM) for nonparametric regression - now enhanced by correction methods based on the ideas of regression calibration, the so-called SIMEX and Markov Chain Monte Carlo (MCMC) correction. All methods are compared in simulation studies regarding Gaussian, binary and Poisson responses. This thesis also discusses the case of multiple error-prone covariates.

Furthermore, a MCMC based correction method for nonparametric regression of binary longitudinal data with covariate error is introduced. This data scenario is often encountered, e.g. in epidemiological applications.

Contents

Chapter 1

Introduction and overview

By 1799, the German mathematician Carl Friedrich Gauss (1777-1855) had developed a completely new technique for fitting an equation to a set of data points: the least squares method. Based on model assumptions about the true state of a system and empirical observations from the system, he succeeded in predicting the re-appearance of the planetoid Ceres after its path hid Ceres behind the sun in 1801. This was accomplished by the assumption of elliptical orbits and a few observations made by the Italian astronomer Giuseppe Piazzi (1746-1826).
This strong instrument of least squares will also be applied in this work for making predictions in the context of regression analysis.
The term 'regression', however, was not established until 1877. It was Sir Francis Galton (1822-1911), who derived and applied linear regression to problems of heredity by e.g. examining characteristics of the sweet pea plant. He performed regression by using the medians of the dependent variable given a grouped independent variable and fitting a line by eye.
Finally, Karl Pearson (1857-1936) and George Udny Yule (1871-1951) linked Galton's regression to Gauss's least squares method and thus released regression from its association with Galton's work. Subsequently, statistical regression entered into a variety of fields.
It was in 1908, when Louis Bauer (1865-1932), an American geomagnetist,

criticized Gauss's spherical harmonic regression model of the terrestrial geomagnetic field: there was no realistic (physical) interpretation of the parameters beyond the first three of the 24 parameters used in this model. One could hardly imagine that roughly 80 years later an area of statistics gains ground, which completely gave up the interpretability of parameters.

The so-called flexible regression methods, which are a subclass of nonparametric methods, completely abandon the interpretability of parameters. On the other hand, they allow for an appropriate analysis of observations that have been generated from a complex data generating process. The flexible regression methods, therefore, relax the strong, yet common, model assumption of the true process being of linear or quadratic form. Consequently, their predictions are rather driven by the observed data than by model assumptions. On that account, it is important to keep the balance between just fitting the particular observations and the ability to generalize the results of the analysis to a predication about the true underlying process.

The application of flexible regression methods for the estimation of complex true functions is a core point in the present work. However, it is only one aspect of relaxing the strong assumptions typically inherent in statistical analysis. The main achievement of this thesis is the development of correction methods for covariate measurement error, which help to improve flexible regression analysis.

It is common assumption in the regression context that the covariates can be measured exactly. Hence, any observational error of the covariates is ignored. However, measurement error is particularly plausible when thinking of studies in the field of economic and social sciences applying statistics to hardly measurable constructs, e.g. motivation to work, personality and self-confidence. Other areas of application include medicine and epidemiology, where effects of (lifelong) risk factors are of interest. But, how should e.g. long term nutrition habits, like fat intake, be observed – are food diaries a reliable and valid measure here? What specific information can be extracted

from air pollution measurements in a city about one's individual exposure to that pollutants?
A general source of measurement error are retrospective surveys. That is, because most people tend to forget e.g. what time they spent watching TV during the last week and thus can merely give a rough guess.
While measurement error in the response variable is usually much easier to handle and not of interest here, accounting for covariate measurement error yields far more complex models and standard analysis is typically not available. For that reason, temptation is great to ignore it. Small error of course affects the analysis only to a minor degree and accurate data collection, if possible, would entirely displace the need for correction. However, the results obtained under violated assumptions are invalid.
Even then, when the data are suspected to contain covariate measurement error, taking that fact into account may be profitable. By comparing the results from the analysis ignoring measurement error with the results when accounting for that error, one can judge whether this suspicion was justified or not. Both results only differ, if covariate measurement error is really an issue in the present case.

The following paragraphs present how this thesis proceeds in developing correction methods for flexible regression. Chapter 2 kicks off with a rather general description of the considered methods and the new developments of the present work. Chapter 3, 4 and 5 contain the details of these approaches for flexible Gaussian, binary and Poisson regression, respectively. Chapter 6 concludes the work and highlights some aspects of future research.
The course of each chapter is summarized and the main points are explained briefly.

Chapter 2 comprises a description of the main techniques employed in this work and its structure is tripartite: firstly, flexible regression using the 'relevance vector machine' (RVM) is introduced. Tipping (2000) presents this Bayesian concept of flexible regression where a set of weighted so-called basis

functions is fitted to the data. Most notably, this method performs data driven selection of relevant basis functions from an arsenal of arbitrarily many basis functions and typically finds an extremely parsimonious model representation. An important point for later error correction is presented here; it lies in the equivalence of two parameter estimation methods based on different paradigms: the specific method of the Bayesian posterior mode estimation applied in the RVM is identical to Fisher scoring in a frequentistic penalized likelihood setting.
Secondly, a Markov Chain Monte Carlo (MCMC) version of the relevance vector machine for binary regression is introduced. It is inspired by an approach by Chakraborty, Gosh & Mallick (2005), but additionally implements Bayesian model averaging as used by Denison, Holmes, Mallick & Smith (2002). A brief introduction to MCMC methodology prepares this development. Including Bayesian sampling methods in the data analysis will prove to be very fruitful in complex data situations.
Finally, the covariate measurement error problem is described. This problem has generated major research interest, driven by the growing awareness of the adverse effects of that error on the statistical analysis (cf. Fuller (1987), Carroll, Ruppert & Stefanski (1995)).
The core ideas of the correction methods developed in this thesis are briefly motivated in this chapter in order to elaborate how these strategies are connected and in which respects they differ. Details are, however, deferred to the following chapters.

Chapter 3 comprises the full details on the developed error correction methods in flexible Gaussian regression. Though the Gaussian regression case may be considered less challenging, error correction in flexible models is vividly concerned with this situation.
The first method developed here is a generalization of the standard regression calibration (cf. e.g. Carroll et al. (1995)) to the specific form of the RVM. Furthermore, this thesis contains a previously unattempted approach for flexible regression, the exact structural quasi likelihood correction (cf. e.g. Carroll et al. (1995)). A simulation study is conducted, where all cor-

rection methods developed here are compared to the naive analysis, ignoring measurement error, and to a state-of-the-art approach by Berry, Carroll & Ruppert (2002). While the Bayesian P-spline approach by Berry et al. (2002) is based on parameter sampling via MCMC techniques, the methods introduced here exclusively rely on algorithmic schemes for parameter optimization. Hence, they require much less computing time while giving error correction that appears to be in most cases better than the MCMC method. An exception in terms of computer efficiency is the simulation based and thus also computer intensive SIMulation EXtrapolation approach (SIMEX), which is motivated from Carroll, Maca & Ruppert (1999) and transferred to the RVM methodology.

Chapter 4 is dedicated to correction methods for flexible binary regression. A part of the methods from the Gaussian case can with slight modification also be applied here. However, this work also develops a more refined approach for the binary case, which will be termed 'expanded basis function calibration' and has not yet been used in the context of flexible binary regression before. It follows the spirit of the structural quasi likelihood, but contains much more algebra to suit the binary case. The results of the attached simulation, however, can not justify the additional costs of this sophisticated method. Instead, the results witness the strength of the SIMEX approach and a here developed approach based on MCMC sampling techniques. This latter method combines the Bayesian treatment of covariate measurement error, which essentially goes back to Richardson & Gilks (1993b), and the MCMC version of the RVM. Another approach, which is rather an ad-hoc development, combines the idea of Bayesian 'data augmentation' with the calibration of basis functions. This attempt proves to be surprisingly successful in the simulation study.
Furthermore, a real data example is included in this chapter where the influence of animal and plant protein intake on mortality is investigated. The complexity lies here in the flexible modeling of two error-prone covariates and additionally accounting for confounder variables in the model; having more than one error-prone covariate in a regression model is a relevant practical

situation, which has very rarely been reflected in the existing literature.
Finally, this chapter contains error correction in flexible models for binary longitudinal data, which is a very important enhancement to the cross-sectional scenario. Though this problem case is highly relevant for practical applications, it has not yet been discussed in the literature. This work shows how the usage of Bayesian sampling techniques allows for the estimation of a flexible binary regression model accounting for subject specific effects, autocorrelated covariate and response observations and covariate measurement error.

Chapter 5 gives the relevant details of how some of the previous methods from the binary case can be adopted for the Poisson case by undergoing only minor (technical) modifications. The escorting simulation study indicates that all correction methods bring a clear improvement, even when the measurement error is small. As in the binary case, there is no unambiguous recommendation which method to favor in the Poisson case.

The broad spectrum of methods visited during this work brings up a number of important, yet unsolved, problems and generates various aspects that appear to be promising for further investigation.
Chapter 6 briefly recapitulates the achievements of this thesis and concludes with a discussion on several perspectives for future research based on these achievements.

Chapter 2

The main techniques

Mostly, the type of functional relationship between independent and dependent variables in a regression context is hardly to predict a priori. A reasonable strategy lies then in adopting such a type of regression model for the analysis that is as flexible as possible to fit the data and find this relationship. Successful application of this philosophy can be found in many examples of real data analysis. These include the investigation of the influence of construction year and floor space on the rent of a flat (cf. Lang & Brezger (2004)) or the effect of calendar time on the forest health (cf. Fahrmeir, Kneib & Lang (2004)). Those methods are particularly attractive in the field of epidemiology where causal relations are highly complex and not nearly investigated and covariates like e.g. the exposure to a certain radiation or pollution may have an unpredictable effect on the human organism. It is again this area of epidemiology which shows a strong demand for methods that take into account that observed covariates can be mismeasured or inaptly operationalized. Examples include Küchenhoff & Carroll (1997) who investigate the critical dust exposure to affect health or Augustin (2002) who studies the effect of animal and plant protein intake on mortality and morbidity, a data example that will be re-analyzed in chapter 4, but then, for the first time, adopting flexible modeling of the intake effects.
This chapter introduces the core concept of flexible regression using the rele-

vance vector machine (RVM), gives an overview of general Bayesian methodology that will be used in this thesis and recalls some fundamental concepts of analysis under covariate measurement error.
Throughout this thesis, only the RVM and its Markov Chain Monte Carlo (MCMC) transformation described here are used for flexible modeling. However, some of the derived results are generalizable to alternative flexible approaches as well. Introduction of covariate measurement error is postponed to the last section of this chapter, where some error models and standard correction methods are recalled. The newly developed strategies to handle measurement error in flexible regression are already sketched here to demonstrate their relation to established approaches, and to make the new achievements obvious.

2.1 Sparsity and smoothness - The relevance vector machine

This section describes the RVM by giving detailed descriptions of the model assumptions and inference methods (cf. Tipping (2000) and Tipping (2001)). This method comes from the field of machine learning, where the respective literature conventionally terms 'target' what a statistician terms 'response'. This thesis will, however, rely on the statistical terminology and notation by denoting 'Y' as the response variable. Neither the model setup nor the inferential methods take into account covariate measurement error at this stage.

2.1.1 The model setup

In regression tasks one is commonly concerned with data sampled from multiple covariates $X_d, d = 1, \ldots, D$ and a single response Y. A typical sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $(\mathbf{x}_i, y_i) \in \mathbb{R}^D \times A$, includes a vector of covariate observations $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ and a scalar response y_i for every unit i. Here, e.g.

$A := \mathbb{R}$ stands for the Gaussian regression, $A := \{0, 1\}$ for the binary regression and $A := \mathbb{N}_0$ for the Poisson regression case.

Based on that sample, one would like to estimate a model that allows for making good predictions for y^* on yet unseen $\mathbf{x}^*$. The responses are assumed being decomposable into a structural and a random part:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \ldots, N, \tag{2.1}$$

where the errors ϵ_i comprise e.g. imprecise measurement of the responses and the impact of unaccounted covariates; it is commonly assumed that $\mathbb{E}(\epsilon_i) = 0$ and thus $\mathbb{E}(Y|X = \mathbf{x}) = f(\mathbf{x})$. The mean function $f(\mathbf{x})$ usually comprises two functions $f(\mathbf{x}) = G(f^*(\mathbf{x}))$, where $G(z)$ denotes the so-called response function, which is conventionally chosen with respect to the distribution of the response and the specific predictor $f^*(\mathbf{x})$.
Making assumptions about the functional form of $f^*(\mathbf{x})$ in (2.1) is the first source of error on the way to a sensible analysis. Classical regression allots a polynomial of certain order, determined by a set of related parameters, to describe this dependency. However, if the true $f^*(\mathbf{x})$ is not representable by the chosen polynomial further analysis will be erroneous. In many practical cases, as in those mentioned in the preface of this chapter, one has no previous idea how to specify $f^*(\mathbf{x})$ and thus wisely prefers to follow a rather flexible approach, like e.g. the one applied throughout this work - the RVM; instead of proposing a rather rigid structural form, the RVM follows the concept of fitting a set of weighted (nonlinear) basis functions to the data; this core idea can also be found in other flexible approaches like smoothing splines (cf. e.g. Wahba (1978), Wahba (1990)), B-Splines (cf. de Boor (1978)), P-splines (cf. Eilers & Marx (1996)) and the support vector machine (SVM) (cf. Vapnik (1998)).
The utilized basis functions are generally not representable by linear combinations of covariates (nor arbitrary products of covariates), but are rather artificial. Each main effect and each interaction between covariates is nonlinearly represented by an individual set of basis functions. The structural part of the model $f^*(\mathbf{x})$ can be most flexibly represented by the double-sum

over D^* main effects plus interactions and their individual representation as a sum of weighted basis functions, respectively. An intercept ω_0 is typically introduced as well. Thus, the decomposition of the response for D^*= number of main effects and interactions reads as

$$y_i = G\left(\sum_{d=1}^{D^*}\sum_{j=1}^{J_d}\omega_{dj}\phi_{dj}(\mathbf{x}_{di}) + \omega_0\right) + \epsilon_i, \quad i = 1, \ldots, N, \tag{2.2}$$

where J_d is the number of knots used for specifying the dth design matrix expanding either a main effect in univariate basis functions or an interaction in multivariate basis hyperplanes, respectively. Depending on whether d indices a main effect or an interaction $\mathbf{x}_{di}$ is a single covariate value or a vector of values associated with the ith observation in the data, respectively.

A popular choice of basis function, which will be adopted throughout this thesis, is the radial basis function (RBF) kernel (here for a main effect and thus x being a scalar)

$$\phi_{dj}(x_{di}) = \exp(-\eta(x_{di} - c_{dj})^2). \tag{2.3}$$

Here, $\phi_{dj}(x_{di})$ denotes the jth (univariate) basis function at position x_{di} for the dth main effect being centered on knot c_{dj} using a covariate specific kernel parameter η. In the case of an interaction between covariates, the multivariate RBF kernel is given by

$$\phi_{dj}(\mathbf{x}_{di}) = \exp\left(-(\mathbf{x}_{di} - \mathbf{c}_{dj})^{\mathrm{T}}\boldsymbol{\eta}(\mathbf{x}_{di} - \mathbf{c}_{dj})\right), \tag{2.4}$$

where the jth basis hyperplane is being centered on a multidimensional knot $\mathbf{c}_{dj}$ at position $\mathbf{x}_{di}$ using a diagonal matrix $\boldsymbol{\eta} = diag(\eta_1, \eta_2, \ldots,)$ with covariate specific kernel parameters. For simplicity, the notation does not distinguish between the main effects and interactions and all of the basis functions are collected in one design matrix yielding

$$y_i = G\left(\Phi(\mathbf{x}_i)\boldsymbol{\omega}\right) + \epsilon_i, \quad i = 1, \ldots, N, \tag{2.5}$$

where $\Phi(\mathbf{x}_i)$ is the ith row, i.e. corresponding to observation $\mathbf{x}_i$, of the complete design matrix

$$\Phi = (\mathbf{1}, \Phi_1, \ldots, \Phi_d, \ldots, \Phi_{D^*})$$

including an intercept and D^* horizontally concatenated matrices $\Phi_d, d = 1, \ldots, D^*$. Again, each design matrix $\Phi_d = (\phi_{d1}, \ldots, \phi_{dJ_d})$ consists of J_d column vectors, each representing a single basis function evaluated at all of the sample observations $\mathbf{x}_{di}, i = 1, \ldots, N$, cf. (2.3) and (2.4). The parameter vector of the weight coefficients corresponding to that design matrix (2.5) is given by

$$\boldsymbol{\omega} = (\omega_0, \omega_{11}, \ldots, \omega_{1J_1}, \omega_{21}, \ldots, \omega_{2J_2}, \ldots, \omega_{D^*1}, \ldots, \omega_{D^*J_{D^*}})^{\mathrm{T}}$$

and contains one intercept weight plus $J = \sum_{d=1}^{D^*} J_d$ weights. The structure of the RVM is additive within each expansion of a non-parametric effect and also between the non-parametric effects, a property that is typically found in generalized additive models (cf. e.g. Fahrmeir & Tutz (2001)). So far and until the end of this subsection no specific assumption is made about the distribution of the response and the error. Estimation of the weight parameters leads to the prediction function $\widehat{f}(\mathbf{x}) = G\left(\Phi(\mathbf{x})\widehat{\boldsymbol{\omega}}\right)$.

A special feature of the RVM as described by Tipping (2000) lies in using every covariate sample as a basis knot on which a basis function is centered, i.e. $J_d = N, \quad d = 1, \ldots, D^*$, which is also found in the smoothing splines. So, principally every observation is conceded to express its impact on fitting the data. However, in later applications contained in this thesis only a subset recruiting the 100 quantiles of the covariate observations will be used. This is in concordance with B-splines and P-splines, where usually some points on a grid are selected to represent the knots of the basis functions. The complete set of basis functions can be viewed as an arsenal of functions, where usually a small subset of vectors sufficiently explains the variation of the response. So, in order to avoid overfitting of the data, the method should automatically infer which subset of functions is relevant for data fitting and exclude the others from the model.

The RVM, hereby, follows an approach of MacKay (1994), termed automatic relevance determination, to find these relevant basis functions. Exclusion of a basis from the model is equivalent to setting the related weight parameter to zero. A preference for a sparse model with only few weights being nonzero is here encoded by placing a Gaussian prior over every weight, centered on

zero with an individual precision (inverse variance) parameter

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{j=0}^{J} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{\alpha_j}{2}\omega_j^2\right). \tag{2.6}$$

This is in exact analogy to the concept of penalization (cf. e.g. Fahrmeir & Tutz (2001)), where, usually only one, but here multiple penalization constants α_j, each associated with an input basis, are introduced. Generally, those penalization constants that correspond to basis vectors with low predictive value are expected to take on large values during the inference process, i.e. the data driven penalization is expected to be high in this case. A basis function is excluded from the model if the related data driven penalization constant tends towards infinity.
MacKay (1994) suggests estimation of the $\alpha_j, j = 0, \ldots, J$ via a likelihood approach, while Tipping (2000) goes for a Bayesian specification by placing a prior distribution over the α_j's, too.
Gamma hyperpriors are specified over these inverse variance parameters, also called scale parameters

$$p(\boldsymbol{\alpha}) \;=\; \prod_{j=0}^{J} \Gamma(a)^{-1} b^a \alpha_j^{a-1} \exp(-b\alpha_j),$$

where $\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t)\mathrm{d}t$. In case the variance $\mathbb{V}(Y|X)$ includes a dispersion parameter σ^2, like in the Gaussian response case, a Gamma hyperprior is also placed over the, usually inverse, dispersion parameter, where it is conventionally defined $\beta \equiv \sigma^{-2}$

$$p(\beta) \;=\; \Gamma(c)^{-1} d^c \beta^{c-1} \exp(-d\beta).$$

Tipping (2001) describes setting the corresponding parameters $a = b = c = d = 0$, a way to proceed that is equivalent to specifying uniform distributions for $\boldsymbol{\alpha}$ and β on a logarithmic scale.

Before the details on the inferential process are presented in the next section, Figure 2.1 visualizes the fitting principle of the relevance vector machine

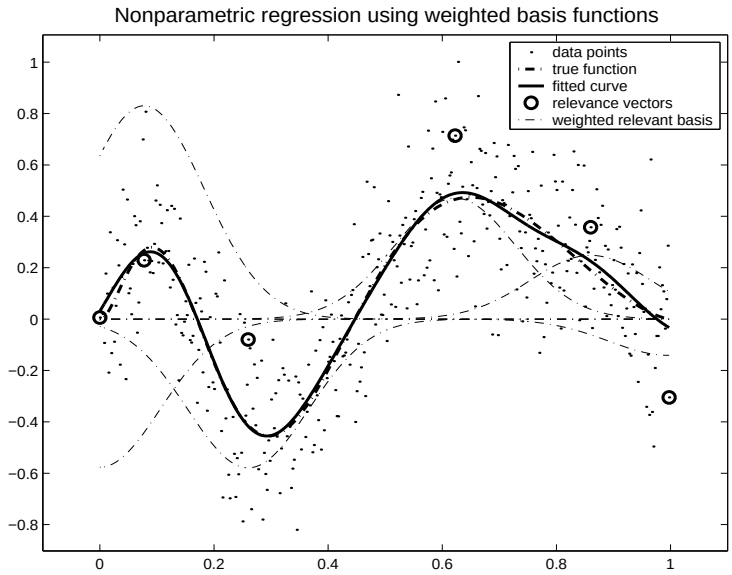

Figure 2.1: Example of RVM regression with true function $f(x) = \sqrt{x(1-x)}\sin\frac{2\pi(1+2^{-3/5})}{x+2^{-3/5}}$. *The six weighted basis functions with nonzero weights are displayed in the background. The associated relevant vectors are circled. Adding up these basis functions yields the prediction function.*

(here for Gaussian regression). While allowing for a very complex model to describe the data (2.2) by specifying a huge arsenal of potentially relevant basis vectors, the RVM model automatically determines truly relevant vectors for sufficiently modeling the data. The final model is then typically very parsimonious.

2.1.2 Inference

The estimation of the unknown parameters $\boldsymbol{\omega}, \boldsymbol{\alpha}$ and β in a Bayesian framework is commonly done via the posterior distribution of these parameters, given the data $\mathbf{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$. Bayes' rule gives

$$p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta)p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta)}{p(\mathbf{y})}. \tag{2.7}$$

Then, predicting y^* for a previously unseen observation $\mathbf{x}^*$, is achieved via the predictive distribution

$$p(y^*|\mathbf{y}) = \int p(y^*|\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta) p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta|\mathbf{y}) \, \mathrm{d}\boldsymbol{\omega} \mathrm{d}\boldsymbol{\alpha} \mathrm{d}\beta.$$

It may come as no surprise that these calculations can not be performed in full analytically and an effective approximation must be sought. Analytic calculation of (2.7) is not feasible, since the normalizing integral $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta) p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta) \, \mathrm{d}\boldsymbol{\omega} \mathrm{d}\boldsymbol{\alpha} \mathrm{d}\beta$ can not be computed. A frequentist approach would circumvent this problem by merely maximizing the numerator of (2.7) using some kind of scoring algorithm, which however comes at the cost of not adequately capturing the uncertainty in these parameter estimates.
Now, Bayesian inference must rely on an effective approximation of the joint posterior distribution which is based on the following decomposition

$$p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \beta|\mathbf{y}) = p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}, \beta) p(\boldsymbol{\alpha}, \beta|\mathbf{y}),$$

where Bayes' rule yields the posterior over the weights and scales

$$p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}, \beta) = \frac{p(\mathbf{y}|\boldsymbol{\omega}, \beta) p(\boldsymbol{\omega}|\boldsymbol{\alpha})}{p(\mathbf{y}|\boldsymbol{\alpha}, \beta)} \tag{2.8}$$

$$p(\boldsymbol{\alpha}, \beta|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\alpha}, \beta) p(\boldsymbol{\alpha}) p(\beta). \tag{2.9}$$

The posterior over the scales (2.9) can only be determined up to a constant since repeated application of the Bayes rule again involves the normalizing constant $p(\mathbf{y})$. Expression (2.9) only allows for finding the mode of the posterior instead of revealing its full form. Tipping (2001) comments on the quality of this approximation, which seems to be very effective in general. Both distributions (2.8) and (2.9) build the core of inference for the RVM. Depending on the specification of the distribution over the responses the trail of inference is pursued. The following inferential schemes distinguish between the Gaussian and non-Gaussian case, including Bernoulli and Poisson distributed responses.

Gaussian regression case

In the Gaussian case, the response function $G(z)$ in model (2.1) is the identity function, and the errors are assumed to be i.i.d. normally distributed, where it is conventionally defined $\beta \equiv \sigma^{-2}$

$$p(\boldsymbol{\epsilon}) = \prod_{i=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}\epsilon_i^2\right).$$

Thus the likelihood is Gaussian

$$p(\mathbf{y}|\boldsymbol{\omega}, \beta) \;=\; (2\pi\beta^{-1})^{-N/2} \exp\left\{-\frac{\beta}{2}(\mathbf{y} - \Phi\boldsymbol{\omega})^{\mathrm{T}}(\mathbf{y} - \Phi\boldsymbol{\omega})\right\}$$

and since the prior over the weights (2.6) is chosen to be Gaussian, too, the marginal likelihood of the data is again - Gaussian:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\alpha}, \beta) &= \int p(\mathbf{y}|\boldsymbol{\omega}, \beta) p(\boldsymbol{\omega}|\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\omega}, \\ &= (2\pi)^{-\frac{N}{2}} \mid C \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}^{\mathrm{T}} C^{-1} \mathbf{y}\right), \end{aligned} \tag{2.10}$$

where the covariance matrix $C = \beta^{-1}\mathbf{I} + \Phi A^{-1}\Phi^{\mathrm{T}}$ and the diagonal matrix $A = diag(\alpha_0, \alpha_1, \ldots, \alpha_N)$ contains the scale parameters from (2.6), previously also characterized as regularization constants. With the Gaussian prior over the weights (2.6), the posterior of the weights is Gaussian

$$p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}, \beta) \;=\; (2\pi)^{-\frac{(J+1)}{2}} \mid \Sigma \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu})\right), \tag{2.11}$$

where the posterior covariance matrix and mean vector are, respectively

$$\Sigma \;=\; (\beta\Phi^{\mathrm{T}}\Phi + A)^{-1}, \tag{2.12}$$

$$\boldsymbol{\mu} \;=\; \beta\Sigma\Phi^{\mathrm{T}}\mathbf{y}. \tag{2.13}$$

Since the posterior distribution is multivariate Gaussian the posterior mode estimator is then given by the mean. The following two inserted paragraphs

comment on this posterior mean estimator from a rather frequentistic perspective.

Connection between RVM inference and penalized likelihood estimation:

A look at the log-posterior of the weights reveals the connection between this Bayesian setup of the relevance vector machine and the principle of penalization, which is applied in flexible regression methods under a frequentistic perspective, like e.g. in the natural cubic splines, see Fahrmeir & Tutz (2001). The log-posterior is according to Bayes' theorem proportional to the log-likelihood plus the logarithm of the Gaussian prior:

$$\log p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) \propto \log p(\mathbf{y}|\boldsymbol{\omega}, \boldsymbol{\alpha}, \sigma^2) - \sum_{j=0}^{J} \alpha_j \omega_j^2. \tag{2.14}$$

In the frequentistic branch of statistics this expression is well known as the 'penalized likelihood' under quadratic penalty function and fixed penalties $\alpha_j, j = 0, \ldots, J$. Estimation of the parameters $\boldsymbol{\omega}$ is done via maximization of this penalized likelihood. This means finding the root of the so-called penalized score function, which is the first derivative of the penalized likelihood (2.14) with respect to $\boldsymbol{\omega}$ (cf. e.g. Fahrmeir & Tutz (2001)). Then, the penalized least squares estimator under quadratic penalty function and fixed penalties is exactly equivalent to the posterior mode estimator in (2.13), i.e. the two rather different concepts of (Bayesian) posterior mode estimation and (frequentistic) penalized likelihood generate the same estimator. In a truly frequentistic approach the variance of this estimator is derived via the so-called sandwich formula (cf. e.g. Fahrmeir & Tutz (2001), p.57). However, in the Bayesian setup, the posterior variance estimator (2.12) is given by the inverse expected (penalized) Fisher matrix, which is computed via the second derivative of the penalized likelihood (2.14) with respect to $\boldsymbol{\omega}$ (cf. e.g. Lin & Zhang (1999), Fahrmeir & Tutz (2001), p.62).
The frequentistic viewpoint to the estimation of the parameters $\boldsymbol{\omega}$, will be used later in the development of correction methods for measurement error.

Unbiasedness of the RVM estimator:

From a frequentist's point of view a strongly desired property of an estimator is unbiasedness. It is clear that in the context of penalization a revised definition of unbiasedness is needed, which has been stated by Fan & Li (2001): an estimator is 'nearly unbiased' if it avoids unnecessary modeling bias in case the true unknown parameter is large. A sufficient condition for unbiasedness in this sense is the first derivative of the penalty function with respect to the weights being zero for large weights. Loosely spoken, constant penalization of large parameter values yields nearly unbiased estimates. So in the spirit of Fan & Li (2001) the resulting estimator of the RVM is not unbiased because in order to achieve unbiased estimates one would have to place an improper prior over the weights, possibly leading to a non-standard posterior distribution of the weights. A special case is the usage of the following prior distribution:

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \begin{cases} \prod_{j=0}^{J} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left\{-\frac{\alpha_j}{2}\omega_j^2\right\} & : \quad \text{if } |\omega_j| < \lambda \\ c(\alpha_j, \lambda) & : \quad |\omega_j| \geq \lambda \end{cases}$$

where $c(\alpha_j, \lambda)$ is a constant making $p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = 1$. This bifid specification of the prior leads to a Gaussian posterior distribution of the weights as above in (2.11) with the following specification of moments:

$$\begin{aligned} \Sigma^* &= (\beta \Phi^{\mathrm{T}} \Phi + A^*)^{-1} \\ \boldsymbol{\mu} &= \Sigma^* \Phi^{\mathrm{T}} \mathbf{y}, \end{aligned}$$

where $A^* = diag(\alpha_0^*, \ldots, \alpha_J^*)$ and

$$\alpha_j^* = \begin{cases} \alpha_j & : \quad \text{if } |\omega_j| < \lambda \\ 0 & : \quad \text{if } |\omega_j| \geq \lambda \end{cases}$$

One would now have to specify a value for λ and to know the value of the true ω_j to configure the posterior mean and variance estimator. Several ad-hoc strategies are conceivable to run this method but this trail will not be continued in this thesis.

Finally, an estimation scheme for the hyperparameters $\boldsymbol{\alpha}$ and β is described. In contrast to the posterior over the weights in (2.11) the posterior over the scale parameters in (cf. 2.8) is only defined up to a constant. A type II maximum likelihood approach (Good (1965)) is taken to optimize the scales, where the idea is to mimic the maximum likelihood approach at the marginal level. For other cases than uniform hyperpriors (in some scale) this is a variation on the type II maximum likelihood making allowance for the specific distributions of the hyperparameters. Harville (1974) showed that this optimization procedure is equivalent to the concept of 'restricted maximum likelihood' (REML) as introduced by Patterson & Thompson (1971). Throughout this thesis uniform hyperpriors over a logarithmic scale are used and thus the objective function to be optimized is solely the logarithm of the marginal likelihood (2.10)

$$\mathcal{L} \;=\; -\frac{1}{2}\left(\log \mid C \mid + \mathbf{y}^{\mathrm{T}} C^{-1} \mathbf{y}\right), \tag{2.15}$$

with the covariance matrix

$$C = \beta^{-1}\mathbf{I} + \Phi A^{-1} \Phi^{\mathrm{T}}.$$

Solutions for $\boldsymbol{\alpha}$ and β are described in turn whereas there are two fundamentally different approaches in deriving the optimal hyperparameters $\alpha_j, j = 0, \ldots, J$.

<u>$\boldsymbol{\alpha}$-Rule 1&2:</u>

In an 'early' version of the RVM (cf. Tipping (2001)), the updating rule for the entries in $\boldsymbol{\alpha}$ is achieved by differentiating the marginal log-likelihood (2.15) with respect to $\log \alpha_j$, which gives

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \log \alpha_j} = \frac{1}{2}\left(1 - \alpha_j(\mu_j^2 + \Sigma_{jj})\right), \tag{2.16}$$

where Σ_{jj} denotes the jth diagonal element of the covariance matrix Σ (2.12) and μ_j the jth element in the mean vector (2.13). Setting this to zero and solving for α_j gives

$$\alpha_j^{new} = \frac{1}{\mu_j^2 + \Sigma_{jj}}. \tag{2.17}$$

Alternatively, setting (2.16) to zero and defining quantities

$$\gamma_j = 1 - \alpha_j \Sigma_{jj} \tag{2.18}$$

leads to the following update

$$\alpha_j^{new} = \frac{\gamma_j}{\mu_j^2}. \tag{2.19}$$

Although this update does not benefit from the local maximization of $\mathcal{L}$ it was observed to lead to much faster convergence, and in practical applications a hybrid scheme of both updating rules is used. This scheme starts from the full model including all of the potential basis functions making this approach relatively computationally intensive at the beginning. An α_j-scale is manually set to infinity when it either exceeds a pre-specified threshold or becomes less or equal zero leading to pruning of the corresponding basis. Once a basis is excluded from the model it can not be reintroduced, so this is a strictly degenerative process.

$\boldsymbol{\alpha}$-Rule 3:

Tipping & Faul (2002) more recently investigated a decomposition of the marginal likelihood (2.15) into two terms where only one is dependent on the jth basis vector. This is the key to a non-iterative analytic optimization of the marginal likelihood with respect to a single hyperparameter α_j. The two stationary points of the marginal likelihood are found to be

$$\alpha_j = \begin{cases} \frac{s_j^2}{q_j^2 - s_j} & \text{if } q_j^2 - s_j > 0 \\ \infty & \text{else} \end{cases}. \tag{2.20}$$

For simplicity the quantities $q_j = \phi_j^{\mathrm{T}} C_{-j}^{-1} \mathbf{y}$ and $s_j = \phi_j^{\mathrm{T}} C_{-j}^{-1} \phi_j$ have been defined here. In (2.20), $C_{-j} = C - \alpha_j \phi_j \phi_j^{\mathrm{T}}$ denotes the covariance matrix of the marginal log-likelihood (2.15) with the influence of basis vector ϕ_j removed. The index *new* is suppressed in (2.20) to underline the non-iterative character of this solution. This maximum of the objective function for an individual parameter is, of course, dependent on the values of all other hyperparameters.

Separating the influence of the jth basis in the marginal likelihood is also a key point in order to derive the following stepwise optimization algorithm: Firstly, the likelihood gain of the following actions for every basis function is assessed:

- **add** the basis to the model (if not already present in the model)
- **remove** the basis (if basis is in the actual model)
- **update** the hyperparameter of this basis (if basis is present in the actual model).

Then only that action associated with the highest likelihood gain is realized. This scheme usually starts with only a single basis, and most desirable, from a computational point of view, even the maximum number of basis being in the model during this scheme is usually only a fraction of all potential basis functions. Alternative versions of this scheme check the possible actions only for a random subset of basis functions, a measure that further increases computational speed.
The required quantities q_j and s_j in (2.20) can be conveniently computed by utilizing the so-called Woodbury identity involving the posterior moments (2.13) and (2.12), cf. Tipping & Faul (2003).

β-Rule:

The updating rule for the error variance is derived from differentiating the objective function (2.15) with respect to $\log \beta$

$$\frac{\partial \mathcal{L}(\beta)}{\partial \log \beta} = \frac{1}{2} \left(N - \beta \parallel \mathbf{y} - \Phi \boldsymbol{\mu} \parallel^2 - \mathrm{tr}(\Sigma \Phi^{\mathrm{T}} \beta \Phi) \right), \tag{2.21}$$

where $\parallel \mathbf{y} - \Phi \boldsymbol{\mu} \parallel^2 := (\mathbf{y} - \Phi \boldsymbol{\mu})^{\mathrm{T}} (\mathbf{y} - \Phi \boldsymbol{\mu})$. Here, $\mathrm{tr}(\Sigma \Phi^{\mathrm{T}} \beta \Phi)$ is the trace of the matrix product that can be re-written as the sum over all parameters γ_j defined earlier in (2.18). Equating this to zero gives

$$\sigma^{2\,new} = \frac{\parallel \mathbf{y} - \Phi \boldsymbol{\mu} \parallel^2}{N - \sum_j \gamma_j}. \tag{2.22}$$

It is important to stress that neither of the presented updating rules leads to an analytic solution for all hyperparameters at once, and thus the optimization procedure is practically iterative, alternating between finding the moments of the posterior distribution (2.11) and optimizing the marginal likelihood (2.10) with respect to the hyperparameters until some convergence criteria is satisfied.

Non-Gaussian regression case

Under the term non-Gaussian regression the binary classification case and Poisson regression is subsumed in this thesis. The main difference to the Gaussian case is that the response function $G(z)$ in model (2.1) is no longer the identity function but e.g. the Gaussian cumulative distribution function for binary, or the exponential function for count data. Therefore, the likelihood is non-Gaussian and thus the integration involved in determining the marginal likelihood is infeasible and so is the posterior over the weights. Note that in the binary and Poisson case the variance of ϵ is completely specified by the mean $\mathbb{E}(Y|X)$, as long as under-/overdispersion is not considered.
The general approach to infer the parameters $\boldsymbol{\omega}$ in this case is via the approximative decomposition

$$p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}) \propto p(\mathbf{y}|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{\alpha}), \tag{2.23}$$

where the specification of the likelihood depends on the distribution over the responses and the prior over the weights is Gaussian (2.6). Tipping (2001) originally follows the Laplace's method (cf. e.g. Tierney & Kadane (1986), MacKay (2003)) approximating $p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha})$ by a Gaussian centered on its mode $\boldsymbol{\omega}_{MP}$ with its covariance matrix being equal to the inverse observed Fisher matrix (i.e. negative inverse Hessian) at the mode. However, in this thesis the inverse expected Fisher matrix will be used instead (cf. e.g. Lin & Zhang (1999), Fahrmeir & Tutz (2001)). The moments of the Gaussian

approximation to the posterior are then given by

$$\Sigma = (\Phi^{\mathrm{T}} B \Phi + A)^{-1}, \tag{2.24}$$

$$\boldsymbol{\mu} = \boldsymbol{\omega}_{MP} = \Sigma \Phi^{\mathrm{T}} B \mathbf{y}^* \tag{2.25}$$

with working observations

$$\mathbf{y}^* = \Phi\boldsymbol{\omega} + D^{-1}(\mathbf{y} - G(\Phi\boldsymbol{\omega})). \tag{2.26}$$

The diagonal matrix B consist of elements

$$B_{ii} = \left(\frac{\partial G(\Phi(\mathbf{x}_i)\boldsymbol{\omega})}{\partial(\Phi(\mathbf{x}_i)\boldsymbol{\omega})} \right)^2 / \mathbb{V}(y_i|x_i), \quad i = 1, \ldots, N$$

involving the first derivative of the response function with respect to the linear predictor and $\mathbb{V}(y_i|x_i)$, which is the variance of the responses according to their distribution. The diagonal matrix D consists of elements

$$D_{ii} = \left(\frac{\partial G(\Phi(\mathbf{x}_i)\boldsymbol{\omega})}{\partial(\Phi(\mathbf{x}_i)\boldsymbol{\omega})} \right), \quad i = 1, \ldots, N.$$

In contrast to the Gaussian response case, finding the posterior mean of (2.23) is iterative and uses some scoring method like e.g. Fisher scoring.
The resulting posterior mean estimator (2.25), exactly corresponds to the 'iteratively weighted least squares' (IWLS) representation of the Fisher scoring maximization (cf. e.g. Fahrmeir & Tutz (2001)). The maximization of (2.23) can be again considered from the frequentistic perspective as optimization of a likelihood, coming from a generalized linear model, under quadratic penalization. In contrast to a truly frequentistic approach, where the variance of the mean estimator is derived via the so-called sandwich formula Fahrmeir & Tutz (2001), p.57), (2.24) is here approximated by the inverse expected Fisher matrix.
The relation between estimating the moments, particularly the mean, of the posterior approximation and inference in generalized linear models under penalization is a crucial point and provides solutions for a broad spectrum of response distributions.

Like in the Gaussian case the posterior approximation and the posterior mode estimation of the hyperparameters alternate. However, the marginal likelihood is no longer Gaussian here. Tipping (2001) suggests a pragmatic approach and assumes that $p(\mathbf{y}^*|\boldsymbol{\alpha}, \boldsymbol{\omega})$ is approximately Gaussian, which is basically again a Laplace approximation, giving the following objective function

$$\begin{aligned} \mathcal{L} \quad &= \quad -\frac{1}{2}\left(\log \mid C \mid + {\mathbf{y}^*}^{\mathrm{T}} C^{-1} \mathbf{y}^*\right) \\ &\text{where} \quad C = B + \Phi A^{-1} \Phi^{\mathrm{T}} \end{aligned} \tag{2.27}$$

with working observations $\mathbf{y}^*$ as defined above (2.26). This optimization procedure based on the marginal log-likelihood is also used in the concept of 'restricted maximum likelihood' (REML) (cf. Harville (1974), Patterson & Thompson (1971)). Finding the hyperparameters proceeds by differentiating (2.27), equating to zero and solving for the respective hyperparameter.
For the α_j's this yields the same updating rules as in the Gaussian case, stated in (2.17) and (2.19), but with the elements of $\boldsymbol{\mu}$ and Σ replaced by their estimators (2.25) and (2.24), respectively.
The one step maximization scheme (2.20) that will be used in the present work utilizes the covariance matrix from (2.27) and the vector of working observations $\mathbf{y}^*$ (2.26).

Since the mean function fully implies the variance function in the binary and the Poisson response case, there is no additional updating scheme for the variance. However, potential under-/overdispersion can disturb the setting of the exponential family since the data might suggest a different variance function $\mathbb{V}^*(y_i|x_i) = \sigma^2\, \mathbb{V}(y_i|x_i)$ than implied by the mean function where σ^2 denotes the dispersion parameter accounting for more or less variance in the responses than implied by the mean model. Dispersion is easily introduced into the RVM by including σ^2 in matrix B which consequently modifies (2.25) and (2.24). An estimator for the dispersion parameter can be derived by differentiation of the objective function with respect to σ^2

$$\frac{\partial \mathcal{L}(\sigma^2)}{\partial \sigma^2} = \frac{1}{2}\left[-\frac{N}{\sigma^2} + \frac{(\mathbf{y}^* - \Phi\boldsymbol{\mu})^{\mathrm{T}} B (\mathbf{y}^* - \Phi\boldsymbol{\mu})}{(\sigma^2)^2} - \mathrm{tr}(\Sigma \Phi^{\mathrm{T}} B \Phi)\right], \tag{2.28}$$

and equating to zero which gives

$$\left(\sigma^2\right)^{new} = \frac{\left(\mathbf{y}^* - \Phi\boldsymbol{\mu}\right)^{\mathrm{T}} B \left(\mathbf{y}^* - \Phi\boldsymbol{\mu}\right)}{N - \sum_j \gamma_j} \left(\sigma^2\right)^{old}.$$

As before $\gamma_j = 1 - \alpha_j \Sigma_{jj}$. Note that the old value $(\sigma^2)^{old}$ is also contained in B and thus in $\boldsymbol{\mu}$ and γ_i. The potential need of a dispersion parameter for non-Gaussian regression cases is however not considered in the practical applications presented in this thesis.

2.2 Flexible regression and Markov Chain Monte Carlo

The optimization scheme in the previous section relied on several approximations. For instance, the estimation of the scales was based on the marginal likelihood instead of the posterior distribution of the scales which was not analytically available. Furthermore, to find the weight estimates in the non-Gaussian case a Laplace approximation to the weights posterior was applied, cf. (2.24) and (2.25). Both approximations are due to the complexity of the joint posterior including intractable integrals. The approximations have been utilized to iteratively maximize the joint posterior, an approach, which is, however, not at the heart of true Bayesian analysis. Following the approximative approach, e.g. uncertainty in the smoothing parameters α_i is not captured in the prediction.
In this section a full Bayesian version of the flexible regression is introduced using MCMC sampling techniques while at the same time preserving the two main features of the RVM:

- the structural part of the model contains the expansion of the covariates in terms of weighted radial basis functions
- relevant basis functions are automatically selected.

Therefore, the main topics of Bayesian inference using Markov Chain Monte Carlo (MCMC) methods are first briefly presented. A good basis to understand MCMC methods is given by Jackman (2000) giving insight by presenting practical applications. Green (2001) Green (2001) theoretically and practically reviews a broad range of available MCMC tools. Denison et al. (2002) and Dellaportas & Roberts (2003) discuss the material in view of flexible regression and spatial statistics, respectively.

The second subsection describes in more detail flexible regression applying MCMC methodology in the binary response case, which is extensively used in chapter 4 of this thesis. A detailed description of MCMC methodology in the Gaussian response case is not undertake here, although a Bayesian P-spline approach (cf. Berry et al. (2002)) will also be used as a reference method in the attached simulation study. Instead the interested reader is conferred to Denison et al. (2002) who present a very detailed explanation of Bayesian flexible regression in the Gaussian case and, of course, to Berry et al. (2002).

2.2.1 Main topics of Bayesian inference using MCMC

Bayesian inference is based on the joint posterior distribution of the unknown parameters gathered in the vector $\boldsymbol{\theta}$. Following Bayes' theorem, the joint posterior distribution is derived from the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ and the prior distribution of the unknowns $p(\boldsymbol{\theta})$ as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}. \tag{2.29}$$

Expanding Bayesian models to more complex problems, following the hierarchical Bayesian approach, a hierarchy of p models each utilizing a parameter vector $\boldsymbol{\theta}_i, i = 1, \ldots, p$ may be needed. $\boldsymbol{\theta}_i, i = 1, \ldots, p$ denote here parameters that are introduced additionally to the vector $\boldsymbol{\theta}$. A neat explanation of hierarchical modeling, compared to empirical Bayes, is given by Vidakovic (2005), and the following lines are inspired by these lecture notes.

In a hierarchical setting, the prior is represented via a conditional hierarchy of so-called hyperpriors

$$p(\boldsymbol{\theta}) = \int p_1(\boldsymbol{\theta}|\boldsymbol{\theta}_1)p_2(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)p_3(\boldsymbol{\theta}_2|\boldsymbol{\theta}_3), \ldots, p_p(\boldsymbol{\theta}_{p-1}|\boldsymbol{\theta}_p)p_{p+1}(\boldsymbol{\theta}_p)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2 \ldots \mathrm{d}\boldsymbol{\theta}_p,$$

where only hyperparameter vectors of proximate hierarchy levels are dependent. Notice that in the hierarchy of data, parameters and hyperparameters,

$$\mathbf{y} \rightarrow \boxed{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2 \rightarrow \boldsymbol{\theta}_3 \ldots \rightarrow \boldsymbol{\theta}_p$$

$\mathbf{y}$ is independent of $\boldsymbol{\theta}_i$ given $\boldsymbol{\theta}$. That means

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\theta}_i) \stackrel{d}{=} p(\mathbf{y}|\boldsymbol{\theta}) \text{ and } p(\boldsymbol{\theta}_i|\boldsymbol{\theta}, \mathbf{y}) \stackrel{d}{=} p(\boldsymbol{\theta}_i|\boldsymbol{\theta}),$$

where '$\stackrel{d}{=}$' means equality in distribution. The joint distribution can then be represented as

$$p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) = p(\mathbf{y}|\boldsymbol{\theta})p_1(\boldsymbol{\theta}|\boldsymbol{\theta}_1)p_2(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \ldots p_p(\boldsymbol{\theta}_{p-1}|\boldsymbol{\theta}_p)p_{p+1}(\boldsymbol{\theta}_p).$$

Therefore, in order to fully specify the model, only neighboring conditionals and the closure distribution $p_{p+1}(\boldsymbol{\theta}_p)$ are needed. Reasons for the prior decomposition include feasibility of the analysis and objectiveness in that sense that the data should determine the hyperparameters.

Although the posterior distribution of the unknown parameters $p(\boldsymbol{\theta}|\mathbf{y})$ or $p(\boldsymbol{\theta}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p|\mathbf{y})$ in the hierarchical case might be written in closed form (at least up to a constant), the moments of that distribution are usually not computable analytically. Direct sampling from the joint posterior may be difficult due to its high dimensionality, so that simple Monte Carlo evaluation of the moments is not possible. MCMC instead simulates from a Markov Chain whose invariant distribution is $p(\boldsymbol{\theta}|\mathbf{y})$. There are essentially two basic sampling schemes used in MCMC: the very general Metropolis Hastings algorithm and its special case, Gibbs sampling.

Sampling schemes

The Metropolis-Hastings algorithm is an extension by Hastings (1970) of the original work by Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953). The Markov Chain is constructed by random generation of J samples $\boldsymbol{\theta}^{[0]}, \boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[J]}$, where $\boldsymbol{\theta}^{[0]}$ is the starting state of all unknown parameters involved in the model. Let $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}, \mathbf{y})$ denote the conditional density of $\boldsymbol{\theta}_i$ given all other parameters $\boldsymbol{\theta}_{\setminus i} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_p)$. In contrast to the previous paragraph, the $\boldsymbol{\theta}_i, i = 1, \ldots, p$ denote components of the full vector of unknowns $\boldsymbol{\theta}$ and not additionally introduced vectors.
Instead of sampling all unknown parameters stacked in $\boldsymbol{\theta}^{[j]}$ at once, one usually successively generates subvectors $\boldsymbol{\theta}_i^{[j]}$ having smaller dimensionality by performing the following two steps:

- Generate a candidate $\boldsymbol{\theta}_i'$ from an arbitrary proposal density $q_i(\boldsymbol{\theta}_i'|\boldsymbol{\theta})$, where the prime symbol denotes the proposal and $\boldsymbol{\theta}$ denotes the current state of all unknown parameters
- Set

$$\boldsymbol{\theta}_i^{[j]} = \begin{cases} \boldsymbol{\theta}_i' \text{ with probability } \alpha = \min\left[1, \frac{p(\boldsymbol{\theta}_i'|\boldsymbol{\theta}_{\setminus i}, y) q_i(\boldsymbol{\theta}_i|\boldsymbol{\theta}_i', \boldsymbol{\theta}_{\setminus i})}{p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}, y) q_i(\boldsymbol{\theta}_i'|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\setminus i})}\right] \\ \boldsymbol{\theta}_i^{[j-1]} \text{ otherwise} \end{cases} .$$

The key feature of this algorithm is that the possibly intractable normalizing constant of the full conditional density $p(\cdot|\boldsymbol{\theta}_{\setminus i}, \mathbf{y})$, evaluated at $\boldsymbol{\theta}_i'$ in the numerator and at $\boldsymbol{\theta}_i$ in the denominator, cancels out in the so-called acceptance probability α. The order of the subvector updating is arbitrary. The proposal q_i can be chosen arbitrarily, but must be capable of allowing to reach all areas of positive probability under $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}, \mathbf{y})$ (irreducibility postulate). However, specifying a poor proposal density might hurt the efficiency of this procedure, since either the constructed Markov Chain explores the space too slow or the rejection rate is too high. Finding a decent proposal is an outstanding problem and there is no universal answer. Three standard recipes applied in this work are described in the following.

Considering the special structure of the acceptance probability α in the sampling scheme, a favorable proposal should closely approximate the conditional density $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}, \mathbf{y})$, whereupon the acceptance probability is close to one. The special and very popular case of setting $q_i = p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}, \mathbf{y})$ presumes that the conditional density is recognizable as a standard distribution one could sample from. Every proposal drawn from that special distribution is accepted since the acceptance probability becomes exactly one. This particular choice was first proposed by Geman & Geman (1984) and is now known as the Gibbs sampler.
If all conditional densities are of standard form, the MCMC algorithm using Gibbs sampling is very similar to the 'iterated conditional modes' approach by Besag (1986). Here, the posterior mode is found by sequentially updating the parameters with the modal value of the full conditional, instead of sampling the parameters from the respective full conditional. However, the greatest benefits of the Bayesian framework are assumed to result from a truly Bayesian inference strategy implying sampling techniques in order to find the (empirical version of the) joint posterior.

If the conditional density is not recognizable as a standard distribution, numerical approximation can be used to construct the proposal. Therefore, e.g. the Laplace method can be used to approximate the log conditional density by a Gaussian distribution centered on its mode and variance equal to the negative inverse Hessian matrix evaluated at the mode. This method has already been used in the non-Gaussian RVM regression case in Subsection 2.1.2 to approximate the posterior distribution of the weights.

Another strategy of generating candidates $\boldsymbol{\theta}_i'$, when the conditional density is not recognizable, is drawing from a (multivariate) Gaussian centered on the current value $\boldsymbol{\theta}_i$ and variance equal to some multiple of the negative inverse Hessian matrix at the posterior mode. This is called a 'symmetric random walk Metropolis' algorithm. Since the proposal is symmetric it cancels out in the acceptance probability of the MH-algorithm. The random walk is biased towards the mode of the conditional density, since all 'uphill moves' increas-

ing $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}, \mathbf{y})$ are accepted, while some 'downhill moves' are rejected. A high acceptance probability is not always desirable for this type of proposal since it indicates a propensity to avoid the tails of the distribution.

Convergence

It is important to run the sampling scheme long enough to ensure the chain has approached stationarity - so that the samples $\boldsymbol{\theta}^{[j]}$ come from the invariant distribution of the Markov chain, which is the joint posterior $p(\boldsymbol{\theta}|\mathbf{y})$ by construction. Common practise is to discard the first number of draws as a burn-in period and base inference on the so-called 'Monte Carlo' sample. This recruits subsequent draws for which convergence is assumed to be achieved.
Popular alternatives use subsamples of the Monte Carlo sample using e.g. every tenth observation to obtain approximately independently identically distributed draws.
Assessing the convergence of the Markov chain and the optimal number of samples is a well known and discussed problem. Andrieu, de Freitas, Doucet & Jordan (2003) present a summary of some of the available approaches to this problem. Common tools to check for convergence include visual inspection of sampling paths, plots of autocorrelation and comparing a set of chains based on different starting values. The log of the posterior density at the current state may indicate, if the method is still working its way to a more representative part of the distribution. An instructive discussion on this topic (and other relevant MCMC topics) is available from Kass, Carlin, Gelman & Neal (1998).

Ways to improve convergence include blocking of parameters, i.e. joint updating of several parameters at once in order to reduce the dependency between iterates.

Estimators

There is a range of estimators available using the samples from the Markov chain. The so-called histogram estimator consistently estimates the posterior mean of the parameter vector by calculating the sample means of its iterates. An estimation of the posterior mean of functionals is in an analog way derived as

$$\mathbb{E}(h(\boldsymbol{\theta}_i|\mathbf{y})) \approx m^{-1} \sum_{j=j^*}^{J} h(\boldsymbol{\theta}_i^{[j]}|\mathbf{y}), \tag{2.30}$$

where h is a function and m indicates the number of draws used for inference. Here, $j^* = J-m+1$ is the number of the first sample after the burn-in period has been completed and that is assumed to come from the joint posterior.
Tierney (1994) and Gelfand & Smith (1990) discuss conditioning, also called Rao-Blackwellisation. This is available for those parameters, where the conditional density is known or at least its first moment, the conditional mean. An consistent estimator of the mean of the posterior density can then be derived by averaging the conditional means instead of the generated samples from this full conditional

$$\mathbb{E}(\boldsymbol{\theta}_i|\mathbf{y}) \approx m^{-1} \sum_{j=j^*}^{J} \mathbb{E}\left(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}^{[j]}, \mathbf{y}\right). \tag{2.31}$$

This so-called 'mixture' estimator has always smaller variance than the histogram estimator for independent samples (cf. Gelfand & Smith (1990)) and its efficiency for dependent MCMC samples depends basically on the correlation structure of the draws (cf. Tierney (1994)).
Similarly, if the entire conditional density is known, an estimator for the marginal posterior density $p(\boldsymbol{\theta}_i|\mathbf{y})$ is obtained by replacing the conditional mean in (2.31) by the complete full conditional density

$$p(\boldsymbol{\theta}_i|\mathbf{y}) \approx m^{-1} \sum_{j=j^*}^{J} p\left(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\setminus i}^{[j]}, \mathbf{y}\right).$$

This estimator uses the full conditional density and outperforms most general density estimators in estimating the tails of the density.

Finally, the prediction density $p(y^*|\mathbf{y}, x^*)$ can be estimated from the set of prediction densities conditional on the parameters $\boldsymbol{\theta}^{[j]}$. An consistent estimator is given by

$$p(y^*|\mathbf{y}) \approx m^{-1} \sum_{j=j^*}^{J} p\left(y^*|\boldsymbol{\theta}^{[j]}, \mathbf{y}\right).$$

Data augmentation

The data augmentation technique can be applied when e.g. the likelihood of the data can be represented in terms of an expectation with respect to some latent structure. While this expectation might not be easily computable, it may well be the case that the conditional likelihood, given the latent structure, is much simpler. More specifically, in some statistical problems the likelihood might be difficult to obtain. Whereas, it may be much easier and efficient to sample from a joint distribution, augmented with some auxiliary variable, than from the distribution of the model parameters alone. More formally, the likelihood $p(y|\boldsymbol{\theta})$ might not be available and consequently foreclosing a MCMC approach because the posterior comprises this likelihood

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

However, it may well be the case that introduction of an additional parameter of unknowns $\boldsymbol{\theta}^*$ makes the conditional likelihood $p(\mathbf{y}|\boldsymbol{\theta}^*, \boldsymbol{\theta})$ more handy and thus allowing for applying MCMC techniques via the augmented posterior

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}^*, \boldsymbol{\theta})p(\boldsymbol{\theta}^*, \boldsymbol{\theta}).$$

Treating the latent structure $\boldsymbol{\theta}^*$ as unobserved augmented data and sampling this data together with the other unknowns from the MCMC algorithm can be most appealing, since sampling the unknown parameters then only involves the conditional likelihood having the simpler form. The success of this approach relies heavily on a certain amount of ingenuity in identifying the latent parameter that can be used to simplify the scheme. Two examples

of data augmentation, which are also applied in this work, are given for a better understanding.
Albert & Chib (1993) present an approach for binary probit regression models using this auxiliary variable technique. Introducing a (truncated normally distributed) latent variable as dependent variable, and treating their 'observations' as additional unknown parameters, the conditional distributions of the other model parameters remain the same as in the Bayesian Gaussian linear regression model. The latent variable can here be viewed as utility/risk, whereas a high value of the variable means an increased probability of an event to occur.
The Bayesian way of measurement error correction, as used by e.g. Berry et al. (2002) in the context of flexible regression, can also be seen as data augmentation. By introducing the true but unobservable covariate observations as additional unknown parameters, the likelihood (based on these parameters) most conveniently remains a standard distribution easy to handle in the optimization scheme.

Bayesian model selection

If there is uncertainty about which model, from a set of competing models $M_1, M_2, \ldots$, describes the data generating process best, one may wish to compare these models. In this thesis the model alternatives will only differ in the specific sets of basis functions they adopt, while in general all model characteristics as prior distributions, error distributions, type of basis may specify a certain model.
The so-called Bayes' factor compares the marginal likelihood of the data under two competing model

$$BF_{ij} = \frac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)} \tag{2.32}$$

and one judges that model as the best one that has the highest Bayes' factor. Kass & Rafferty (1995) present a comprehensive treatment of Bayes' factors. Estimates of the marginal likelihood for each model in (2.32) can rarely be

obtained analytically, but Chib (1995) and Chib & Jeliazkov (2001) propose an approximation of this quantity directly from the MCMC output under only slight changes in the construction of the chain. However, when the set of models under consideration is huge the computational effort to compute all Bayes factors is immense. Finding an optimal set of basis functions from a huge pool of possible basis functions represents such a case.
Instead of searching over the complete model space, which is impractical for most problems, model selection can be done in a subspace of all models. Search algorithms include greedy searches, EM algorithms, simulated annealing and genetic algorithms, see Denison et al. (2002) and the references herein. The previous strategies attempt to find a single optimal model among those under consideration, however, one can rarely expect to replicate the truth with exactly one model among the proposed alternatives.

Bayesian model averaging

Another strategy, which will be used in the course of this thesis, accepts that none of the proposed models is true. A mixture model of all proposed alternative models is considered with the prior over each model formulating the relative degree of believe on each model, see e.g. Smith & Kohn (1996). This is sometimes called Bayesian model averaging (BMA). Bayesian model averaging accounts for model uncertainty the same way, as diversification of an investment portfolio accounts for the stock market uncertainties.
Let an auxiliary parameter vector $\boldsymbol{\gamma}^+$ index the different models. Samples of either, the parameters and model indicators, are drawn from the joint posterior $p(\boldsymbol{\theta}_{\boldsymbol{\gamma}^+}, \boldsymbol{\gamma}^+|\mathbf{y})$. In this thesis the potential model alternatives are characterized by the number and position of utilized basis functions and thus the vector $\boldsymbol{\gamma}^+ = (\gamma_1^+, \gamma_2^+, \dots \gamma_T^+)$, with $\gamma_i^+ = \{0, 1\}$ and model space $\Gamma = \{0, 1\}^T$ indicates whether a particular basis function, from a potential set of T basis functions, is in the model or not.
The most popular sampling scheme for model averaging is the reversible jump algorithm proposed by Green (1995). While the general Metropolis-Hastings

sampler requires a deterministic scan over all elements γ_i^+ in the vector $\boldsymbol{\gamma}^+$, the reversible jump algorithm can handle variable, i.e. non fixed, dimensions. This is an advantageous property since it allows for focusing entirely on the non zero elements in γ^+: basically, this algorithm proceeds similar to the Metropolis Hastings algorithm described in the paragraph 'Sampling schemes', except that $\boldsymbol{\gamma}^+$ is re-defined as $\boldsymbol{\gamma}$, which now contains the positions of the nonzero elements in $\boldsymbol{\gamma}^+$ – e.g. $\boldsymbol{\gamma} = \{1, 5, 6\}$ indicates that the first, fifth and sixth basis function is in the model. So, the dimension of $\boldsymbol{\gamma}$ is not fixed any longer. The acceptance probability needs to be modified by introducing a Jacobian term to take into account the change in dimension. In case of a discrete model space, as this is the relevant case in this thesis, this Jacobian term is not required as can be seen from Denison et al. (2002).
Once the sampler has converged, posterior averaging across the model space leads to the expectation of the predictive distribution

$$\begin{aligned} \mathbb{E}(y^*|x^*, \mathbf{y}) &= \sum_{s=1}^{2^T} \mathbb{E}(y^*|x^*, \mathbf{y}, \boldsymbol{\gamma}_s^+) p(\boldsymbol{\gamma}_s^+|D) \\ &\approx \frac{1}{m} \sum_{j=J-m+1}^{J} \mathbb{E}(y^*|x^*, \mathbf{y}, \boldsymbol{\gamma}^{[j]}). \end{aligned} \tag{2.33}$$

Here, $\boldsymbol{\gamma}_s^+$ denotes one particular specification of 0's and 1's in $\boldsymbol{\gamma}^+$ of 2^T possible specifications. The predictive distribution contains $\mathbb{E}(y^*|x^*, \mathbf{y}, \boldsymbol{\gamma}^{[j]})$, which is based on all visited models $\boldsymbol{\gamma}^{[j]}$ during the sampling scheme after the burn-in period. This approach is obviously motivated by a phrase delivered from the Greek philosopher Epicurus (ca. 341-271 BC) that says: „if more than one theory is consistent with the data, keep them all".

In the face of potential overfitting of the data by a large set of basis functions it is important to understand that the Bayesian framework contains a naturally penalty of complex models via the specification of priors over the coefficients. This is sometimes called Ockham's razor, see also MacKay (2003) for an introduction on this topic. It goes back to a phrase attributed to the Franciscan friar William of Ockham (ca. 1285-1349) „pluralitas non

est ponenda sine necessitate“, roughly translating to „plurality should not be posited without necessity“. The specific form of the prior distribution on the model coefficients strongly influences model selection. In case of very vague priors this leads to selection of the least complex model. This is an example of Lindley's paradox (cf. Lindley (1957)), which states that when comparing models of different complexity with diffuse priors on the model coefficients, then the simpler model is always favored, irrespective of the data. Choosing a 'good' prior on the coefficients is a sensible task as it controls the complexity of the favored model, comparable to a smoothing parameter. Denison et al. (2002) most interestingly explain the effect of prior choices on the final model fit.

2.2.2 The flexible Bayesian probit regression model

In the course of this thesis the Markov Chain Monte Carlo (MCMC) techniques of parameter estimation are only applied and further investigated in the case of binary regression. For the interested reader, a broad discussion on the Bayesian Gaussian model is given by Denison et al. (2002).

The binary regression case applied in this thesis follows the probit regression approach by Albert & Chib (1993), later refined by Holmes & Held (2006). Both apply data augmentation in order to obtain closed form conditional densities for parameter sampling. While the original sample consists of outcomes $y_i = \{0, 1\}$, a latent variable Z is introduced into the model

$$y_i = \begin{cases} 1 & : \text{ if } z_i > 0 \\ 0 & : \text{ otherwise} \end{cases} \tag{2.34}$$

$$z_i = \Phi(\mathbf{x}_i)\boldsymbol{\omega} + \epsilon_i \tag{2.35}$$

$$\epsilon_i \sim \mathrm{N}(\epsilon_i|0, 1). \tag{2.36}$$

Now, the stochastic auxiliary variable Z determines the response Y. In analogy to the RVM model setup in (2.5), $\Phi(\mathbf{x}_i)$ is the ith row of the design matrix including an intercept and the expansion of covariate(s) in terms of

radial basis functions. The appealing feature of data augmentation is that by introducing the latent variable Z all full conditionals take the form of standard distributions, as will be explained in the next section.

2.2.3 MCMC inference in the flexible binary case

In analogy to the RVM specification in Section 2.1.1, a multivariate Gaussian prior distribution is defined over the weights, whereas the exact form of the prior depends on the specific choice for model selection. There are two conceivable strategies: the first imitates the automatic relevance determination approach, described in Section 2.1.1 where hyperparameters need to be estimated that determine the precision of the weights. The second is associated with Bayesian model averaging, where these hyperparameters take on a fixed value a priori, however, additional indicator variables are introduced into the model controlling the complexity of the model. While the MCMC samples of the hyperparameters give merely an indication of the importance of the associated basis function, the indicator samples provide a clear declaration whether to include or to drop a basis at the current state of the MCMC sampling scheme. Both strategies are described in turn together with their respective sampling scheme in the following paragraphs.

Hyperparameter approach

The first approach follows the original RVM as presented in Section 2.1.1. A Gaussian prior distribution is specified over the weights

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{j=0}^{J} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{\alpha_j}{2}\omega_j^2\right), \tag{2.37}$$

with individual smoothing parameters assembled in the vector $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_J)^{\mathrm{T}}$ and a Gamma prior specified over these scales

$$p(\boldsymbol{\alpha}) \;=\; \prod_{j=0}^{J} \Gamma(a)^{-1} b^a \alpha_j^{a-1} \exp(-b\alpha_j), \tag{2.38}$$

where $\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t) \mathrm{d}t$.
If it is reasonable to include a dispersion parameter into the model to account for under-/overdispersion in the observed responses then it would also be necessary to specify a prior distribution for that parameter. This is however not realized in this work. In contrast to Section 2.1.1 the specification of the Gamma distribution must now be proper in order to receive a proper posterior distribution, a postulation which goes back to Casella & George (1992), who investigated the interaction between prior and posterior distribution.

Bayesian analysis of the probit regression model (2.34)-(2.36) now aims at finding the joint posterior distribution over the unknown parameters and auxiliary variable $p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \mathbf{z}|\mathbf{y})$. There is no analytic solution to this problem, but under certain conditions one is able to draw samples from this posterior distribution and so can determine an empirical version of the joint posterior. The Gibbs sampler is a standard sampler, where samples are alternately drawn from the conditional distributions of one parameter given the other(s); and these successively drawn samples form a Markov chain with the joint posterior being the (unique) invariant distribution of the chain, see Section 2.2.1.
The Gibbs sampler is completely specified by the so-called full conditional distributions. The priors for the coefficients and the scales have been chosen in a conjugate way, so that the full conditionals are recognizable as standard distributions. The usual way of deriving the full conditional of a certain parameter (vector) is recruiting those terms of the joint posterior distribution that contain this certain parameter (vector).
The joint posterior for the problem at hand is proportional to the product of the likelihood times the prior distributions

$$\begin{aligned} p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \mathbf{z}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{z}) p(\mathbf{z}|\boldsymbol{\omega}, \boldsymbol{\alpha}) p(\boldsymbol{\omega}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) \\ \Leftrightarrow p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \mathbf{z}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{z}) \times \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (z_i - \Phi(\mathbf{x}_i)\boldsymbol{\omega})^2 \right) \\ &\quad \times \exp\left(-\frac{1}{2} \sum_{j=0}^{J} \omega_j^2 \alpha_j - \sum_{j=0}^{J} b_j \alpha_j \right) \times \prod_{j=0}^{J} \alpha_j^{a-1}. \end{aligned}$$

By picking the relevant terms from the joint posterior distribution, the full conditional densities are recognized as standard distributions

$$\begin{aligned} p(\boldsymbol{\omega}|\boldsymbol{\alpha}, \mathbf{z}) \quad &= \quad \mathcal{N}(\mu_{\boldsymbol{\omega}|\cdot}, \Sigma_{\boldsymbol{\omega}|\cdot}) \\ &\text{where} \quad \mu_{\boldsymbol{\omega}|\cdot} = \Sigma_{\boldsymbol{\omega}|\cdot}\Phi^{\mathrm{T}}\mathbf{z} \\ &\qquad\quad \Sigma_{\boldsymbol{\omega}|\cdot} = (A + \Phi^{\mathrm{T}}\Phi)^{-1}, \end{aligned} \tag{2.39}$$

$$p(\alpha_j|\omega_j, \mathbf{z}) \quad = \quad Ga\left(a + \frac{1}{2}, b + \frac{1}{2}\omega_j^2\right) \tag{2.40}$$

$$\begin{aligned} p(z_i|\mathbf{z}_{-i}, y_i) \quad &\propto \quad \begin{cases} \mathrm{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i > 0) & \text{if } y_i = 1 \\ \mathrm{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i < 0) & \text{otherwise} \end{cases} \\ &\text{where} \quad \mu_{z_i|\cdot} = \Sigma_{z_i|\cdot}\Phi(\mathbf{x}_i)\mu_{\boldsymbol{\omega}|\cdot} - w_i z_i \\ &\qquad\quad \Sigma_{z_i|\cdot} = 1 + w_i \\ &\qquad\quad w_i = h_i/(1 - h_i). \end{aligned} \tag{2.41}$$

In (2.40), $Ga(\cdot)$ denotes the univariate Gamma density (cf. (2.38)). In (2.41), z_i denotes the current value for z_i and $\mathbf{z}_{-i}$ the vector $\mathbf{z}$ with z_i removed. The indicator function is denoted by $I(\cdot)$ indicating if the condition, given in parentheses, is valid. In $p(z_i|\mathbf{z}_{-i}, y_i)$, h_i denotes the ith diagonal element of the Bayesian hat matrix, $h_i = H_{ii}, H = \Phi\Sigma_{\boldsymbol{\omega}|\cdot}\Phi^{\mathrm{T}}$.

1 The usage of $|\cdot$ in the moments of the full conditionals generally indicates the conditioning on other unknown parameters and the observed data. This notational shortcut will be used for all MCMC approaches in this thesis since full conditional distributions are typically conditional on a large set of other unknowns.

2 Especially in the MCMC context, the notation $p(\cdot) =$ *type of distribution* will be used, when a lot of place can be spare by doing so, compared to giving the detailed specification of the respective distribution.

Controlling the weights by individual random scales α_j is inspired by the original RVM regression setup in Section 2.1.1. There, a type II maximum likelihood approach (cf. Good (1965)) is taken to optimize the scales and a scale is manually set to infinity when it exceeds a pre-specified threshold leading to pruning of the corresponding basis.
To mimic that approach in a MCMC setup brings an unexpected inconvenience: as mentioned earlier the user set parameters a, b from the Gamma prior over the scales (2.38) must be specified in order to give a proper prior distribution. This is in contrast to the original RVM specification, where both parameters have been set to zero being equivalent to an uniform prior on the log scale.
However, Chakraborty et al. (2005) follow this scheme and sample these hyperparameters as unknown parameters within a MCMC scheme. Doing so, firstly, a measure of the uncertainty of the scales can be given and secondly the predictive distribution automatically captures the uncertainty in the scales. This reflects a fully Bayesian analysis while on the other hand no selection of basis function is undertaken in their approach.
However, sparsity is a core characteristic of the RVM, but it remains an open problem how to transfer this concept from the original RVM to its MCMC mimicry adopting these scales. As for the original setting, many α_j's approach infinity during the optimization algorithm and as they exceed a certain threshold the corresponding basis functions are cut from the model. A comparable strategy in the MCMC case would be to base the decision whether a basis is relevant or not on the posterior distribution of the α_j's. This means assessing the complete set of basis functions and finally labeling these functions non-relevant, where the empirical mean of the related scale parameter exceeds a certain threshold. Though ending up with a post-sampling indicator of the relevance of a basis, the prediction function is still based on the complete set of basis functions. And from a computational perspective even worse, each MCMC loop is based on the full model instead of only a subset of functions. This property is a pronounced burden in the calculation of e.g. the moments in the full conditional density of the weights (2.39), since $\Sigma_{\boldsymbol{\omega}|\cdot}$ is a $(J+1) \times (J+1)$ matrix that needs to be inverted.

An alternative ad-hoc strategy is to decide within every MCMC loop, based on the actual sampled values α_j, whether to keep or to discard the corresponding basis. The main difficulty is, to specify a concrete threshold to include or remove a basis.
After one of the previous described strategies is chosen and values of the Gamma prior and starting values for the unknown variables have been set, samples from (2.39), (2.40) and (2.41) are drawn in turn. After having collected a large number of samples, the initial set of samples (burn-in) is discarded, leaving inference to the remaining samples assumed to come from the joint posterior distribution. Several estimators can be applied to this sample, as explained in Section 2.2.1.

The first strategy to handle the hyperparameters is not designed to simplify the model during the MCMC algorithm, while the second founds the basis elimination on an arbitrary threshold. One usually prefers to have a clear indicator whether to add or to remove a basis function that makes specification of a threshold needless. The following section describes a MCMC version of the RVM, which accomplishes this claim with the aid of Bayesian model averaging.

Indicator approach - Bayesian model averaging

Determination of relevance is here no longer accomplished by the random hyperparameters α_j. Instead, these hyperparameters are replaced by a single pre-specified value $1/v$, which is not subject to randomness, in the weights prior (2.37).
Bayesian model averaging, as described in Section 2.2.1, assumes that there is no unique optimal model, but instead the best approximation to the truth is a mixture of models from some model space. In the course of this thesis a particular component model is fully characterized by the set of utilized basis functions as indicated by the parameter vector $\boldsymbol{\gamma}^+$. The respective part of

the Bayesian probit model (2.35) is then given as

$$z_i = \Phi_{\gamma^+}(\mathbf{x}_i)\boldsymbol{\omega}_{\gamma^+} + \epsilon_i,$$

where the vector of coefficients is now modified to $\omega_{\gamma^+} = (\omega_0\gamma_0^+, \omega_1\gamma_1^+, \ldots, \omega_T\gamma_J^+)^{\mathrm{T}}$ and $\gamma_j^+ = \{0, 1\}, j = 1, \ldots, J+1$, such that $\gamma_j^+ = 1$ if the jth basis function is in the model and $\gamma_j = 0$ if it is not. Hereby, $J+1$ denotes the size of the complete stock of potential basis functions, while usually only a maximum set of size T, with $T \leq J+1$ is admitted in the model. This reflects the belief that a wide range of different basis function is needed to fit a wide range of functional forms, while normally a subset of only few selected basis functions represents the truth quite decently. $\Phi_{\gamma^+}(\mathbf{x}_i)$ is defined in an analog way to the weights, with the indicators working on the columns of the design matrix representing the basis functions.
In this formulation all elements in $\boldsymbol{\gamma}^+$ are unknowns having to be sampled during the sampling scheme. A MH-step would scan over all elements in $\boldsymbol{\gamma}^+$ and sample each individual γ_j^+. If the number of possible basis functions is very large the computational effort to approximate the posterior density would be immense and impractical. A more convenient way is to redefine $\boldsymbol{\gamma}^+$ as $\boldsymbol{\gamma} = \{\text{locations of nonzero elements in } \boldsymbol{\gamma}^+\}$. Typically, the dimension of $\boldsymbol{\gamma}$ is much smaller, however, its dimension is no longer fixed and thus a sampling scheme allowing for variable dimensions is needed, which is the reversible jump algorithm.
A prior distribution on the basis function set is adopted via a prior over $\boldsymbol{\gamma}$ where Denison et al. (2002) suggest application of a discrete uniform that takes

$$p(\boldsymbol{\gamma}) = \binom{J+1}{dim(\boldsymbol{\gamma})}^{-1} \times \frac{1}{T+1}. \tag{2.42}$$

Here, $dim(\boldsymbol{\gamma})$ denotes the number of elements in $\boldsymbol{\gamma}$, $J+1$ the number of the candidate locations for the basis functions, and T the maximum number of basis functions allowed. The '+1' in 'J+1' comes from the intercept in the model and the '+1' in 'T+1' means that also an empty model with $dim(\boldsymbol{\gamma}) = 0$ is allowed. The maximum number of basis functions T is typically chosen so large that it effectively does not affect the posterior. Given

model dimension $dim(\boldsymbol{\gamma})$, each combination of $dim(\boldsymbol{\gamma})$ basis functions from the stock of $J+1$ candidates has equal probability and each model complexity is equally likely. This prior does not place an explicit penalty on the model complexity. However, as described in connection with Bayesian model selection/averaging, the marginal likelihood already contains a penalty on the dimension, which depends on the prior variance of the coefficients (cf. Ockham's razor in Section 2.2.1). So this uniform prior seems adequate and will be chosen throughout this thesis, while also alternative types, like e.g. Poisson and truncated geometric prior are used in practise.
The reversible jump algorithm (cf. Green (1995)), which is based on the Metropolis-Hastings sampler, is briefly described in the following. Only the two move types BIRTH and DEATH are applied here to traverse the posterior probability surface, naming jumps in higher and lower model dimensions, respectively. The current model is assumed to be of dimension $t := dim(\boldsymbol{\gamma})$.

BIRTH. Proposal of adding a randomly chosen new basis (including intercept) from those that are not present in the current model with probability b_t.

DEATH. Proposal of removing a randomly chosen basis (including intercept) from those that are present in the current model with probability d_t.

The proposal probabilities depend on the current complexity of the model, denoted by t, in that sense that they are chosen as $b_t, d_t = 0.5$ for $0 < t < T$ and $b_0, d_T = 1$ and $b_T, d_0 = 0$. A specific basis function to be born or to die is randomly selected with uniform probability from the basis functions currently being excluded or contained, respectively.
Under this specification, the acceptance probability of a proposed move from model $\boldsymbol{\gamma}$ of dimension t to model $\boldsymbol{\gamma}'$ of dimension t' is

$$\alpha = \min\left\{1, \frac{p(\mathbf{z}|\boldsymbol{\gamma}')}{p(\mathbf{z}|\boldsymbol{\gamma})} \times R\right\}, \tag{2.43}$$

with the ratio of the marginal likelihoods multiplied by a constant R, where R is the ratio of probabilities given by d'_t/b_t for a BIRTH and b'_j/d_j for a

DEATH. The specific form of the acceptance probability (2.43) is due to the marginalization of (i.e. integration over) the weight vector allowing for a block update of $\boldsymbol{\gamma}$ together with $\boldsymbol{\omega}$. This block update proceeds by first generating a new sample of $\boldsymbol{\gamma}$ and then generating the new coefficients for these sampled basis functions. For details on deriving this expression see Denison et al. (2002) and Ranyimbo & Held (2006).
The specific from of the marginal likelihood is presented below since it relies on the moments from the full conditional distribution of the weights. By realization of a uniform random number u the proposed jump from t to t' dimensions is accepted if $u < \alpha$, where α is the acceptance probability (2.43). A very appealing feature of this method is that the model complexity is small with high probability if the data suggest a simple model.

Samples of the unknown parameter vector $\boldsymbol{\omega}_{\boldsymbol{\gamma}}$ and the observations from the latent variable $\mathbf{z}$ are obtained from the conditional densities (2.39) and (2.41), but now based on the model as defined by the model selection parameter $\boldsymbol{\gamma}$. It is stressed here again, that the former hyperparameter vector $\boldsymbol{\alpha}$, is now replaced by a pre-specified and fixed prior precision v^{-1} (i.e. prior variance v) in the prior over the weights (2.37), and is not subject any form of uncertainty or randomness. Consequently, it does not have to be sampled in the following sampling scheme:

$$
\begin{aligned}
p(\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\mathbf{z}) \quad &= \quad \mathrm{N}(\mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}, \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}) && (2.44)\\
&\text{where} \quad \mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} = \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}\Phi_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z} \\
&\qquad\quad \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} = (v^{-1}\mathbf{I} + \Phi_{\boldsymbol{\gamma}}^{\mathrm{T}}\Phi_{\boldsymbol{\gamma}})^{-1}, \\
p(z_i|\mathbf{z}_{-i}, y_i) \quad &\propto \quad \begin{cases} \mathrm{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i > 0) & \text{if } y_i = 1 \\ \mathrm{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i < 0) & \text{otherwise} \end{cases} && (2.45)\\
&\text{where} \quad \mu_{z_i|\cdot} = \Sigma_{z_i|\cdot}\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_i)\mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} - w_i z_i \\
&\qquad\quad \Sigma_{z_i|\cdot} = 1 + w_i \\
&\qquad\quad w_i = h_i/(1 - h_i).
\end{aligned}
$$

Here, h_i is the ith diagonal element of the Bayesian hat matrix, $h_i = H_{ii}, H = \Phi_{\boldsymbol{\gamma}}\Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}\Phi_{\boldsymbol{\gamma}}^{\mathrm{T}}$ and $I(\cdot)$ again denotes the indicator function.

Finally, the marginal likelihood which is used in the acceptance probability of the $\boldsymbol{\gamma}$-sampling scheme (2.43) is based on the moments of the full conditional of the parameter vector $\boldsymbol{\omega}_{\boldsymbol{\gamma}}$ (2.39). It is given by

$$p(\mathbf{z}|\boldsymbol{\gamma}) = (2\pi)^{-\frac{N}{2}} \frac{\mid \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \mid^{\frac{1}{2}}}{\mid v_{\boldsymbol{\gamma}} \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\mathbf{z}^{\mathrm{T}}\mathbf{z} - \mu^{\mathrm{T}}_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \Sigma^{-1}_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}\right)\right), \tag{2.46}$$

where the exponential term $\mathbf{z}^{\mathrm{T}}\mathbf{z}$ cancels out in the acceptance probability. Here, $v_{\boldsymbol{\gamma}}$ denotes the prior covariance matrix over the weights for a model as defined by $\boldsymbol{\gamma}$.

Model selection occurs here in each cycle of the sampling scheme. Yet, not a single most probable model is sought, but instead, following the Epicurean spirit described in the respective paragraph of Section 2.2.1: every type of model that is able to explain the data is considered in order to make the final predictions. This is called Bayesian model averaging and yields prediction by averaging the results over all models visited during the sampling scheme. The different models are characterized by the parameter vector of varying length $\boldsymbol{\gamma}$, naming these basis functions that are in the model.
The histogram estimator (cf. (2.30)) for the posterior mean prediction y^* at an unseen $\mathbf{x}^*$ is based on the samples $y^{*[j]} := G\left(\Phi(\mathbf{x}^*)\boldsymbol{\omega}^{[j]}\right)$, where $G(t)$ is the Gaussian cumulative distribution function. It is given by

$$\mathbb{E}(y^*|\mathbf{x}^*, \mathbf{y}) \approx \frac{1}{m} \sum_{j=J-m+1}^{J} (y^{*[j]}|\mathbf{y}, \mathbf{z}^{[j]}, \boldsymbol{\omega}^{[j]}\boldsymbol{\gamma}^{[j]}), \tag{2.47}$$

where it is made explicitly that the samples $y^{*[j]}$ are conditioned on the current samples for $\mathbf{z}^{[j]}$, $\boldsymbol{\omega}^{[j]}$ and $\boldsymbol{\gamma}^{[j]}$.
The right hand side of (2.47) reflects the approximation of the integral over the parameter space $(\boldsymbol{\omega}, \mathbf{z})$ and the model space $\boldsymbol{\gamma}$. Here, the final estimating function is based on potentially different subsets of basis functions in each MCMC loop. This contrasts the original RVM where the final prediction model displays reduced complexity.
A slight modification of the estimator (2.47) lies in computing the posterior

mean estimator for the linear predictor $z^* := \Phi(\mathbf{x}^*)\boldsymbol{\omega}$ via a histogram estimator and then to link this estimate to the desired probability by using the response function $G\left(\widehat{\mathbb{E}}(z^*|\mathbf{y})\right)$. While this estimator is equivalent to (2.47) in the Gaussian response case, it is slightly different otherwise. This latter estimator will be used in the simulation study of the binary case (cf. Section 4.1.4).

2.3 Covariate measurement error and its correction

So far, all of the covariates have been assumed to be collectable without any error. A more realistic view is to allow for error-prone covariates, where sources of error include e.g. imperfect measurement devices or defective operationalization of factors. Measurement error of the response is not discussed here, since it is usually much less problematic.
Popular examples of covariate measurement error are described by Fuller (1987) and Carroll et al. (1995), ranging from bioassay experiments with plants to the investigation of earthquakes. Particularly in the area of econometrics this problem has received a lot of attention (cf. Schneeweiß (1990)) where relevant econometric variables, like e.g. intensity of motivation, are hardly observable. Also in the field of medicine and epidemiology this error has generated major research interest (cf. Willett (1998)) where e.g. individual exposures to a certain radiation or nutrition habits of study participants need to be recorded and their influence on disease is investigated.
It is stressed, that there is almost no kind of measurement that is free from potential measurement error. This error may seem negligible in some cases, however, in a lot of cases it is not and the impact of this error on the analysis is indisputable in the existent literature (cf. all of the references given above). Statistical analysis ignoring such inherent error is referred to as 'naive analysis' and, for instance, Carroll et al. (1999) emphasize that when measurement error is ignored „conventional parametric and nonparametric techniques are

no longer valid". That means, the parameter estimates from that so-called naive analysis, where measurement error is ignored, are usually biased.
Statistically speaking, the fundamental parameters $\boldsymbol{\omega}$ in the 'ideal mean model', formulated in the true, however, unobservable covariate ξ

$$\mathbb{E}(Y|\xi) = f(\xi, \boldsymbol{\omega}) \tag{2.48}$$

are usually not retained in the 'observed mean model', where the error-prone covariate X replaces ξ

$$\mathbb{E}(Y|X) = f(X, \boldsymbol{\omega}^*), \tag{2.49}$$

with $\boldsymbol{\omega} \neq \boldsymbol{\omega}^*$.

Note:
The fundamental mean model parameters collected in $\boldsymbol{\omega}$ are here and in the remainder of this section considered as being non-random and thus the moments need not to be conditioned on $\boldsymbol{\omega}$. This frequentistic perspective as sketched in the paragraph 'Connection between RVM inference and penalized likelihood estimation' in Section 2.1.2 is inherent in all calibration methods applied to the RVM concept.

The following paragraphs summarize some 'classical' as well as very recent approaches to error correction briefly. The reader, not yet familiar with the terminology in that area is recommended to skip this at first reading and resume reading where the models for measurement are introduced.

There are a range of approaches to error correction being employed in (non-) linear regression. Among these are regression calibration which was suggested by Carroll & Stefanski (1990) and Gleser (1990) and an approach based on simulation and extrapolation called SIMEX by Cook & Stefanski (1994). Though both methods usually do not yield consistent estimators, they are able to reduce the measurement error induced bias to a great extent, investigated e.g. by Wolf (2004). These two methods are the first ones to be readily implemented in a statistical software package, which may also speak for their popularity, see `http://www.stata.com/merror/` for further

information. An extensive description of both methods is also provided by Carroll et al. (1995).
Also likelihood methods, which are particularly appealing because of the optimality properties of maximum likelihood estimates, have been investigated (cf. Carroll et al. (1995), Schafer & Purdy (1996), Küchenhoff & Carroll (1997)). The likelihood of the observed data can, except for the Gaussian and probit regression model (cf. Fuller (1987), Carroll & Gallo (1984)), not be obtained analytically and usually requires numerical integration – a fact that is, however, less appealing.
A very recent approach by Carroll, Midthune, Freedman & Kipnis (2006) investigates 'seemingly unrelated regression' to obtain more precise estimates of the inherent attenuation, i.e. deflation to zero, when basing analysis on error-prone data. Being interested in the attenuation introduced by e.g. error-prone protein intake data and having a second related error-prone variable like e.g. energy intake available, fitting simultaneous measurement error models for protein and energy may result in better estimates for the systematic bias of the protein measurement tool.
However, particularly in the flexible regression case the problem of covariate measurement error has not received extensive attention, yet. A most discouraging result has been stated by Fan & Truong (1993) who investigated consistent estimators, which have the desired property of deviating from the true function on a compact interval with probability zero when the sample size goes to infinity. A crucial point for practical usefulness is the rate of convergence describing the order in terms of the sample size N an estimator approaches the true function. They found that the optimal rate of convergence of an consistent nonparametric estimator under measurement error is $(\log(N))^2$, which is impractically slow.
In a very recent work on that subject, Schennach (2004) presents a $\sqrt{N}$ consistent estimation of nonlinear models with covariate measurement error. The integral equations relating the distribution of true and observed variable are converted into algebraic equations by Fourier transform. Resolving these equations allows to identify arbitrary moments of the true unobservable variable which can then be used in moment based estimators like e.g.

nonlinear least squares or general extremum estimators. The key to the fast $\sqrt{N}$ convergence rate is here that the proposed moment estimator contains a function, which is able to downweight the noisy tail of the estimated characteristic function of the latent ξ.
Another possible resort to Fan & Truong (1993) is the consideration of approximately consistent estimators in flexible, yet parametric, subclasses of the nonparametric family. The RVM is a member of this subclass among others, like e.g. regression splines Carroll et al. (1999) describe a correction method, based on the popular regression calibration approach, for flexible regression using (penalized) regression splines. They also present a SIMEX version of kernel estimators and spline smoothing.
A Bayesian approach to covariate measurement error in flexible regression is suggested by Berry et al. (2002) based on MCMC sampling of the true but unknown covariate observations together with the other unknowns. The basic principle of Bayesian error correction goes back to Richardson & Gilks (1993a) and Richardson & Gilks (1993b), and an overview is given by Richardson (1996).

For the sake of simplicity, only a single error-prone variable will be considered in the following sections. Extension to the case of multiple covariates is straightforward as long as there are no dependencies between the covariates – a strong restriction indeed. The case of multiple covariates in flexible regression is more complicated and, to the author's knowledge, has been studied in the existent literature only by Ganguli, Staudenmayer & Wand (2005). They develop a likelihood-based method for fitting additive models where random coefficient penalized splines are used to estimate the smooth functions. The problem of analytically intractable integrals in the likelihood is overcome by using a nested Monte Carlo Expectation Maximization algorithm. There, the samples from the unknown covariate are drawn in a Metropolis Hastings step and then used in a nested 'expectation maximization' (EM) algorithm.
Throughout this thesis it is further claimed that the error-prone variable is continuous. Measurement error in categorical variables, typically called misclassification, is discussed by Küchenhoff, Mwalili & Lesaffre (2005) and is

beyond the scope of this thesis.

Firstly, an overview of the standard measurement error models is given and then the new correction methods for flexible regression are motivated and briefly described in order to elaborate how the different strategies are connected and in which respects they differ.

2.3.1 Models for measurement error

The fundamental prerequisite for error correction is the specification of an error model, which relates the observed error-prone variable X to the true variable ξ one would ideally like to observe. Two general types of error processes as presented by Carroll et al. (1995) are illustrated in the following.

The classical additive error

The concept of classical measurement error is appropriate, when e.g. one tries to determine ξ, but the technical device does not allow for a correct measurement. Measurement deviates randomly from the true value, but is expected to be correct on average. A common model for that type of error process links the true and the observed covariate, sometimes also called surrogate variable, in an additive way

$$X = \xi + \delta, \quad (\delta, \xi) \sim \text{indep.}, \quad \mathbb{E}(\delta|\xi) = 0, \tag{2.50}$$

which is frequently extended to $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ and $\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$. This assumption will be used later for several correction methods, e.g. basis function calibration and structural quasi likelihood. A more flexible specification for the distribution of ξ is the mixture of normals as applied in Davidian & Gallant (1993) and Carroll et al. (1999). While in model (2.50) the variable X denotes a single error-prone measurement of ξ, it may generally be the case that a series of replicate measurements are available. Apart from the particular usage of X in the error model (2.50) above, the variable X may

just as well represent the average over m possible replicate measurements, which is an unbiased estimator for ξ in model (2.50). Of particular interest for some of the later error correction methods, is the conditional density of ξ given X. For δ and ξ being Gaussian, this conditional distribution is given as

$$\begin{aligned} p_{\xi|X} &= \frac{1}{\sqrt{2\pi}\sigma_{\xi|X}} \exp\left(-\frac{1}{2}\frac{(\xi-\mu_{\xi|X})^2}{\sigma^2_{\xi|X}}\right) \\ & \mu_{\xi|X} = \mu_\xi + \lambda\cdot(X-\mu_\xi), \quad \lambda := \frac{\sigma^2_\xi}{\sigma^2_\xi+\sigma^2_\delta/m} \\ & \sigma^2_{\xi|X} = \frac{\sigma^2_\xi\sigma^2_\delta/m}{\sigma^2_\xi+\sigma^2_\delta/m}, \end{aligned} \tag{2.51}$$

where the surrogate variable X may now stand for an average over m possible replicate measurements.

If the real data provides m_i replicate measurements of ξ_i for subject i, the measurement error variance can be estimated from the data by the usual components of variance analysis

$$\widehat{\sigma}^2_\delta = \frac{\sum_{i=1}^N \sum_{j=1}^{m_i}(x_{ij}-x_i)^2}{\sum_{i=1}^N (m_i-1)}, \tag{2.52}$$

with $x_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}$ denoting the average over a subjects's replicates.

Furthermore, μ_ξ and σ^2_ξ can be consistently estimated from the sample using the analysis of variance formulas

$$\begin{aligned} \widehat{\mu}_\xi &= \sum_{i=1}^N m_i x_i / \sum_{i=1}^N m_i = \widehat{\mu}_X \\ \widehat{\sigma}^2_\xi &= \left[\sum_{i=1}^N m_i(x_{i\cdot}-\mu_X)^2 - (N-1)\widehat{\sigma}^2_\delta\right]/\nu \\ \nu &= \sum_{i=1}^N m_i - \sum_{i=1}^N m_i^2 / \sum_{i=1}^N m_i. \end{aligned} \tag{2.53}$$

The classical additive error model (2.50) is despite its simplifying character the prevalent case to be considered when new correction methods are studied. It is well known in the literature that even this classical type of error,

expected to be zero on average, leads to biased parameter estimates in a regression context. More general error models account for the observed variable possibly being biased, which might be caused by some other variables or an intercept, representing e.g. erroneous gauging (cf. Carroll et al. (1995)).

The Berkson error

A slightly different model renders the true variable being the result from the observed variable plus some independent random deviation. This seems reasonable when e.g. individual exposure to a certain radiation is measured by a stationary recording device. These measurements typically do not reflect one's personal individual exposure, but merely represent one determinant of exposure. The true individual exposure varies among persons according to their personal habits of being outdoor, ventilating the rooms, etc.
These foregoing considerations are captured in the model

$$\xi = X + \delta, \quad (\delta, X) \sim \text{indep.}, \quad \mathbb{E}(\delta|X) = 0, \tag{2.54}$$

which is frequently extended to $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$. Here in (2.54), X denotes again a single error-prone observation of ξ, though it may generally be the case that a series of replicate measurements are available. If not stated otherwise, the variable X may just as well represent the average over m available replicate measurements, as in the following specification of the conditional density of the latent variable given the observation(s)

$$p_{\xi|X} = \frac{1}{\sqrt{2\pi}\sigma_\delta} \exp\left(-\frac{1}{2}\frac{(\xi - X)^2}{\sigma_\delta^2}\right). \tag{2.55}$$

The Berkson model is a special case of the more general 'regression calibration model' additionally allowing for a bias, which can be introduced into the model by an intercept and possibly further covariates (cf. Carroll et al. (1995)).

It is not always clear, which error model to use and then the choice between them is made on the basis of convenience. In practical applications one

may also find hybrid forms of both error models. Heid, Gerken, Wellmann, Küchenhoff, Kreienbrock & Wichmann (2002) investigate the effect of radon exposure and lung cancer where measurements of the exposure come from two sources: a questionnaire and a measuring instrument. Either contains substantial measurement error, however the type of error differs between both methods.

2.3.2 Methods for error correction

After a sensible choice of the error model has been made, there are several approaches to measurement correction depending on the error model and the type of statistical analysis. In this thesis, the focus lies on error correction for flexible regression for Gaussian, binary and Poisson responses and the error is assumed to come from the classical error model (2.50) unless stated otherwise.

This work develops a number of new approaches that are highlighted here, briefly. The so-called 'basis function calibration' method is a transformation of the structural regression splines (cf. Carroll et al. (1999)) to suit the popular radial basis functions (RBF) applied by the relevance vector machine (RVM).
Furthermore, an exact structural quasi likelihood method for flexible Gaussian regression will be developed and additionally a refined approximative method in the spirit of expanded regression calibration for flexible regression in non-Gaussian cases (cf. Carroll et al. (1995)).
The nonparametric SIMEX as presented by Carroll et al. (1999) will be adopted here for the RVM regression.
The core idea of Bayesian measurement error correction, as described by Richardson (1996), will be introduced into the MCMC version of the RVM from Section 2.2.3.

The major part of the methods applied in this thesis are roughly described in

the following, while a detailed description follows in the respective chapter, where they are used for the first time.
However, this work also develops 'basis function calibration' for two error-prone covariates and an approach for measurement error correction in a flexible model for binary longitudinal data. The details on these latter, very specific, methods are postponed to chapter 4.

Standard regression calibration

The main purpose of regression calibration is to find an approximation to the observed mean model $\mathbb{E}(Y|X)$ in (2.49), while retaining the fundamental model parameters $\boldsymbol{\omega}$ of the ideal mean model $\mathbb{E}(Y|\xi)$ in (2.48) (cf. Carroll et al. (1995)).
This approximation is implemented by replacing the latent ξ in the ideal model by its expectation given the surrogate, i.e. $\mu_{\xi|X} = \mathbb{E}(\xi|X)$, which yields

$$\mathbb{E}(Y|X) \approx f\left(\mu_{\xi|X}, \boldsymbol{\omega}\right).$$

The basic idea is that under small measurement error ξ will be close to its expectation $\mu_{\xi|X}$. However, even with small measurement error, ξ may not be close to X. Thus naively replacing ξ by X may lead to large bias, hence the need for calibration. Though standard regression calibration itself is not applied in this work, it is the foundation of basis function calibration and structural quasi likelihood, both discussed later. This justifies the following presentation, which is a bit more detailed compared to the other methods presented in this chapter.

Regression calibration is a pre-procession of the covariate observations and these modified observations are then used for the analysis. This analysis, however, is then performed by the standard routine as if no measurement error was in the data. Each true but unobservable covariate observation ξ_i[1]

[1]Although the variable ξ is latent, i.e. *unobservable*, this work occasionally needs to refer to ξ_i and therefore uses the term *observation*. This pun is not intended, but necessary.

is replaced by $\mu_{\xi|x_i} = \mathbb{E}(\xi|X = x_i)$ and then the standard analysis is carried out. If there are replicate measurements available, then X and the realization x_i may again represent the average of these replicates.
The desired quantities $\mu_{\xi|x_i}, i = 1, \ldots, N$ are readily derived for the Berkson error model from (2.55) as

$$\mu_{\xi|x_i} = x_i. \tag{2.56}$$

And for the classical error model with ξ and δ being normally distributed they are straightforwardly derived from (2.51) as

$$\mu_{\xi|x_i} = \mu_\xi + \lambda \cdot (x_i - \mu_\xi), \quad \lambda := \frac{\sigma^2_{\xi|x_i}}{\sigma^2_\delta/m} \tag{2.57}$$

$$\sigma^2_{\xi|x_i} = \frac{\sigma^2_\xi \sigma^2_\delta/m}{\sigma^2_X} = \frac{(\sigma^2_X - \sigma^2_\delta/m)\sigma^2_\delta/m}{\sigma^2_X}, \tag{2.58}$$

where λ has been slightly re-arranged compared to (2.51). The error variance σ^2_δ needs to be known or estimated from e.g. validation/replication data using formula (2.52). The mean of the latent variable μ_ξ can be calculated from its analysis of variance formula (2.53).
This approach yields consistent parameter estimates in the linear regression model. In other non-linear regression cases this is merely an approximate working model for the observed data which can be checked via residual plots and possibly modified accordingly.
Since assumptions about the distributions of ξ are necessary to perform regression calibration this method is categorized as a 'structural method' in contrast to 'functional methods', which do not need any distributional assumptions about ξ. Carroll et al. (1995) present the more general case of multiple, possibly error-prone, correlated covariates and generalizations to nonlinear and nonadditive measurement error.

Adopting that strategy for the relevance vector machine regression reveals an interesting aspect of how regression calibration modifies the radial basis functions (RBF). This is shortly highlighted in the following two paragraphs.

Regression calibration and basis function modification

The ideal RVM mean model is

$$\mathbb{E}(Y|\xi) \;=\; G\left(\sum_{j=1}^{J}\omega_j\phi_j(\xi)+\omega_0\right). \tag{2.59}$$

The replacement of ξ by its calibrated version $\mu_{\xi|X}$ in effect proposes the following approximation to the observed mean model under retainment of the true mean model parameters $\omega_0,\ldots,\omega_J$

$$\mathbb{E}(Y|X) \;\approx\; G\left(\sum_{j=1}^{J}\omega_j\phi_j(\mu_{\xi|X})+\omega_0\right). \tag{2.60}$$

Through this work, the focus lies on the classical measurement error model (2.50). Here is, how standard regression calibration for that case affects the RVM.
Straightforward use of the calibrated covariate $\mu_{\xi|X}$ from (2.57) modifies each univariate radial basis function $\phi_j(\xi)=\exp(-\eta(\xi-c_j)^2)$ in the ideal model (2.59) into

$$\begin{aligned}\phi_j\left(\mu_{\xi|X}\right) &= \exp(-\eta(\mu_{\xi|X}-c_j)^2)\\ &= \exp(-\eta\cdot(\mu_\xi+\lambda\cdot(X-\mu_\xi)-c_j)^2),\end{aligned}$$

designated to be utilized in the working model (2.60). The kernel parameter η determines the width of the Gaussian basis and the knots c_j define the position on which the jth basis function is centered.
Following the original RVM setup (cf. Tipping (2000)), using a RBF type kernel with knots given by the N covariate observations $x_i, i=1,\ldots,N$ (now subject to calibration!) and assuming the same number of replicates for all observations, which means that λ in (2.57) is the same for all observations, the jth basis function at position $c_j=x_j$ is now given as

$$\begin{aligned}\phi_j\left(\mu_{\xi|x_i}\right) &= \exp(-\eta^2(\mu_{\xi|x_i}-\mu_{\xi|x_j})^2)\\ &= \exp(-(\eta\cdot\lambda)^2\cdot(x_i-x_j)^2).\end{aligned} \tag{2.61}$$

Thus, the inference of the weight parameters applying standard regression calibration is equivalent to a naive analysis using a modified basis kernel with

parameter $\eta^* = \eta \cdot \lambda$, instead of η.
What is the intuition behind error correction via basis modification and how does this modification affect the quality of error correction? The following paragraph describes the mechanism behind regression calibration, first on the basis of classical linear regression and finally generalizes to the flexible RVM regression.

A note on the mechanism behind regression calibration

The connection between regression calibration and modification of the utilized basis function reveals an appealing new perspective on how regression calibration works. By choosing the example of classical linear regression, this idea is easily put across.
Figure 2.2 displays the true linear relationship between covariate ξ and response Y and observations (x_i, y_i) generated from that functional form under response error and covariate measurement error. Naive linear regression can

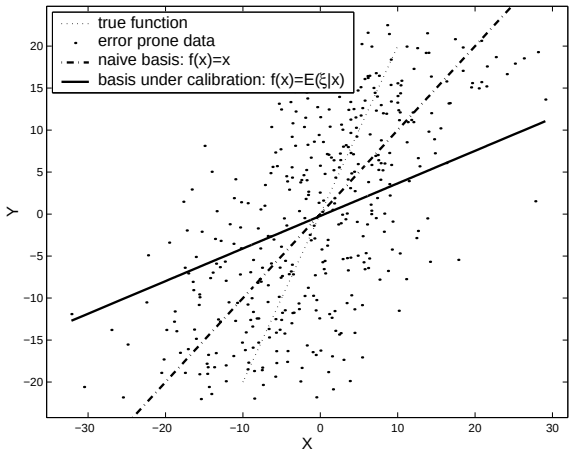

Figure 2.2: Classical linear regression can be seen as fitting a basis function to the observed data. When covariate measurement error is introduced, naive estimation and regression calibration exclusively differ in the type of basis function they fit to the data.

now be viewed as fitting a linear basis $f(x) = x$ to the observed data. This

basis function is the steep, 45 degree line in the figure.
In an analog way, regression calibration in linear regression can then be understood as fitting the more flattened linear basis $f(x) = \mathbb{E}(\xi|x)$ also displayed in Figure 2.2 to the observations. Intuitively, the estimated coefficient under calibration needs to be larger then the naive one in order to fit that flattened basis properly to the observed data. Indeed, the naive estimate and that under calibration for this example are $\widehat{\beta}_{naive} = 0.7073$ and $\widehat{\beta}_{calib} = 1.8099$, respectively. In linear regression and parametric regression in general, the accuracy of both estimates is assessed by simply comparing the parameter estimates $\widehat{\beta}_{naive}$ and $\widehat{\beta}_{calib}$ with the true regression coefficient, which is $\beta_{true} = 2$ in this example.
However, in the case of flexible regression the true functional form is not necessarily in the scope of the applied, flexible model. The quality of the estimates is then checked by comparing the true underlying function with the prediction functions from the naive and corrected analysis, respectively. Prediction means plugging in the parameter estimates in the ideal mean

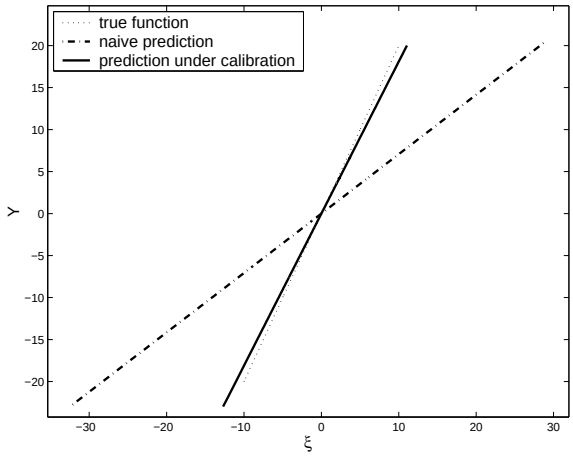

Figure 2.3: The quality of naive and corrected analysis in the flexible regression is conventionally judged by the accuracy of prediction both methods achieve. Here, this principle is visualized for the classical linear regression.

model and evaluating for a set of yet unseen covariate observations that are free from measurement error: for the linear regression example, discussed

here, this means plugging in $\widehat{\beta}_{naive}$ and $\widehat{\beta}_{calib}$, respectively, into the ideal (classical regression) model $Y = \xi\beta$ and then evaluating at a set of predetermined points $\xi_k, k = 1, \ldots, K$. Figure 2.3 displays the true, the naive and the corrected (under regression calibration) estimated relationship between ξ and Y. It can be seen that regression calibration leads to a much steeper estimate of the true relationship, which is obviously more accurate than the naive attenuated estimate.
This principle of calibration fitting a more flattened basis type to the observed data also carries over to the nonlinear RVM regression: the basis under calibration (2.61) is more flattened (less peaked) than the original basis because $\eta^* < \eta$ for $\sigma_\delta^2 > 0$, as can be seen from Figure 2.4. A simulation

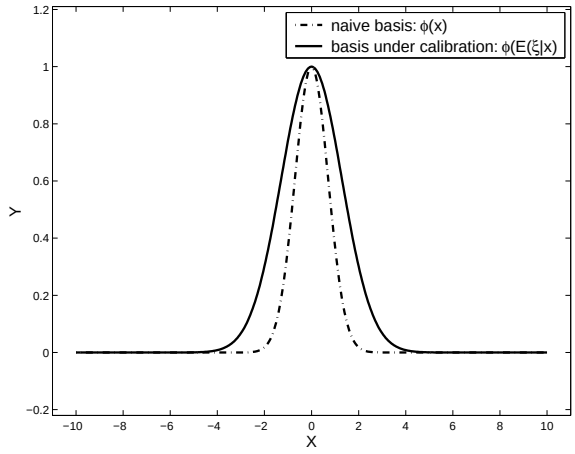

Figure 2.4: When covariate measurement error is introduced into the RVM regression model, naive estimation and regression calibration exclusively differ in the type of basis function they fit to the data.

study indicated that, for the RVM approach, standard regression calibration applying (2.61) is more or less equivalent to the naive analysis. So standard regression seems not to be a promising correction approach in a flexible regression setup. However, the principle of regression calibration can be carried forward to a more successful method which will be termed as 'basis function calibration' and will be presented in the next paragraph.

Basis function calibration

Carroll et al. (1999) present a structural approach to regression splines, which extends the idea of standard regression calibration to flexible regression. Rummel (2004) combines the core of this approach with the relevance vector machine and terms this 'basis function calibration'. Using the radial basis function type is the prominent new challenge, however, making this approach particularly profitable. The idea of basis function calibration is sketched in the following.

An intuitive improvement of standard regression calibration, which fits the working model (2.60) to the data lies in finding a better approximation of the observed mean $\mathbb{E}(Y|X)$, i.e. a more accurate working model under retainment of the parameters of the ideal mean model $\boldsymbol{\omega}$ (2.59). Indeed an exact model will be developed for the Gaussian RVM regression.
In the context of flexible regression the predictor f^* is formulated in terms of nonlinear basis functions $\phi_j(\xi)$ rather than in ξ itself. Consequently, it is natural to think of replacing each basis function $\phi_j(\xi)$ by $\mathbb{E}(\phi_j(\xi)|X)$, instead of replacing ξ by $\mathbb{E}(\xi|X)$.
The values of the basis functions at a certain observation ξ_i in the ideal model are organized in row vectors

$$\Phi(\xi_i) = [\phi_0, \phi_1(\xi_i), \ldots, \phi_J(\xi_i)].$$

Thus, it is convenient to define the calibrated version of this vector for each observation as

$$\mu_{\Phi(\xi)|x_i} \quad := \quad [\phi_0, \mathbb{E}(\phi_1(\xi)|x_i), \ldots, \mathbb{E}(\phi_J(\xi)|x_i)]. \tag{2.62}$$

This is still a row vector and will be used for the new approximate observed mean model. The basic idea is again, similar to the standard regression calibration, that under small measurement error the latent row vector $\Phi(\xi_i)$ will be close to $\mu_{\Phi(\xi)|x_i}$. However, even with small measurement error, $\Phi(\xi_i)$ may not be close to $\Phi(x_i)$. Thus naively replacing $\Phi(\xi_i)$ by $\Phi(x_i)$ may lead to large bias, hence the need for calibration.

Note that if $\phi_j(\cdot)$ is a linear basis $\mathbb{E}(\phi_j(\xi|X))$ simplifies to $\phi_j\left(\mathbb{E}(\xi|X)\right)$, which is equivalent to standard regression calibration as discussed in the previous paragraph. Choosing the typical radial basis function (RBF) type, the computation of the elements $\mathbb{E}(\phi_j(\xi)|X)$ in the calibrated vector $\mu_{\Phi(\xi)|X}$ is more complex. Chapter 3 presents details on the calculations for univariate radial basis functions under the structural assumption of $\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$ and $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$. A more flexible specification of ξ following a mixture of normals, as adopted by e.g. Davidian & Gallant (1993) and Carroll et al. (1999) leaves the core of the method unchanged and the required computations of the elements $\mathbb{E}(\phi_j(\xi)|X)$ still tractable. However, it requires the estimation of the associated parameters of this mixture distribution.
In the particular case of Gaussian regression, i.e. $G(z) = Id(z) = z$, the replacement of the row vector $\Phi(\xi)$ by $\mu_{\Phi(\xi)|X}$ (cf. (2.62)) yields an exact representation of $\mathbb{E}(Y|X)$ in terms of the parameter vector $\boldsymbol{\omega}$ of the ideal mean model, i.e.

$$\mathbb{E}(Y|X) \;=\; \mu_{\Phi(\xi)|X}\boldsymbol{\omega}. \tag{2.63}$$

Here, the parameters collected in $\boldsymbol{\omega}$ are viewed as non-random and thus conditioning on the left hand side is superfluous. In any other case than the Gaussian regression case, where $G \neq Id$, this is an approximate, so-called working model for the mean function of the observed data, i.e.

$$\mathbb{E}(Y|X) \;\approx\; G(\mu_{\Phi(\xi)|X}\boldsymbol{\omega}). \tag{2.64}$$

Inference for the RVM using basis function calibration proceeds then in a standard way as described in Section 2.1.2, but now, of course, with the new design matrix, defined as

$$\Phi_c = \begin{pmatrix} \mu_{\Phi(\xi)|x_1} \\ \mu_{\Phi(\xi)|x_2} \\ \cdots \\ \mu_{\Phi(\xi)|x_N} \end{pmatrix}$$

utilizing the calibrated row vectors from (2.62).

Rummel (2004) presents evidence of the improvement that can be achieved by using this method in the Gaussian case. In that case this method retains an analytical solution for the posterior mean estimator for the weights $\boldsymbol{\omega}$ (cf. 2.11), which is a neat property since a time consuming scoring algorithm can be prevented.
This working model (2.64) for basis function calibration will later also be used for binary and Poisson distributed responses treated in this work.

Viewing this measurement error problem in terms of the mean model is halfway to a more general class of correction methods that consider correction of the so-called 'score function' and thus addressing both, mean and variance function.

Structural quasi likelihood

Earlier in Section 2.1.2, the connection between finding the posterior mean of the weights in the RVM regression setup and solving a penalized score function has been established. As described by Carroll et al. (1995), chapter 7, the essence of likelihood based measurement error correction lies in finding the likelihood, conditional on the observed data

$$p_{Y|X}(y|x,\boldsymbol{\omega}) \;=\; \int p_{Y|\xi}(y|\xi,\boldsymbol{\omega})p_{\xi|X}(\xi|x)\mathrm{d}\xi, \tag{2.65}$$

where parameters, besides the fundamental mean model parameters $\boldsymbol{\omega}$, are suppressed here for the sake of clarity. Further parameters inherent in (2.65), include variance parameters in the likelihood function and parameters of the conditional distribution $p_{\xi|X}$, like e.g. the measurement error variance.
Maximizing the modified likelihood $p_{Y|X}(y|x,\boldsymbol{\omega})$ from (2.65) with respect to $\boldsymbol{\omega}$ yields error corrected parameter estimates for $\boldsymbol{\omega}$. However, the likelihood of the observed data $p_{Y|X}(y|x,\boldsymbol{\omega})$ does rarely represent a standard density and its specific form is usually not even computable due to the complexity of the integration.

An appealing facilitation of parameter estimation is the quasi likelihood

method introduced by Wedderburn (1974) and further developed by McCullagh (1983). In that approach, the unknown parameters are estimated from the mean and variance function only, rather than from the full likelihood. This is the basis for the structural quasi likelihood correction as presented by Carroll et al. (1995) and applied e.g. by Augustin (2002) in the context of survival analysis. Here, the error correction affects only the mean and the variance function of the model not the full likelihood.
The 'ideal mean model' is recalled here from (2.48) as

$$\mathbb{E}(Y|\xi) = f(\xi, \boldsymbol{\omega}) \tag{2.66}$$

and the 'ideal variance model' is defined as

$$\mathbb{V}(Y|\xi) = \sigma^2 g^2(\xi, \boldsymbol{\omega}, \theta), \tag{2.67}$$

based on the true but latent covariate ξ. As before the respective model parameters are assumed being non-random and thus do not appear in the conditional moments on the left hand sides.
Considering the RVM from a frequentist's view with fixed $\boldsymbol{\omega}$ its ideal mean function is represented by a sum of weighted basis functions, which is in matrix notation

$$\mathbb{E}(Y|\xi) = G(\Phi(\xi)\boldsymbol{\omega}),$$

while the ideal variance function is specified according to the type of response distribution and possibly accounting for over- and underdispersion and heteroscedasticity.
Now, given the specifications of the mean model $\mathbb{E}(Y|\xi)$ and the variance model $\mathbb{V}(Y|\xi)$ together with realizations $(y_i, \xi_i), i = 1, \ldots, N$, where it is assumed for a moment that the latent ξ_i could be observed, the quasi score function for the parameters $\boldsymbol{\omega}$ is computed as

$$s^{\xi}(Y, \xi, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\partial\, \mathbb{E}(y_i|\xi_i)}{\partial \boldsymbol{\omega}} \frac{y_i - \mathbb{E}(y_i|\xi_i)}{\mathbb{V}(y_i|\xi_i)}. \tag{2.68}$$

$\mathbb{E}(Y|\xi_i)$ and $\mathbb{V}(Y|\xi_i)$ account e.g. for the type of response distribution. Equating (2.68) to zero yields the parameter estimates $\widehat{\boldsymbol{\omega}}$. If the underlying model

is a generalized linear model in canonical form this coincides with the usual score function derived as first derivative of the log-likelihood with respect to the unknown parameters. Even with other models, under appropriate regularity conditions, the estimators obtained from (2.68) are still consistent, asymptotically normal and this estimation method is the most efficient one among those being linear in Y.
Estimation of the additional variance parameter σ^2 and the nuisance parameters θ in (2.67) can be achieved by solving

$$s^{\xi}(Y, \xi, \sigma^2, \theta) = \sum_{i=1}^{N} \frac{\partial \log(\mathbb{V}(y_i|\xi_i))}{\partial(\sigma^2, \theta^{\mathrm{T}})^{\mathrm{T}}} \left(\frac{(y_i - \mathbb{E}(y_i|\xi_i))^2}{\mathbb{V}(y_i|\xi_i)} - \frac{N - (J+1)}{N} \right) = 0, \tag{2.69}$$

where $J+1$ denotes here the number of current weights in the model. However, the variance parameter estimation for the RVM, where σ^2 is a random variable, will again be derived via an approximation to the marginal likelihood as before (cf. Section 2.1.2). Consequently, (2.69) will not be considered further.

An appropriate quasi score function for the relevance vector machine needs to account for the specified prior over the fundamental parameters $\boldsymbol{\omega}$. The adopted Gaussian prior (cf. 2.6) works as a quadratic penalty on the coefficients, and under this supplement the quasi score function from (2.68) is slightly modified to the penalized quasi score function

$$s^{\xi}(Y, \xi, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\partial \, \mathbb{E}(y_i|\xi_i)}{\partial \boldsymbol{\omega}} \frac{y_i - \mathbb{E}(y_i|\xi_i)}{\mathbb{V}(y_i|\xi_i, \boldsymbol{\omega}, \theta)} - \boldsymbol{\omega} A.$$

In the structural quasi likelihood approach error correction is realized by substituting $\mathbb{E}(Y|\xi)$ and $\mathbb{V}(Y|\xi)$ by their observed counterparts

$$\mathbb{E}(Y|X) = \int \mathbb{E}(Y|\xi) p_{\xi|X} \mathrm{d}\xi \tag{2.70}$$

$$\mathbb{V}(Y|X) = \int \mathbb{V}(Y|\xi) p_{\xi|X} \mathrm{d}\xi \tag{2.71}$$

under retainment of the fundamental mean and variance model parameters (cf. e.g. Carroll et al. (1995)). This yields the modified estimation equation

$$\sum_{i=1}^{N} \frac{\partial \mathbb{E}(y_i|x_i)}{\partial \boldsymbol{\omega}} \frac{y_i - \mathbb{E}(y_i|x_i)}{\mathbb{V}(y_i|x_i, \boldsymbol{\omega}, \theta)} - \boldsymbol{\omega} A = 0. \tag{2.72}$$

Like in other structural correction methods it is again indispensable to specify distributions over ξ and δ to compute the conditional distribution $p(\xi|X)$, which is required for calculating $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$.

Through the connection between solving a penalized quasi score function and Bayesian inference in the RVM as described earlier in Section 2.1.2, finding the root of the modified score function (2.72) yields a corrected estimate of the posterior mode of the distribution $p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2)$. The posterior covariance matrix is given by the inverse expected Fisher matrix, which is derived via the first derivative of the penalized score vector (2.72) with respect to $\boldsymbol{\omega}$. This estimator does, however, not take into account that the moments of $p(\xi|X)$ might have been estimated and not given a priori. Accounting for that uncertainty properly remains an open problem.

Theoretically, it is similarly possible to compute the corrected moments of the marginal likelihood and use these in (2.69) to estimate the variance parameters σ^2 and $\boldsymbol{\alpha}$ – at least if a uniform prior has been specified over these parameters. However, an approximation for the marginal likelihood will be sought and maximized as described in Section 2.1.2.

The analytic computation of the mean (2.70) and variance function (2.71) of the observed data is restricted to Gaussian responses. This approach itself and how it is embedded in the estimation of the hyperparameters is discussed in greater detail in chapter 3 of this thesis.

For all other response functions $G \neq Id$ this structural quasi likelihood approach is not passable since observed mean and variance functions can not be exactly related to their ideal counterparts. This means that the observed $\mathbb{E}(Y|X)$ and observed variance $\mathbb{V}(Y|X)$, both in terms of the parameter $\boldsymbol{\omega}$ of the ideal model, can not be computed offhand. This fact is particularly attributable to the expansion of the covariate in nonlinear basis functions. Numerical integration may be of some help at this point, which is, however,

not investigated during the course of this work.
In order to improve the analysis of non-Gaussian responses, i.e. when $G \neq Id$, the following related 'expanded basis function calibration' approach merely seeks to find good approximations to the observed models instead an exact representation .

Expanded basis function calibration

The structural quasi likelihood approach seeks an exact representation of the observed mean and variance models. In the case of non-Gaussian RVM regression the structural quasi likelihood is not feasible, and approximate models are attractive. The expanded basis function calibration introduced here is a derivative of the approximate quasi likelihood method presented by Carroll & Stefanski (1990), which was, however, exclusively used in non-flexible regression until now. This method was later also termed expanded regression calibration by Carroll et al. (1995). They suggest a method, utilizing standard regression calibration and Taylor series expansion, to improve the approximation of the observed moments. This thesis extends that approach to basis function calibration in the flexible RVM.
The original expanded regression calibration as described by Carroll et al. (1995) is based on standard regression calibration and is very briefly presented here. Though 'expanded regression calibration' itself will not be applied in this work, it may give insight into the basic idea behind the concept which is later applied to radial basis functions in the 'expanded basis function calibration'.

The ideal mean and variance models are recalled as

$$\begin{aligned} \mathbb{E}(Y|\xi) &= f(\xi, \boldsymbol{\omega}) \\ \mathbb{V}(Y|\xi) &= \sigma^2 g^2(\xi, \boldsymbol{\omega}, \theta), \end{aligned}$$

with dispersion parameter σ^2 and a function $g^2(\xi, \boldsymbol{\omega}, \theta)$ possibly accounting for heteroscedasticity.

The essential idea behind regression calibration as already described earlier is to replace ξ by $\mu_{\xi|X} = \mathbb{E}(\xi|X)$. This yields the following working model for the observed data (retaining the fundamental mean and variance model parameters)

$$\mathbb{E}(Y|X) \approx f(\mu_{\xi|X}, \boldsymbol{\omega}) \tag{2.73}$$

$$\mathbb{V}(Y|X) \approx \sigma^2 g(\mu_{\xi|X}, \boldsymbol{\omega}, \theta). \tag{2.74}$$

That is, where the 'expanded' comes in. The working models (2.73) and (2.74) are now refined by a second order Taylor series expansion of the true functions $f(\xi, \boldsymbol{\omega})$ and $\sigma^2 g(\xi, \boldsymbol{\omega}, \theta)$ around $\mu_{\xi|X}$.
In the expanded basis function calibration this Taylor series expansion will be around $\mu_{\Phi(\xi)|X}$, which is the calibrated row vector from (2.62) as defined earlier. Exact calculation of $\mu_{\Phi(\xi)|X}$ and $\Sigma_{\Phi(\xi)|X} = \mathbb{V}(\Phi(\xi)|X)$ will be required and will be shown to be feasible under the assumption of ξ and δ being Gaussian. This approach is again structural since the latent covariate ξ is assumed to follow a certain distribution.
Once the approximate observed mean and variance model are computed, parameter estimation is accomplished by finding the root of the penalized quasi score under usage of the approximations for $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$. This is in analogy to the structural quasi likelihood approach presented previously. Expanded basis function calibration will be used for the non-Gaussian cases, where no analytic solution for the observed mean and variance model is available. The details on expanded regression calibration are postponed to chapter 4 for binary RVM regression, and the amendments for the Poisson case are developed in chapter 5.

SIMEX

The idea of SIMEX (SIMulation EXtrapolation) was originally proposed by Cook & Stefanski (1994) and further developed by Carroll, Küchenhoff, Lombard & Stefanski (1996). Carroll et al. (1995) present a detailed description of SIMEX which has become a standard method in correction for covariate

measurement error in parametric models. An extension to flexible regression is given by Carroll et al. (1999) and applied to the RVM model for the first time in Rummel (2005).
SIMEX is exclusively applicable in the case of classical measurement error, i.e. independent additive error (in some scale, e.g. , log), as presented in the respective paragraph of Section 2.3.1. However, and most notably, SIMEX is a so-called functional correction method that is realizable without any distributional assumptions about the latent covariate ξ. In contrast to the calibration methods, the conditional distribution $p(\xi|X)$ is not required here. The core of SIMEX is studying the effect of measurement error on the observed mean $\mathbb{E}(Y|X)$ in a simulation study and afterwards extrapolating on the error-free case.

For the classical additive measurement error model, as described in (2.50), artificial random errors $\delta_i^* \sim \mathcal{N}\left(0, \frac{\sigma_{\delta^*}^2}{m_i}\right)$ are generated and added to the original covariate samples $x_i, i = 1, \ldots, N$, where the surrogate x_i may again stand for the mean of m_i replicate measurements. Then a standard RVM analysis is performed using these 'new' data containing the additional error. Repeating this scheme sufficiently often with varying error variances $\sigma_{\delta^*}^2 = c \cdot \sigma_\delta^2$ (in multiples of the original error variance) allows to study the effect of the (additional) measurement error on the estimated mean function $\widehat{f}(x)$. Figure 2.5 displays the typical attenuation of the prediction $\widehat{f}(\xi_k)$ at position ξ_k with increasing artificial error variance $\sigma_{\delta^*}^2$.
Finally one extrapolates on the case of zero measurement error. Different extrapolation schemes are conceivable, however, extrapolation itself remains a dubious task.
Again the true error variance σ_δ^2 has to be known or estimated from replication or validation data in order to know how far to extrapolate the scheme. But, no assumptions about the distribution of ξ have to be made and even the assumption of normally distributed measurement error is not critical in practise (cf. Carroll et al. (1995), chapter 4).
Beyond that, the SIMEX approach is readily applicable to any kind of response model i.e. response function $G(z)$. However for responses being re-

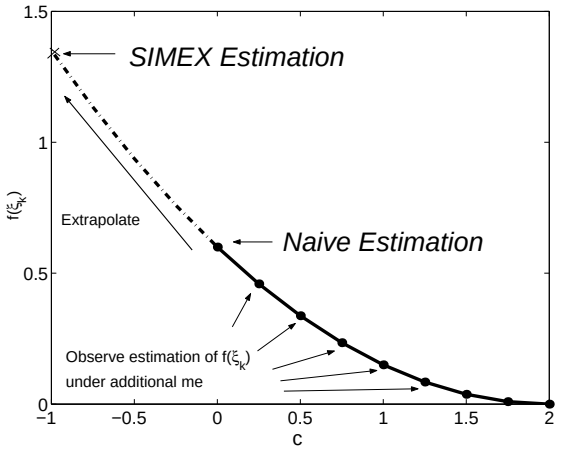

Figure 2.5: By inflating the variance for artificially generated error by $c \times \sigma_\delta^2$, the effect of additional error on the estimation $\widehat{f}(\xi_k)$ can be studied. Setting $c = 0$ is equivalent to using the original data in the analysis. The curve can be extrapolated to the case of zero measurement error for instance by using quadratic regression.

stricted to a certain domain, the SIMEX estimates might take on inadmissable values like $\widehat{f}_{SIMEX}(X) < 0$ or $\widehat{f}_{SIMEX}(X) > 1$ in a binary regression. An ad-hoc modification instead corrects the unlinked predictor $f^*(\mathbf{x})$, which is defined over the domain of $\mathbb{R}$ and finally links the corrected predictor to obtain the corrected estimation $G\left(\widehat{f}^*_{SIMEX}(\mathbf{x})\right)$. So very minor modification has to be done to suite the binary and Poisson regression cases in chapter 4 and 5.
But, since in fact one generates multiple artificial data sets and RVM estimates for each chosen error variance SIMEX becomes a computational heavy method.

Markov Chain Monte Carlo error correction

A radically different method to correct for measurement error goes back to Richardson & Gilks (1993a) and Richardson & Gilks (1993b). They follow a fully Bayesian approach of filling in the latent ξ_i's into the model by treat-

ing them as additional unknown parameters that have to be estimated. An overview of Bayesian error correction is presented by Richardson (1996) and the underlying idea is used e.g. by Gössl & Küchenhoff (2001) in a logistic regression problem with unknown change point and by Berry et al. (2002) in a flexible regression approach using P-splines.

This, typically MCMC based, measurement error correction exploits the strength of data augmentation (cf. Section 2.2.1) by additionally introducing the true but unobservable $\xi_i, i = 1, \ldots, N$ into the MCMC scheme as unknown parameters. The essential feature of this procedure is that, once these latent quantities are generated from their full conditional distribution, sampling from the other full conditionals remains basically unchanged from the case without measurement error. The correction comes here automatically from the Bayes machinery.
The sampling routine for the latent observations ξ_i can be modularly introduced into the MCMC-RVM from Section 2.2.2, essentially without affecting the other conditional densities in form. A remaining challenge is, however, to generate observations from the latent true covariate. The main features of Bayesian error correction in a flexible Bayesian probit model are presented in the following
The flexible Bayesian binary probit regression model from Section 2.2.2 enriched by basis selection (indicated by the subscript $\boldsymbol{\gamma}$) and with the latent variable ξ is recalled as

$$\begin{aligned} y_i &= \begin{cases} 1 & : \text{ if } z_i > 0 \\ 0 & : \text{ otherwise} \end{cases} \\ z_i &= \Phi_{\boldsymbol{\gamma}}(\xi_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}} + \epsilon_i, \\ \epsilon_i &\sim \mathcal{N}(\epsilon_i|0,1). \end{aligned}$$

For the parameter estimation, the samples of the weights, of the latent observations from both latent variables Z and ξ, and of the parameter $\boldsymbol{\gamma}$ indicating model complexity, have to be drawn from the joint posterior distribution $p(\boldsymbol{\omega}_{\boldsymbol{\gamma}}, \mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\gamma}|\mathbf{y})$. Though the joint posterior itself is not available, this is accomplished via the MCMC algorithm sampling in turn from the full con-

ditionals of the unknown parameters (cf. paragraph 'Sampling schemes' in Section 2.2.1). The modular design of the hitherto existing sampling scheme, without error correction (cf. Section 2.2.2) needs to be extended in order to account for measurement error, now.
Firstly, the covariate model, i.e. the prior distribution over the ξ_i's needs to be specified. In the simplest case, this is chosen to be a normal distribution

$$\xi_i \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$$

with respective normal and inverse Gamma hyperpriors over its moments:

$$\mu_\xi \sim \mathcal{N}(f, g^2), \tag{2.75}$$
$$\sigma_\xi^2 \sim IG(A_\xi, B_\xi). \tag{2.76}$$

The inverse Gamma is defined as

$$f(x|A,B) = \frac{1}{\Gamma(A)B^A x^{A+1}} \exp\left(-\frac{1}{Bx}\right) I(0 \leq x < \infty). \tag{2.77}$$

More flexible variants of specifying the distribution of ξ_i include the specification of a mixture of normals (cf. Roeder & Wasserman (1997)), which in turn increases the number of unknown parameters in the sampling scheme.
For convenience the covariate observations are assumed to come from the classical additive measurement error model (cf. 2.50) under the additional assumption of normally distributed measurement errors

$$x_{ij} = \xi_i + \delta_{ij}, \quad (\delta_{ij}, \xi_i) \sim \text{indep.}, \quad \delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2). \tag{2.78}$$

In the case of replication data, $j = 1, \ldots, m_i$ indexes these replications for person i. If the variance of the measurement error σ_δ^2 is unknown it can also be sampled in the algorithm if replication data is available; a typical prior over this variance is again the inverse Gamma distribution (2.77).
For the classical additive measurement error (2.50), as will be considered throughout this work, it holds that

$$\sigma_x^2 = \sigma_\xi^2 + \sigma_\delta^2.$$

Now, in order to account for this additional information in the sampling scheme, σ_δ^2 is reparametrized as $\sigma_\delta^2 = \frac{1-\lambda}{\lambda} \cdot \sigma_\xi^2$ with the so-called attenuation factor $\lambda = \frac{\sigma_\xi^2}{\sigma_X^2}$ (cf. the implementation of the methods of Berry et al. (2002) and also Carroll, Ruppert, Crainiceanu, Tosteson & Karagas (2004)). Thus, the respective distribution of the error-prone observations following model (2.78) becomes:

$$p(x_{ij}|\xi_i) \propto \exp\left(-\frac{1}{2\frac{1-\lambda}{\lambda} \cdot \sigma_\xi^2} \sum_{i=1}^{N} \sum_{j=1}^{m_i} (x_{ij} - \xi_i)^2\right)$$

Consequently, σ_δ^2 does not have to appear in the sampling scheme, but instead λ does. A uniform prior is specified for λ on the interval $[\lambda_L, \lambda_H]$.
The conditional distributions of the parameters $\boldsymbol{\omega}$, μ_ξ, σ_ξ^2, the model indicators $\boldsymbol{\gamma}$ and the vector of observations for the latent ξ and Z, i.e. $\boldsymbol{\xi}$ and $\mathbf{z}$ are constructed by collecting those terms from the joint posterior that contain the respective parameters of interest.
The joint posterior for $\boldsymbol{\theta} := (\boldsymbol{\omega}, \boldsymbol{\gamma}, \mathbf{z}, \boldsymbol{\xi}, \mu_\xi, \sigma_\xi^2, \lambda)$, under consideration of the prior distributions, as already specified in Section 2.2.3, is given by

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y}) \quad &\propto \quad p(\mathbf{y}|\mathbf{z}) \times \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(z_i - \Phi_{\boldsymbol{\gamma}}(\xi_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}})^2\right) \\
&\times \quad \binom{J+1}{dim(\boldsymbol{\gamma})}^{-1} \frac{1}{T+1} \exp\left(-\frac{1}{2\frac{1-\lambda}{\lambda} \cdot \sigma_\xi^2} \sum_{i=1}^{N} \sum_{j=1}^{m_i} (x_{ij} - \xi_i)^2\right) \\
&\times \quad \exp\left(-\frac{1}{2\sigma_\xi^2}\sum_{i=1}^{N}(\xi_i - \mu_\xi)^2\right) \exp\left(-\frac{1}{2g^2}(\mu_\xi - f)^2 - \frac{1}{B_\xi \sigma_\xi^2},\right) \\
&\times \quad \sigma_\xi^{-2(n+A_\xi+1)} \left(\frac{\lambda}{1-\lambda}\right)^{n/2} I(\lambda_L < \lambda < \lambda_H)
\end{aligned}$$

The model indicators $\boldsymbol{\gamma}$ are again sampled in a reversible jump MH step as presented in the respective Subsection in 2.2.3, based on the marginal likelihood. This is now conditioned on the latent ξ_i's, which means using the design matrix Φ constructed from the samples $\xi_i, i = 1, \ldots, N$.
The weights $\boldsymbol{\omega}$ and the latent variable vector $\mathbf{z}$, are sampled as above in

(2.44) and (2.45) and now also conditioned on the latent ξ_i, i.e. the design matrix based on the ξ_i's. So, in every step of the MCMC algorithm this design matrix has to be recalculated from the current samples.
The full conditional distributions of the parameters of the error model, μ_ξ and σ_ξ^2, are recognized as standard distributions

$$\begin{aligned} p(\mu_\xi|\boldsymbol{\xi}, \sigma_\xi^2) &= \mathcal{N}\left(\frac{\left(\sum_{i=1}^N \xi_i\right) g^2 + f\sigma_\xi^2}{Ng^2 + \sigma_\xi^2}, \frac{\sigma_\xi^2 g^2}{Ng^2 + \sigma_\xi^2}\right) \\ p(\sigma_\xi^2|\mathbf{x}, \boldsymbol{\xi}, \mu_\xi, \lambda) &= IG\left(A_{\xi|\cdot}, \frac{1}{B_{\xi|\cdot}}\right) \\ A_{\xi|\cdot} &= A_\xi + \frac{1}{2}\sum_{i=1}^N m_i + \frac{N}{2}, \\ B_{\xi|\cdot} &= B_\xi^{-1} + \frac{\lambda}{2(1-\lambda)}\sum_{i=1}^N \sum_{j=1}^{m_i}(x_{ij} - \xi_i)^2 + \frac{1}{2}\sum_{i=1}^N (\xi_i - \mu_\xi)^2, \end{aligned}$$

where A_ξ, B_ξ are from the prior specification (2.76) and $A_{\xi|\cdot}, B_{\xi|\cdot}$ denote the moments of the full conditional distribution.
Samples for the attenuation parameter λ relating σ_δ^2 to σ_ξ^2 are generated with the aid of a gridded Gibbs estimator that has also been used by Berry et al. (2002) and Carroll et al. (2004) in this situation. The full conditional of this parameter is given by

$$\begin{aligned} p(\lambda|\boldsymbol{\xi}, \sigma_\xi^2) &\propto I(\lambda_L < \lambda < \lambda_H)\left(\frac{\lambda}{1-\lambda}\right)^a \exp\left(-\frac{\lambda \cdot b}{2(1-\lambda)\sigma_\xi^2}\right) \\ a &= \sum_i m_i/2 \\ b &= \sum_i m_i(\overline{x}_i - \xi_i)^2 + \widehat{\sigma_\delta^2}\sum_i (m_i - 1), \end{aligned} \tag{2.79}$$

with $\overline{x}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}$ denoting the average over a subjects's replicates. Here, $\widehat{\sigma_\delta^2}$ is the methods of moments estimate for the measurement error variance (cf. (2.52)). The set $\lambda \in [\lambda_L, \lambda_H]$ can be discretized into a number of different values, then (2.79) is computed for these values, a discrete distribution function is constructed from the results and λ is sampled from this distribution

function. Alternatively, one can also implement a MH step based on (2.79). Finally, the conditional density of $\boldsymbol{\xi}$ is given as a product of the full conditionals of its elements ξ_i. These are given, up to a constant, by

$$\begin{aligned} p(\xi_i|z_i, \mathbf{x}_i, \mu_\xi, \sigma_\xi^2, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \lambda) &\propto \exp\left(-\frac{1}{2}(z_i - \Phi_{\boldsymbol{\gamma}}(\xi_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}})^2\right) \\ &\times \exp\left(-\frac{1}{2\frac{1-\lambda}{\lambda}\cdot\sigma_\xi^2}\sum_{j=1}^{m_i}(x_{ij} - \xi_i)^2\right) \\ &\times \exp\left(-\frac{1}{2\sigma_\xi^2}(\xi_i - \mu_\xi)^2\right), \end{aligned} \tag{2.80}$$

which is not recognized as standard density. The ξ_i's are independent a posteriori and $\mathbf{x}_i := (x_{i1}, \ldots, x_{im_i})$ denotes the vector of replicates for person i. A Metropolis Hastings (MH) step is needed to sample observations from (2.80). Choosing a symmetric random walk proposal leads to the acceptance probability of the proposed move being equal to

$$\alpha = \min\left\{1, \frac{p(\xi_i'|z_i, \mathbf{x}_i, \mu_\xi, \sigma_\xi^2, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \lambda)}{p(\xi_i|z_i, \mathbf{x}_i, \mu_\xi, \sigma_\xi^2, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \lambda)}\right\}, \tag{2.81}$$

with the prime symbol indicating the proposal. The random walk proposal cancels out in (2.81), since it is symmetric. It is sufficient to be able to evaluate the conditional density (2.80) at both positions ξ_i and ξ_i' to perform MH sampling.

It may be stressed here again that, given the current values for $\boldsymbol{\xi}$, μ_ξ, σ_ξ^2, and λ, the remaining unknowns ($\boldsymbol{\omega}$, $\mathbf{z}$ and $\boldsymbol{\gamma}$) are still sampled from Gibbs sampling and MH sampling (cf. Section 2.2.3) – but now with the design matrix Φ based on the samples $\xi_i, i = 1, \ldots, N$ instead based on the error-prone observations $x_i, i = 1, \ldots, N$.

2.3.3 A failure - Corrected score

As stated earlier in Section 2.1.2, finding the posterior moments of the weights is equivalent to solving a penalized score function and calculating the in-

verse expected Fisher information matrix, respectively. The following small paragraph checks the possibility of applying the corrected score method (cf. Stefanski (1989)) to flexible regression using radial basis functions. For the sake of clarity penalization is not considered here.
The first derivative of the log likelihood yields the score function

$$s^{\xi}(Y, \xi, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\partial \, \mathbb{E}(y_i|\xi_i)}{\partial \boldsymbol{\omega}} \frac{y_i - \mathbb{E}(y_i|\xi_i)}{\mathbb{V}(y_i|\xi_i)}.$$

Of particular interest for parameter estimation are unbiased estimation functions having $\mathbb{E}(s^{\xi}(Y, \xi, \boldsymbol{\omega})) = 0$.
Let $s^{\xi}(Y, \xi, \boldsymbol{\omega})$ denote the score function given the observations y_i and based on the latent ξ_i and let $s^{\xi}(Y, X, \boldsymbol{\omega})$ denote the naive score function, which arises from $s^{\xi}(Y, \xi, \boldsymbol{\omega})$ after replacing ξ by X.
In general $\mathbb{E}(s^{\xi}(Y, X, \boldsymbol{\omega})) \neq 0$ and thus the root of $s^{\xi}(Y, X, \boldsymbol{\omega})$ is an inconsistent estimator of the true weights $\boldsymbol{\omega}$. This is an functional approach, which, in contrast to the structural approaches, does not use any information about the distribution of the latent covariate ξ and the error distribution and thus can be performed with high generality.

The idea of corrected score, as described by Stefanski (1989), is to search for a function $s^{X}(Y, X, \boldsymbol{\omega})$, with the property $\mathbb{E}(s^{X}(Y, X, \boldsymbol{\omega})|\xi) = s^{\xi}(Y, \xi, \boldsymbol{\omega})$. Every such function is called a corrected score function, since, by the law of iterated expectation it yields expectation being zero.
Therefore, in case of the classical additive error (2.50), $'\xi + \delta'$ $(=X)$ is plugged into the ideal score function (here for the RVM with $\mathbb{E}(y_i|\xi_i + \delta_i) = G(\Phi(\xi_i + \delta_i)\boldsymbol{\omega})$)

$$s^{\xi}(Y, \xi_i + \delta_i, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\delta G(\Phi(\xi_i + \delta_i)\boldsymbol{\omega})}{\delta \boldsymbol{\omega}} \frac{y_i - G(\Phi(\xi + \delta)\boldsymbol{\omega})}{\mathbb{V}(y_i|\xi_i + \delta)}$$

and one has to compensate for the effects of δ, which have entered after the replacement, such that $\mathbb{E}(s^{X}(Y, X, \boldsymbol{\omega})|\xi) = s^{\xi}(Y, \xi, \boldsymbol{\omega})$.
Since in the RVM the covariate is encoded in a set of radial basis functions and $\mathbb{E}(\Phi(\xi_i + \delta_i))$ is not decomposable into $\Phi(\xi_i)$ and a term being independent

of ξ, there seems to be no chance for compensation for the effects of the measurement error on the score function in this setting.

Chapter 3

Covariate measurement error in flexible Gaussian regression

This chapter is concerned with covariate measurement error in the flexible Gaussian regression model. Since the main techniques, including the response model, error model and correction methods, have already been presented in chapter 2, the task is merely putting the pieces together. Problem orientated details of the methods are, of course, stated in more detail in this chapter. While the Gaussian response case is trivial in standard statistical analysis, appropriate covariate measurement error correction in flexible Gaussian regression models is still of major interest.
Firstly, the Gaussian relevance vector machine (RVM) is recalled from chapter 2 and then Section 3.1 develops the error correction methods for this specific model. Relevant aspects of the methods and literature head the respective subsections. A simulation concludes this chapter.

The Gaussian RVM regression model is of the form

$$Y = \Phi(\xi)\boldsymbol{\omega} + \epsilon, \tag{3.1}$$

where the error is assumed to be $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Furthermore, a prior distribution over the parameters of the mean model is specified

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{j=0}^{J} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{\alpha_j}{2}\omega_j^2\right).$$

Gamma hyperpriors are specified over the variance σ^2 and those hyperparameters collected in $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_J)^{\mathrm{T}}$ (cf. Section 2.1.1).
Thus the 'ideal mean model' and the 'ideal variance model', motivated from (2.66) and (2.67) in Section 2.3, are here

$$\mathbb{E}(Y|\xi) = \Phi(\xi)\boldsymbol{\omega} \tag{3.2}$$
$$\mathbb{V}(Y|\xi) = \sigma^2. \tag{3.3}$$

More generally the 'ideal variance model' can be formulated as

$$\mathbb{V}(Y|\xi) = \sigma^2 g^2(\xi, \boldsymbol{\omega}, \theta), \tag{3.4}$$

with dispersion parameter σ^2, nuisance parameters collected in θ and $g^2(\cdot)$ possibly accounting for heteroscedasticity. This is however not considered in this work. The mean model parameters and the dispersion parameter, though chosen to be random parameters in a Bayesian context, are suppressed in the conditional mean and variance on the left hand side in (3.2) and (3.4) for notational clarity and will be throughout this chapter.

3.1 The arsenal of correction methods

Due to the simple model structure (3.1), a variety of promising correction methods is available. The main points and relations between the methods have already been roughly discussed in Section 2.3.2 and here follow the details.
Basis calibration is an ad-hoc idea in the spirit of standard regression involving the reformulation of the observed mean model only. However, most

conveniently it guarantees an analytic solution of the posterior mean over the fundamental model parameters. This method only improves estimation of the the fundamental model parameters of the mean model, i.e. those parameters collected in the vector $\boldsymbol{\omega}$.
The structural quasi likelihood method, however, obtains an approximately unbiased estimation equation for the posterior mean. This requires reformulation of both the observed mean and variance model. An important point to justify the adoption of these two methods lies in the equivalence of parameter estimation via the posterior mode estimation as applied in the RVM and solving a penalized quasi score function (cf. Section 2.1.2).
Finally, the simulation based SIMulation EXtrapolation (SIMEX) method is discussed, which in contrast to the former approaches makes no assumption about the distribution of the latent variable.
All methods are compared to the naive RVM and a competing state-of-the-art strategy using Markov Chain Monte Carlo (MCMC) techniques to estimate the parameters of a P-spline model.

3.1.1 Basis function calibration

The here developed basis function calibration can be seen as a generalization of the well known standard regression calibration (cf. Carroll & Stefanski (1990) and Gleser (1990)). To the author's knowledge, this generalization has only been used in the context of regression splines (cf. Carroll et al. (1999)). Rummel (2004), for the first time, shows that this approach does most successfully work in flexible regression using radial basis functions. The core idea is briefly motivated from the perspective of the standard regression calibration.

The standard regression calibration seeks an approximate model for the observed data in terms of the fundamental model parameters by replacing ξ in the ideal model (3.2) by $\mu_{\xi|X} = \mathbb{E}(\xi|X)$. For the Gaussian RVM this yields

the following approximation to the observed mean function

$$\mathbb{E}(Y|X) \approx \sum_{j=1}^{J} \omega_j \phi_j(\mu_{\xi|X}) + \omega_0 = \Phi(\mu_{\xi|X})\boldsymbol{\omega}. \tag{3.5}$$

In the RVM model (3.1), the linear predictor is formulated in terms of the radial basis functions $\phi(\xi)$ rather than ξ. Consequently, it is more natural to think about replacing all row vectors in the design matrix

$$\Phi(\xi_i) = [\phi_0, \phi_1(\xi_i), \ldots, \phi_J(\xi_i)]$$

by the calibrated row vectors, defined as

$$\mu_{\Phi(\xi)|x_i} := [\phi_0, \mathbb{E}(\phi_1(\xi)|x_i), \ldots, \mathbb{E}(\phi_J(\xi)|x_i)]. \tag{3.6}$$

Though the RVM uses radial basis functions, the required computation of $\mathbb{E}(\phi_j(\xi)|X)$ is analytically tractable in particular cases, depending on the specification of the conditional distribution $p(\xi|X)$.
Under the structural assumptions $\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$ and $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ and an additive relation between ξ and X it follows

$$p(\xi|X) = \mathcal{N}(\mu_{\xi|X}, \sigma_{\xi|X}^2). \tag{3.7}$$

The specific figures of the moments $\mu_{\xi|X}$ and $\sigma_{\xi|X}$ may account for replicate measurements, potential heteroscedasticity introduced by the error process and depend on the kind of error model - classical additive error or Berkson error, see also the respective paragraphs in Section 2.3.1.
It is important to note that the computation of the moments in (3.7) involves the measurement error variance σ_δ^2. This has to be known or estimated from replication or validation data. The latter will be done in the simulation study in Section 3.2, where two replicates will be available. The distribution $p(\xi|X)$ being Gaussian is a special case, however a popular one. Since it will be exclusively used in this work, the robustness of this method under this assumption will be tested.
Now, given $p(\xi|X)$ being Gaussian, the desired quantity $\mathbb{E}(\phi_j(\xi)|X)$ in (3.6)

can be re-written as the following integral

$$\begin{aligned}\mathbb{E}(\phi_j(\xi)|X) &= \int_{-\infty}^{\infty} \exp(-\eta(\xi - c_j)^2) p(\xi|X) \mathrm{d}\xi \\ &= \sqrt{2\pi} \int_{-\infty}^{\infty} \varphi\left(\sqrt{2\eta}(\xi - c_j)\right) \frac{1}{\sigma_{\xi|X}} \varphi\left(\frac{\xi - \mu_{\xi|X}}{\sigma_{\xi|X}}\right) \mathrm{d}\xi,\end{aligned}$$

where $\varphi(\cdot)$ denotes the standard Gaussian density.
Then, after substituting $t := \sqrt{2\eta}(\xi - c_j)$ and rearranging this is

$$\mathbb{E}(\phi_j(\xi)|X) = \frac{\sqrt{\pi}}{\sqrt{\eta}\sigma_{\xi|X}} \int_{-\infty}^{\infty} \varphi(t) \varphi\left(\frac{t}{\sqrt{2\eta}\sigma_{\xi|X}} + \frac{c_j - \mu_{\xi|X}}{\sigma_{\xi|X}}\right) \mathrm{d}t.$$

Here, the knots c_j are required to be located at fixed positions, not subject to any form of randomness as in the original RVM setup, see Tipping (2000). Integration of the product of two Gaussians is feasible (cf. e.g. Appendix of Küchenhoff (1995)) and the calibrated basis function can be written as

$$\mathbb{E}(\phi_j(\xi)|X) = \frac{\sqrt{\pi}}{\sqrt{\eta}\sigma_{\xi|X}} \varphi\left(\frac{b}{\sqrt{1 + c^2}}\right) \frac{1}{\sqrt{1 + c^2}},$$

where

$$\begin{aligned} b &= \left(\frac{c_j - \mu_{\xi|X}}{\sigma_{\xi|X}}\right) \\ c &= \frac{1}{\sqrt{2\eta}\sigma_{\xi|X}}. \end{aligned}$$

Inserting b and c leads to the representation

$$\mathbb{E}(\phi_j(\xi)|X) = \frac{1}{\sqrt{2\eta\sigma_{\xi|X}^2 + 1}} \exp\left(-\frac{\eta}{2\eta\sigma_{\xi|X}^2 + 1}(c_j - \mu_{\xi|X})^2\right). \quad (3.8)$$

This shows that calibration of the basis is threefold: firstly, replacing X by $\mu_{\xi|X}$, secondly, applying a wider width parameter (i.e. a smaller η) and finally, re-scaling the basis function. If $\sigma_{\xi|X}^2 = 0$ the calibrated basis function (3.8) is equivalent to the original radial basis function.
Here it may be helpful to revive the connection between error correction

and basis modification, established in Section 2.3.2: basis calibration like regression calibration exclusively differs from the naive analysis by adopting a modified basis function type for inference. According to this interpretation, Figure 3.1 visualizes this threefold modification of the original radial basis function leading to (3.8). Intuitively, the use of the new basis functions

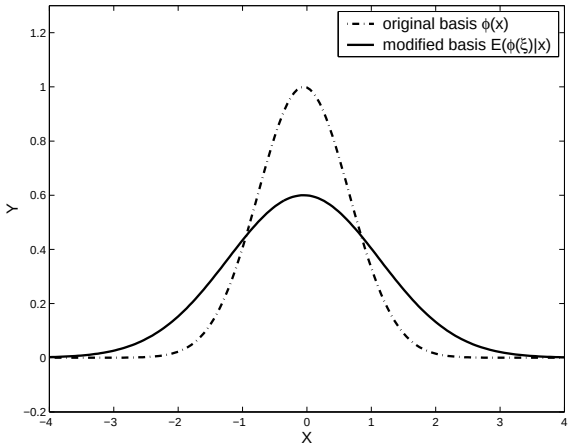

Figure 3.1: The RVM under basis calibration fits a modified basis functions to the observed data. Compared to the original radial basis type, this new basis is more flattened and wide.

leads to larger estimates for the coefficients. In this way, one alleviates the oversmoothing that is typically inherent in the naive analysis.
The Gaussian regression represents a special case, where the replacement of the latent row vector $\Phi(\xi)$ by $\mu_{\Phi(\xi)|X}$ yields an exact representation of the observed mean model $\mathbb{E}(Y|X)$ in terms of the ideal mean model parameters $\boldsymbol{\omega}$. To make this clear, the law of iterated expectations is used to rewrite $\mathbb{E}(Y|X)$ in terms of the parameters of the ideal mean model

$$\begin{aligned}
\mathbb{E}(Y|X) &= \mathbb{E}(\mathbb{E}(Y|X,\xi)|X) \\
&= \mathbb{E}(\mathbb{E}(Y|\xi)|X) \\
&= \mathbb{E}\left(\left(\textstyle\sum_{j=1}^{J}\omega_j\phi_j(\xi)+\omega_0\right)|X\right) \\
&= \textstyle\sum_{j=1}^{J}\omega_j\,\mathbb{E}(\phi_j(\xi)|X)+\omega_0 \\
&= \mu_{\Phi(\xi)|X}\boldsymbol{\omega}. \qquad (3.9)
\end{aligned}$$

As usual in this context, it is assumed that the measurement error is independent of the response variable. This property is called non-differentiability of the measurement error and means thats there is no additional information in the error about the response. Thus, conditioning on X is dispensable in the second line of the above re-formulation (3.9). The fundamental mean model parameters ω_j from (3.2) are retained under the modification of introducing the calibrated basis functions $\mathbb{E}\left(\phi_j(\xi)|X\right)$. This is in contrast to the standard regression calibration in (3.5), which is a working model in a literal sense.

The estimation of the parameters $\boldsymbol{\omega}$ again proceeds via the posterior distribution (2.11) (cf. Section 2.1.2). And the estimation of $\boldsymbol{\alpha}$ and σ^2 $(=\beta^{-1})$ is again performed by optimizing the marginal likelihood (2.10) (cf. Section 2.1.2).
However, the posterior distribution of the weights and the marginal likelihood are now based on the calibrated design matrix

$$\Phi_c = \begin{pmatrix} \mu_{\Phi(\xi)|x_1} \\ \mu_{\Phi(\xi)|x_2} \\ \cdots \\ \mu_{\Phi(\xi)|x_N} \end{pmatrix}, \tag{3.10}$$

containing the row vectors from (3.6).
It is important to note, that the standard errors based on the resulting posterior covariance matrix of $\boldsymbol{\omega}$ (2.11) do not properly account for the inherent covariate measurement error, which is a open problem, yet.
This method will also later be used for binary and Poisson regression, though it merely yields an approximate mean model for the observed data.

From a computational point of view it is important to note that computation of the calibrated design matrix is a preprocessing step that needs to be accomplished only once, before the parameter estimation procedure. Furthermore this method retains an analytical solution of the posterior moments of the weights in the Gaussian case. This is a nice property since time consuming

scoring algorithms can be prevented, which, however, come into play when additionally considering the observed variance as in the following approach.

3.1.2 Structural quasi likelihood

The connection of estimating the moments of the posterior distribution of the mean model parameters $\boldsymbol{\omega}$ and finding the root of a penalized score function has been described in Section 2.1.2. Through this link it is advantageous to cast the measurement error problem more generally in the form of so-called mean and variance models (cf. Carroll & Ruppert (1988)). There, a corrected parameter estimation is based upon a score function employing the observed mean and variance functions $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$. Carroll et al. (1995) term these methods later 'quasi likelihood and variance function methods'.
The RVM model setup remains here truly Bayesian, adopting hyperparameters and seeking to find the posterior mean of the unknowns. However, it will prove fruitful here to view the measurement error problem as occurring in a penalized likelihood setting, where the methods for mean and variance models can be applied to achieve a corrected estimator. Here, the art is Bayesian but the instruments are frequentistic!

Given the specifications of the ideal mean model (3.2) and the ideal variance model (3.4), together with realizations $(y_i, \xi_i), i = 1, \ldots, N$, the penalized quasi score function for the parameters $\boldsymbol{\omega}$ in the RVM regression model is given by

$$s^{\xi}(Y, \xi, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\partial\, \mathbb{E}(y_i|\xi_i)}{\partial \boldsymbol{\omega}} \frac{y_i - \mathbb{E}(y_i|\xi_i)}{\mathbb{V}(y_i|\xi_i)} - \boldsymbol{\omega} A. \tag{3.11}$$

Here, A again denotes the diagonal matrix having the vector $\boldsymbol{\alpha}$ as its diagonal. Then, if the ξ_i's were available, equating (3.11) to zero yields the parameter estimate $\widehat{\boldsymbol{\omega}}$. This penalized quasi score function (3.11) coincides with the usual penalized score function (without 'quasi') derived as first derivative of the log-likelihood plus the log-prior with respect to $\boldsymbol{\omega}$.

The baseline of the structural quasi likelihood approach is to substitute $\mathbb{E}(Y|\xi)$ and $\mathbb{V}(Y|\xi)$ in (3.11) by their observed counterparts under retainment of the fundamental mean and variance model parameters. In case of an un-penalized score function, this procedure yields an approximately unbiased estimation function for $\boldsymbol{\omega}$. Unbiasedness in penalized settings, is discussed in Section 2.1.2 and has a slightly different meaning. Despite it has been revealed in Section 2.1.2 that the RVM eo ipso is not unbiased, a good deal of error correction is expected to come from the application of that approach. Therefore, the relation between the observed conditional moments and the ideal conditional moments is required, which retains both parameters, $\boldsymbol{\omega}$ and σ^2, of the ideal mean and variance model (cf. (3.2), (3.4)).
Assuming non-differentiability of the measurement error and applying the theorem of iterated expectations, the mean and variance functions of the observed data are given as

$$\begin{aligned}\mathbb{E}(Y|X) &= \mathbb{E}(\mathbb{E}(Y|\xi, X)|X)\\ \mathbb{V}(Y|X) &= \sigma^2 + \mathbb{V}(\mathbb{E}(Y|\xi, X)|X).\end{aligned}$$

In the specific Gaussian RVM case, where $\Phi(\xi)$ denotes the row vector of all $J+1$ basis functions (including the intercept) at position ξ, this is

$$\mathbb{E}(Y|X) = \mathbb{E}(\Phi(\xi)|X)\boldsymbol{\omega} = \mu_{\Phi(\xi)|X}\boldsymbol{\omega} \tag{3.12}$$

$$\mathbb{V}(Y|X) = \sigma^2 + \mathbb{V}(\Phi(\xi)\boldsymbol{\omega}|X). \tag{3.13}$$

For notational clarity, the row vector $\mu_{\Phi(\xi)|X} := \mathbb{E}(\Phi(\xi)|X)$ is introduced. This quantity is recognized from the basis function calibration approach (cf. (3.6)). The required assumptions allowing for the computability of $\mu_{\Phi(\xi)|X}$ are echoed later this paragraph. Thus, only the observed variance (3.13) is yet unknown and of further interest.
Applying the variance decomposition formula to (3.13), this can equivalently be written as

$$\mathbb{V}(Y|X) = \sigma^2 + \boldsymbol{\omega}^{\mathrm{T}}\,\mathbb{E}(\Phi(\xi)^{\mathrm{T}}\Phi(\xi)|X)\boldsymbol{\omega} - \boldsymbol{\omega}^{\mathrm{T}}\mu_{\Phi(\xi)|X}^{\mathrm{T}}\mu_{\Phi(\xi)|X}\boldsymbol{\omega}. \tag{3.14}$$

It is stressed again, that the weights are assumed to be fixed in the expansion of the observed variance since it has been implicitly conditioned on the

weights, which is however suppressed for notational clarity.
The second summand in (3.14) reflects integration of elements in a matrix constructed by the vector product $\Phi(\xi)^{\mathrm{T}}\Phi(\xi)$. The third summand represents a vector product of the calibrated vectors $\mu_{\Phi(\xi)|X}$. Interchanging summation and integration and expanding (3.14) into a elementwise sum allows for a more convenient reformulation

$$\begin{aligned}\mathbb{V}(Y|X) &= \sigma^2 + \sum_{k=0}^{J}\sum_{j=0}^{J}\left(\mathbb{E}(\phi_j(\xi)\phi_k(\xi))|X) - \mu_{\phi_j(\xi)|X}\mu_{\phi_k(\xi)|X}\right)\omega_j\omega_k \\ &= \sigma^2 + \boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\Phi(\xi)|X}\boldsymbol{\omega}, \end{aligned} \tag{3.15}$$

where $\mu_{\phi_j(\xi)|X} := \mathbb{E}(\phi_j(\xi))|X)$ denotes the $(j+1)$th element of the row vector $\mu_{\Phi(\xi)|X}$ and $\Sigma_{\Phi(\xi)|X} := \mathbb{V}(\Phi(\xi)|X)$ denotes the covariance matrix of the $J+1$ latent basis functions given X. The matrix $\Sigma_{\Phi(\xi)|X}$ describes the dependence structure of the true, but latent, basis functions given the observed data. Since the 'first' basis function ϕ_0, the intercept, is independent of X, the first column and row of $\Sigma_{\Phi(\xi)|X}$ have zero entries.
For heteroscedastic response error, i.e $g(\xi, \boldsymbol{\omega}, \theta) \neq 1$, the observed variance would also include an integral over $g^2(\xi, \boldsymbol{\omega}, \theta)$

$$\mathbb{V}(Y|X) = \sigma^2 \int g^2(\xi, \boldsymbol{\omega}, \boldsymbol{\theta}) p_{\xi|X} \mathrm{d}\xi + \boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\Phi(\xi)|X}\boldsymbol{\omega}, \tag{3.16}$$

which can be a complex task in some special cases, but usually is easy to calculate. However, only the case of homoscedastic response errors is considered throughout this thesis.

Having knowledge about the conditional distribution $p_{\xi|X}$ is again essential in order to perform the necessary integration inherent in $\Sigma_{\Phi(\xi)|X}$ in (3.15). Therefore, in the classical error model, distributions over the latent variable and the measurement error must be specified. In the Berkson case only the distribution over the error is needed to be able to compute $p_{\xi|X}$.
Alternatively, one can directly specify $p(\xi|X)$ in a flexible way, e.g. by a mixture of normals or mixture distributions as proposed by Davidian & Gallant (1993). Carroll et al. (1999) assume a mixture of normals to appropriately

represent $p(\xi)$ in a classical (Gaussian) error setup. They perform MCMC sampling to determine posterior probabilities for the number of mixture components and the estimates for the associated moments. This information can be used to determine the density $p(\xi|X)$ which is, again a mixture of normals and can be used for the correction of the mean and variance function, respectively.

The required integrations in order to compute $\mu_{\Phi(\xi)|X}$ and $\Sigma_{\Phi(\xi)|X}$ in the observed models (cf. (3.12) and (3.15)) are analytically tractable in particular cases, depending on $p(\xi|X)$.
As described in Section 2.3.1, the following assumptions are made in this work: $\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$, $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ and the measurement error is additive. Then, it follows that the required conditional distribution is of the from

$$p(\xi|X) = \mathcal{N}(\mu_{\xi|X}, \sigma_{\xi|X}^2).$$

The specific figures of the moments $\mu_{\xi|X}$ and $\sigma_{\xi|X}$ may again account for replicate measurements, potential heteroscedasticity introduced by the measurement process and depend on the kind of error model - classical additive error or Berkson error, cf. Section 2.3.1. In any case its computation involves the measurement error variance that has to be known or estimated from replication or validation data. The robustness of this method under the normality assumption for the conditional distribution will be investigated later in the simulations.
The mean function (3.12) is formulated in terms of the calibrated basis functions as derived earlier in the basis function calibration (3.8).
However, the exact representation of the observed variance function (3.15) involves computation of the elements in $\Sigma_{\Phi(\xi)|X}$ which is rather complex. The non-zero elements of $\Sigma_{\Phi(\xi)|X}$ corresponding to the J non-intercept basis functions are given as

$$Cov(\phi_j(\xi), \phi_k(\xi)|X) \;=\; \mathbb{E}(\phi_j(\xi), \phi_k(\xi)|X) - \mathbb{E}(\phi_j(\xi)|X)\,\mathbb{E}(\phi_k(\xi)|X). \tag{3.17}$$

For univariate radial basis functions and under the assumption of $p(\xi|X)$ being a Gaussian distribution, the first term in (3.17) can be rewritten as

$$\begin{aligned}\mathbb{E}(\phi_j(\xi),\phi_k(\xi)|X) &= \int_{-\infty}^{\infty}\exp(-\eta(\xi-c_j)^2)\exp(-\eta(\xi-c_k)^2)p(\xi|X)\mathrm{d}\xi \\ &= \frac{\sqrt{2}\pi}{\sqrt{\eta}\sigma_{\xi|X}}\int_{-\infty}^{\infty}\varphi(t)\varphi\left(\frac{t}{\sqrt{2\eta}\sigma_{\xi|X}}+\frac{c_j-\mu_{\xi|x}}{\sigma_{\xi|X}}\right) \\ &\quad \varphi\left(t+\sqrt{2\eta}(c_j-c_k)\right)\mathrm{d}t,\end{aligned}$$

where $\varphi(\cdot)$ denotes the standard Gaussian density and the substitution $t := \sqrt{2\eta}(\xi - c_j)$ has been performed. Rewriting the radial basis functions in terms of normal distributions proceeds as described for the calculation of the calibrated basis functions in Section 3.1.1. The knots c_j and c_k are again required to be located at fixed positions, not subject to any form of randomness. Integrating a product of three Gaussians is feasible (cf. e.g. Appendix of Küchenhoff (1995)) and given by

$$\mathbb{E}(\phi_j(\xi),\phi_k(\xi)|X) = \frac{\sqrt{2}\pi}{\sqrt{\eta}\sigma_{\xi|X}}\varphi\left(\frac{b}{\sqrt{1+c^2}}\right)\frac{1}{\sqrt{1+c^2}}\varphi\left(\frac{d}{e}\right), \quad (3.18)$$

where

$$\begin{aligned} b &= \left(\frac{2\eta\sigma_{\xi|X}^2(c_j-c_k)+c_j-\mu_{\xi|x}}{\sqrt{\sigma_{\xi|X}^2+2\eta\sigma_{\xi|X}^4}}\right) \\ c &= \sqrt{1+\frac{1}{2\eta\sigma_{\xi|X}^2}} \\ d &= \sqrt{2\eta}(\mu_{\xi|X}-c_k) \\ e &= \sqrt{1+2\eta\sigma_{\xi|X}^2}. \end{aligned}$$

Combining the recent result (3.18) with the formula for the calibrated basis functions (3.8) allows for calculation of the elements of the desired covariance matrix $\Sigma_{\Phi(\xi)|X}$ in (3.15). However, since this matrix depends on the specific realizations, an individual matrix $\Sigma_{\Phi(\xi)|X=x_i}$ for every observation $x_i, i = 1,\ldots,N$ must be calculated.

It is important to stress, that computation of these N covariance matrices is a preprocessing step that is performed only once prior to the optimization algorithm of the RVM. But in the light of typically having a huge set of potential basis functions, the calculation of $J(J+1/2)$ distinct elements in each of N distinct covariance matrices might be expensive.
Alternatively, and more efficiently, one can focus on calculating the required covariances only for those basis functions present in the current model. As discussed in Section 2.1.2 the algorithm starts with having only a single basis function in the model and decides from step to step whether to include or to exclude basis function. Thus, the matrices $\Sigma_{\Phi(\xi)|x_i}$ need only to be updated if a new basis is introduced.

The observed mean $\mathbb{E}(Y|X)$ from (3.12) and the observed variance $\mathbb{V}(Y|X)$ from (3.15) are plugged into the penalized score (3.11) to give

$$s^X(Y, X, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\partial\, \mathbb{E}(y_i|x_i)}{\partial \boldsymbol{\omega}} \frac{y_i - \mathbb{E}(y_i|x_i)}{\mathbb{V}(y_i|x_i)} - \boldsymbol{\omega} A. \tag{3.19}$$

The posterior mean of the parameters $\boldsymbol{\omega}$ is estimated by finding the root of this penalized score function.
Since the observed variance function (3.15) is formulated in terms of the unknown parameters $\boldsymbol{\omega}$, the modified score function (3.19) can no longer be solved analytically in contrast to the basis function calibration.
Instead, the Fisher scoring algorithm utilizing the modified score function (3.19) and the expected Fisher matrix will be applied, as is also used in the non-Gaussian regression case in Section 2.1.2.
Calculation of the expected Fisher matrix includes differentiation of the penalized score with respect to $\boldsymbol{\omega}$ and is given by

$$F(\boldsymbol{\omega}) \;=\; \left(\Phi_c^{\mathrm{T}} B_c \Phi_c + A\right), \tag{3.20}$$

where for the sake of clarity the calibrated design matrix Φ_c, given the observations $x_1, \ldots, x_N$, and the diagonal matrix B_c are defined as

$$\Phi_c = \begin{pmatrix} \mu_{\Phi(\xi)|x_1} \\ \mu_{\Phi(\xi)|x_2} \\ \cdots \\ \mu_{\Phi(\xi)|x_N} \end{pmatrix}, B_c = \begin{pmatrix} \mathbb{V}(y_1|x_1)^{-1} & & \\ & \ddots & \\ & & \mathbb{V}(y_N|x_N)^{-1} \end{pmatrix}. \tag{3.21}$$

Here, $\mathbb{V}(y_i|x_i) = \sigma^2 + \boldsymbol{\omega}^{\mathrm{T}} \Sigma_{\Phi(\xi)|x_i} \boldsymbol{\omega}$ as stated in (3.15). So, B_c is the inverse covariance matrix of the responses given the observations. In contrast to the basis function calibration, which leaves the variance unchanged, this accounts now for heteroscedasticity introduced by the measurement error process. The subscript 'c' has been chosen to indicate that both quantities are affected by calibration.

Using the Fisher scoring algorithm, the corrected posterior mean of the weights and its approximative variance are given as

$$\Sigma = F(\boldsymbol{\omega})^{-1}, \tag{3.22}$$

$$\boldsymbol{\mu} = \Sigma \Phi_c^{\mathrm{T}} B_c \mathbf{y}. \tag{3.23}$$

Taking the inverse expected Fisher matrix as posterior covariance matrix is according to Lin & Zhang (1999) and Fahrmeir & Tutz (2001) and uses Laplace's approximation (cf. e.g. Tierney & Kadane (1986), MacKay (2003)). It is stressed here, that the standard errors based on (3.22) are not properly corrected for the inherent covariate measurement error if the true moments of $p(\xi|X)$ are not given, but instead must be estimated. So, correct standard errors remain an open problem.
The posterior moments (3.22) and (3.23) are then used in the type II maximum likelihood estimation of the hyperparameters. Details on this estimation procedure is described in the following subsection.

Estimating the hyperparameters

Recalling the true underlying model of the data

$$Y = \Phi(\xi)\boldsymbol{\omega} + \epsilon, \quad \epsilon \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2),$$

the 'ideal' likelihood of the data, based on the latent covariate ξ is Gaussian. However, the 'observed' likelihood of the data $\mathbf{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ based on the error-prone covariate observations $\mathbf{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$ is usually **no longer** Gaussian

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}, \sigma^2) \quad \neq \quad \mathcal{N}(\mathbf{y}|\Phi\boldsymbol{\omega}, \sigma^2), \tag{3.24}$$

where the design matrix Φ contains the intercept and J basis vectors evaluated at N observations $x_i, i = 1, \ldots, N$.
This is in contrast to measurement error in classical linear regression, where (3.24) is Gaussian with moments readily available (cf. Carroll et al. (1995), section 7.9.2). When the predictor contains nonlinear basis vectors in ξ this becomes a non-standard density.
In order to derive the observed marginal likelihood of the data $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \sigma^2)$ to perform the hyperparameter optimization, one has to find a sensible approximation for the observed likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}, \sigma^2)$ in order to solve

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \sigma^2) \quad = \quad \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}, \sigma^2)p(\boldsymbol{\omega}|\boldsymbol{\alpha})}{p(\boldsymbol{\omega}|\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}, \sigma^2)}. \tag{3.25}$$

In the following it is shown how a Gaussian approximation to the likelihood leads to a Gaussian marginal likelihood.

Based on the ideal likelihood being Gaussian and the fact that the measurement error introduces heteroscedasticity into the model (cf. (3.15)), a sensible approximation for the observed likelihood is given by a Gaussian distribution with heteroscedastic variance

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}, B_c) \quad \approx \quad \frac{|B_c|^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \Phi_c\boldsymbol{\omega})^{\mathrm{T}} B_c (\mathbf{y} - \Phi_c\boldsymbol{\omega})\right). \tag{3.26}$$

Here, the observed mean and variance function (3.12) and (3.15) are contained in Φ_c and B_c (cf. 3.21).
Now, in order to find the explicit form of the marginal likelihood, it is convenient to rewrite (3.25) as

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\alpha},B_c)p(\boldsymbol{\omega}|\mathbf{y},\mathbf{x},\boldsymbol{\alpha},B_c) \;=\; p(\mathbf{y}|\mathbf{x},\boldsymbol{\omega},B_c)p(\boldsymbol{\omega}|\boldsymbol{\alpha}),$$

and then to expand the right hand side. After having collected all terms containing $\boldsymbol{\omega}$ into the weights posterior, the remainder is the sought marginal likelihood.
Now, expanding the right hand side $p(\mathbf{y}|\mathbf{x},\boldsymbol{\omega},B_c)p(\boldsymbol{\omega}|\boldsymbol{\alpha})$ gives

$$\begin{aligned}(2\pi)^{-N/2}|B_c|^{1/2}\exp\left(-\tfrac{1}{2}(\mathbf{y}-\Phi_c\boldsymbol{\omega})^{\mathrm{T}}B^c(\mathbf{y}-\Phi_c\boldsymbol{\omega})\right)\\ \times(2\pi)^{-(J+1)/2}|A|^{1/2}\exp\left(-\tfrac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}\right).\end{aligned} \tag{3.27}$$

Collecting all terms in $\boldsymbol{\omega}$ yields the weights posterior as

$$\begin{aligned}p(\boldsymbol{\omega}|\mathbf{y},\mathbf{x},\boldsymbol{\alpha},B_c) \quad &\approx \quad (2\pi)^{-\frac{(N+1)}{2}}\,|\,\Sigma\,|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu})\right),\\ \text{where}\quad &\Sigma=(\Phi_c^{\mathrm{T}}B_c\Phi_c+A)^{-1},\\ &\boldsymbol{\mu}=\Sigma\Phi_c^{\mathrm{T}}B_c\mathbf{y}.\end{aligned} \tag{3.28}$$

The posterior moments derived via this approximation of the observed likelihood (3.26) is in concordance with their derivation from Fisher scoring above (cf. (3.22) and (3.23)).
Having the weights collected, the remainder of (3.27) is the marginal likelihood given the observations

$$\begin{aligned}p(\mathbf{y}|\mathbf{x},\boldsymbol{\alpha},B_c) \quad &\approx \quad (2\pi)^{-\frac{N}{2}}\;|\;C\;|^{-\frac{1}{2}}\;\exp\left(-\frac{1}{2}\mathbf{y}^{\mathrm{T}}C^{-1}\mathbf{y}\right),\\ \text{where}\quad &C=B_c^{-1}+\Phi_cA^{-1}\Phi_c^{\mathrm{T}}.\end{aligned} \tag{3.29}$$

The approximate marginal likelihood (3.29) is of the same form as the 'ideal' marginal likelihood (2.10), in the error free case as described in Section 2.1.2. However, the calibrated basis functions Φ_c from (3.21) replace the unobservable Φ. Furthermore, the diagonal matrix B_c from (3.21) accounts for the

heteroscedasticity introduced by the measurement error and replaces the former $\beta\mathbf{I} := \sigma^{-2}\mathbf{I}$.
In case of uniform hyperpriors on a logarithmic scale, estimators for the hyperparameters are derived by differentiating the logarithm of the marginal likelihood (3.29). This yields the objective function

$$\mathcal{L} \approx -\frac{1}{2}\left(\log \mid B_c^{-1} + \Phi_c A^{-1}\Phi_c^{\mathrm{T}} \mid + \mathbf{y}^{\mathrm{T}}(B_c^{-1} + \Phi_c A^{-1}\Phi_c^{\mathrm{T}})^{-1}\mathbf{y}\right), \qquad (3.30)$$

which is similar to the objective function in the error free case (2.15).
In the simulations, the one step optimization scheme of the marginal likelihood (cf. paragraph '$\boldsymbol{\alpha}$-RULE 3' in Section 2.1.2), based on the decomposition of the marginal likelihood by Tipping & Faul (2002) is applied . This yields the non-iterative updating rule for a single α_j

$$\alpha_j = \begin{cases} \frac{s_j^2}{q_j^2 - s_j} & \text{if } q_j^2 - s_j > 0 \\ \infty & \text{otherwise} \end{cases}. \qquad (3.31)$$

For the sake of clarity, $q_j = \phi_{c_j}^{\mathrm{T}} C_{-j} - 1\mathbf{y}$ and $s_j = \phi_{c_j}^{\mathrm{T}} C_{-j}^{-1} \phi_{c_j}$ are defined here, with the column vector $\phi_{c_j} := \mathbb{E}(\phi_j(\xi)|\mathbf{x})$ being the jth calibrated basis functions in the calibrated design matrix $\Phi_c = [\phi_{c_0}, \phi_{c_1}, \ldots, \phi_{c_J}]$. Here, $C_{-j} = C - \alpha_j \phi_{c_j} \phi_{c_j}^{\mathrm{T}}$ denotes the covariance matrix in (3.29) with the influence of basis vector ϕ_{c_j} removed. When $\alpha_j = \infty$, the corresponding basis is removed.

The updating rule for the variance σ^2 $(= \beta^{-1})$ is derived from differentiating the objective function (3.30) with respect to $\log \beta$ and equating to zero. Therefore the diagonal matrix B_c is rewritten as $B_c = \beta(B_c/\beta)$ and differentiation with respect to $\log \beta$ is only applied to the pre-multiplied β in front of the parentheses. This gives the following updating rule for σ^2

$$\sigma^{2\,new} = \frac{(\mathbf{y} - \Phi_c \boldsymbol{\mu})^{\mathrm{T}} B_c (\mathbf{y} - \Phi_c \boldsymbol{\mu})}{N - \sum_j \gamma_j} \frac{1}{\beta^{old}}. \qquad (3.32)$$

The post-multiplied β^{old} comes from the re-writing of $B_c = \beta(B_c/\beta)$, with the β in parenthesis being neglected for differentiation.

Both updating rules, (3.31) and (3.32) rely on the Gaussian approximation of the observed likelihood in (3.26). Most importantly, they take into account that there is additional variance in the responses due to covariate measurement error. This fact is expected to become manifest in the estimation of the variance parameter σ^2. This will be investigated later in the simulations.

3.1.3 SIMEX

All of the calibration methods discussed earlier require the knowledge of the conditional distribution $p(\xi|X)$. However, when the assumption of ξ being Gaussian is not supported and consequently $p(\xi|X)$ is not a normal density, the required calculation of the calibrated design matrix may become nasty (cf. Section 3.1.1).
Here, the competing SIMEX approach (Cook & Stefanski (1994)) comes into play. It relies on an experimental study of the effect of measurement error on the outcome of a naive analysis. The core idea is to predict the estimates of a error free analysis based on an simulation experiment.
SIMEX in a non-parametric regression context is presented by Carroll et al. (1999), which is successfully adopted by Rummel (2005) for the RVM regression. The basic intuition behind that concept has already been sketched in Section 2.3.2.

Particularly in the case of flexible regression models, the effect of measurement error on the estimated prediction function is hard to forecast. Here, it seems reasonable, to study the impact of measurement error on the prediction $\widehat{f}(\xi_k)$ for a set of ξ_k's. The details to perform this correction methods are presented now:

1a) Random errors $\delta_i^* \sim \mathcal{N}\left(0, \frac{\sigma_{\delta^*}^2}{m}\right)$ are generated and added to the observed $x_i, i = 1, \ldots, x_N$. Here, x_i may denote the mean of m replicate measurements $x_{i1}, x_{i2} \ldots, x_{im}$. It is for simplicity assumed that the number of replicates are identical for all objects in the sample.

1b) Then a standard RVM analysis is performed using these 'new' data containing the additional error.

These steps are repeated sufficiently often to obtain a series of estimates $\widehat{f}_1(\xi_k), \widehat{f}_2(\xi_k), \ldots, \widehat{f}_B(\xi_k), \quad B = 50 - 200$ at ξ_k.

2) Averaging over these estimates yields $\widehat{f}(\xi_k) = \sum_{s=1}^{B} \widehat{f}_s(\xi_k)$.

Finally, this whole scheme is repeated for a set of different error variances $\sigma^2_{\delta^*} = c \cdot \sigma^2_\delta$. The multiplication factor c is typically chosen to be $c = 0, 0.5, 1, 1.5, \ldots$, where $c = 0$ corresponds to the naive analysis.
This yields a series of mean estimates $\widehat{f}_{c_1}(\xi_k), \widehat{f}_{c_2}(\xi_k), \ldots$ depending on the factor c of the error variance.
Plotting these mean estimates versus the variance of the measurement error may now reveal a pattern, of how the measurement error affects the estimate for $f(\xi_k)$. Fitting a line to the error contaminated estimates $\widehat{f}_c(\xi_k)$ and extrapolate to $c = -1$ yields the desired SIMEX estimate for the mean function at ξ_k.
Figure 3.2 displays the typical attenuation of the estimation $\widehat{f}_c(\xi_k)$ with increasing artificial error variance $\sigma^2_{\delta^*}$ and the final extrapolation step to zero measurement error at c=-1.
Particular care must be taken to fit an adequate model to the estimates $\widehat{f}_c(\xi_k)$. This should ideally be based on theoretical considerations and model diagnostics. Still, extrapolation is a risky task. In many problems the magnitude of the error variance is such that the curvature in the best extrapolant is small and adequately modeled by the quadratic extrapolant (cf. Carroll et al. (1995), section 4.3.4).

In contrast to the calibration methods described above, and most advantageously, no assumptions about the distribution of the latent ξ has to be made. Moreover, the assumption of δ being normally distributed, which is inherent in step 1a) above, is not critical in practise. However, a basic requirement is the classical additive structure of the measurement (2.50) in some scale, e.g. $g(X) = g(\xi) + \delta$, where δ is independent of ξ and has mean zero and variance σ^2_δ. Approaches for non-additive measurement error rely on transformations

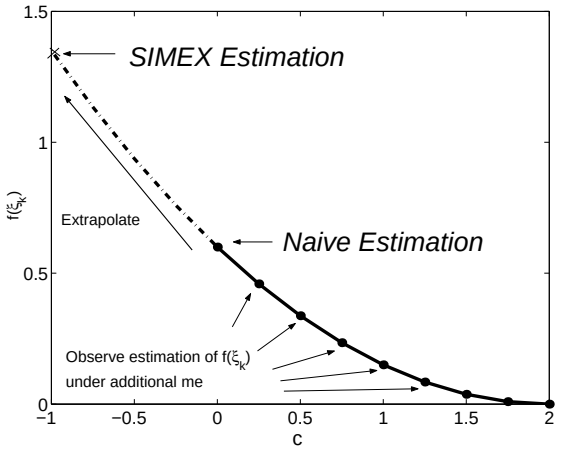

Figure 3.2: Manually inflating the error variance and estimating $\widehat{f}(\xi_k)$ allows for studying the effect of measurement error on the analysis. Plotting $\widehat{f}(\xi_k)$ against the multiples of the variance, fitting a line and extrapolate it to the case of zero measurement error yields the SIMEX estimate.

and are described by Carroll et al. (1995), section 4.4. The measurement error variance can be estimated from replication data or validation data using (2.52) or must be guessed.
The Gaussian response is not restricted to a certain domain, so $\widehat{f}_{SIMEX}(X)$ will always be in $\mathbb{R}$ and no additional precautions need to be taken here. In the following simulation study, a quadratic extrapolation based on the naive analysis (c=0) and the mean estimates over $B = 50$ repetitions for each $c \in \{0.5, 1, 1.5, 2\}$ is used to attain the SIMEX estimates.
It is important to stress once more the computational heaviness of SIMEX, which is a pronounced burden compared to the previous methods.

3.2 Simulation study

The presented correction methods, basis function calibration, structural quasi likelihood and SIMEX are now compared in a simulation study. As a state of the art reference method the approach of Berry et al. (2002) is used. They

present Bayesian P-splines for measurement error problems using a MCMC approach. All methods are investigated in a variety of data scenarios.
Firstly, these data scenarios are described, before the competing methods are contrasted in some essential respects. Finally the results of the simulation study are presented and discussed.

3.2.1 The data

For each data scenario 200 data sets are simulated.
There are always two replicates ($m_i = 2$) available containing classical additive measurement error with $\mu_\delta = 0$. Thus, each surrogate observation $x_i, i = 1, \ldots, N$ represents the average over these two replicates. From these replicates the measurement error variance σ_δ^2 will be estimated by the usual components of variance analysis, cf. (2.52) in Section 2.3.2.
In the first five data cases the $\xi_i, i = 1, \ldots, N$ are generated as independent normal random variables with mean μ_ξ and variance σ_ξ^2. Case 6 and case 7 study deviations from that assumption. Case 8 represents a function that is difficult to fit with the described methods.
The level of measurement error variance is different for the data scenarios. As a consequence of having two replicates, the measurement error variance of the surrogates $x_i = \frac{x_{i1}+x_{i2}}{2}$ is only half the error variance that is stated below in the respective cases.
Predictions of the methods were obtained for 101 grid values in the interval $[a, b]$. The specific limits a and b are expected to contain most of the distribution for ξ.
The responses are generated randomly from the (true) mean functions $m(\xi)$, with variance σ^2. The series of simulation includes the following eight data cases:

Case 1: The mean function of the data is given by

$$m(\xi) = \frac{\sin(\pi\xi/2)}{1 + 2\xi^2(\text{sign}(\xi) + 1)},$$

with $N = 100$, $a = -2.0$, $b = 2.0$, $\sigma^2 = 0.3^2$, $\sigma_\delta^2 = 0.8^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1$

Case 2: Same as case 1 except $N = 200$

Case 3: Same as case 1 except $N = 500$

Case 4: The mean function of the data is given by

$$m(\xi) = 1000\xi_+^3(1-\xi)_+^3,$$

where $\xi_+ = \xi I(\xi > 0)$, with $N = 200$, $a = 0.1$, $b = 0.9$, $\sigma^2 = 0.0015^2$, $\sigma_\delta^2 = (3/7)\sigma_\xi^2$, $\mu_\xi = 0.5$ and $\sigma_\xi^2 = 0.25^2$

Case 5: The mean function of the data is given by

$$m(\xi) = 10\sin(4\pi\xi),$$

with $N = 500$, $a = 0.1$, $b = 0.9$, $\sigma^2 = 0.05^2$, $\sigma_\delta^2 = 0.141^2$, $\mu_\xi = 0.5$ and $\sigma_\xi^2 = 0.25^2$

Violations of the assumptions that ξ and ϵ are normally distributed are studied in the following cases:

Case 6: The same as case 1 above except that ξ is a standardized $\chi^2(4)$ random variable. MSE is evaluated on $[-1.25, 2.00]$.

Case 7: The same as case 1 above except that ξ is a standardized $\chi^2(4)$ random variable and ϵ is generated as a Laplace random variable. MSE is evaluated on $[-1.25, 2.00]$.

A plateau function is difficult to model with the RVM methods using RBF kernels or the MCMC approach using 2nd order truncated power series. This model misspecification is investigated here:

Case 8: The same as case 1 above except that

$$m(\xi) = H(100\xi) + H\left(-100(\xi - 0.5)\right),$$

where $H(\xi) = (1 + \exp(-\xi))^{-1}$.

Figure 3.3 and 3.4 display example data sets for each scenario as well as the mean function. Despite there are two replicate measurements available and usually the average is taken to perform the model estimation, here only a single measurement is displayed. These figures demonstrate the challenge of inferring the underlying mean function - a task that can hardly be performed 'by eye'.

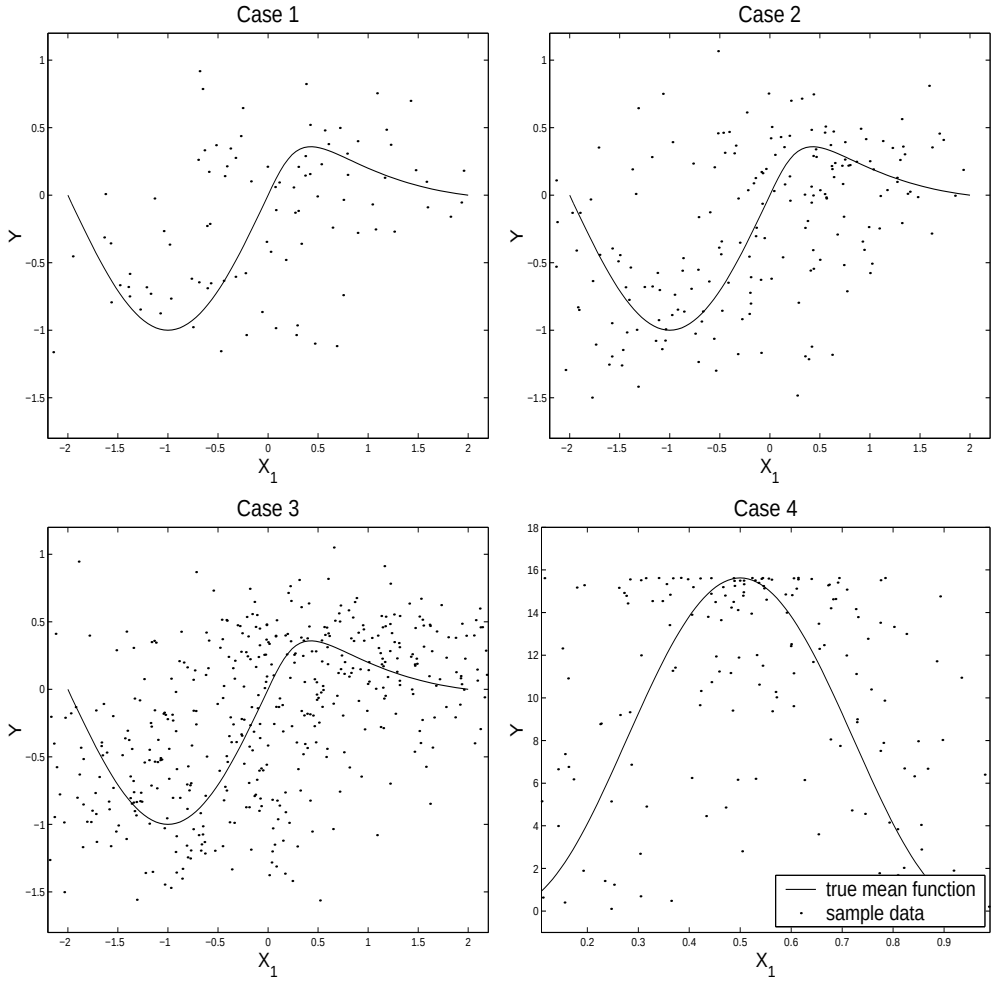

Figure 3.3: Example data sets for cases 1-4 and the respective true mean function. The response is plotted versus one error-prone measurement, i.e. no averaging over the two available replicates is done here. Only the range $[a, b]$ *on which the methods will be evaluated is shown.*

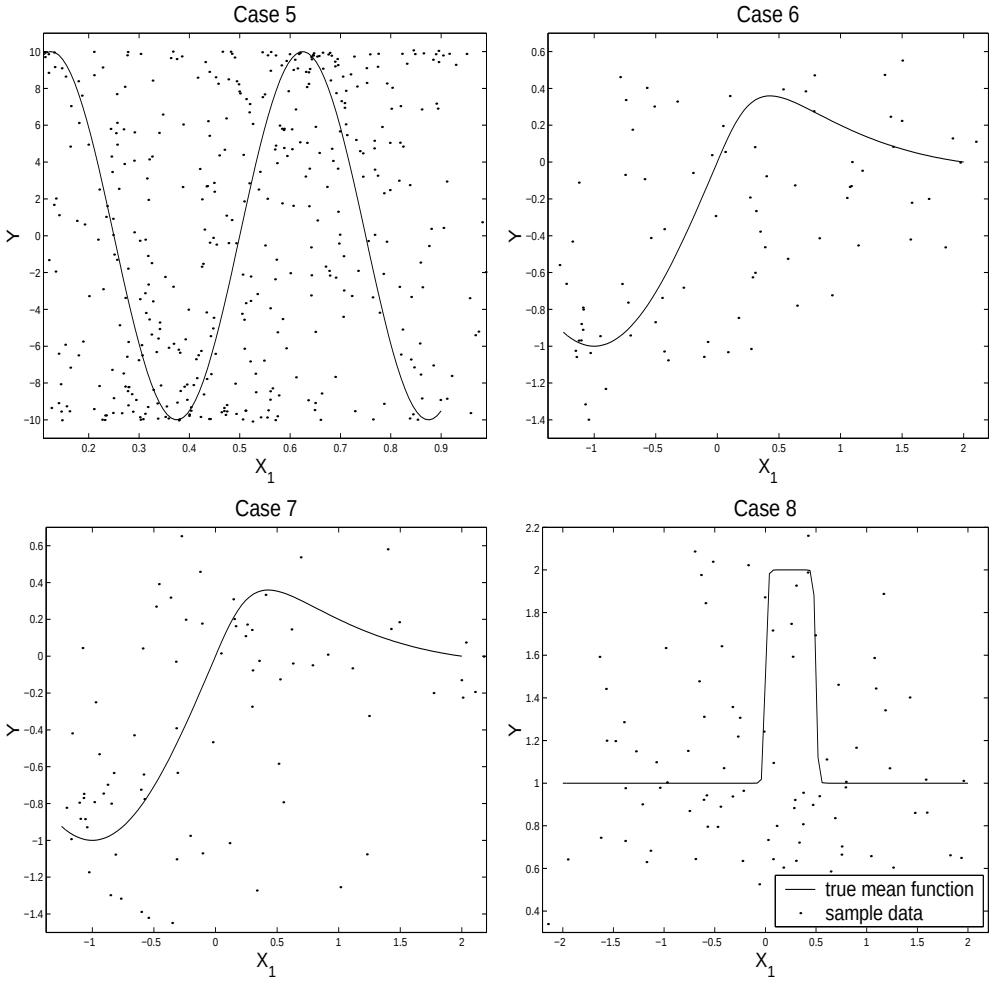

Figure 3.4: Example data sets for cases 5-8 and the respective true mean function. The response is plotted versus one error-prone measurement, i.e. no averaging over the two available replicates is done here. Only the range $[a, b]$ on which the methods will be evaluated is shown.

3.2.2 Specification details of the methods

Some general settings of the presented methods are briefly described in the following paragraph.

Basis calibration, structural quasi likelihood and SIMEX use radial basis functions centered on 100 knots located at the quantiles of the observed data. For SIMEX the knots are located at the quantiles of the artificially generated observations in each simulation step.
The kernel parameter η (cf. (2.3)) is selected from a set of admissible values and is chosen as to give the maximal alignment of the basis functions with the observations

$$\eta := \max_{\eta_m} \parallel \Phi_{\eta_m}^{\mathrm{T}} C^{-1}\mathbf{y} \parallel,$$

where Φ_{η_m} is the design matrix constructed under the mth kernel parameter η_m from the list and C is the covariance matrix of the marginal likelihood (cf. 2.10). Here, $\| \cdot \|$ denotes the quadratic norm.
By including this sub-procedure into the RVM optimization scheme, an optimal $\eta \neq \eta_0$ (η_0 denotes the starting state) can usually be found in the very first iteration. If so, this sub-procedure is blocked for the rest of the optimization algorithm. Otherwise this sub-procedure is active until an optimal $\eta \neq \eta_0$ is found. This ad-hoc approach is motivated by Tipping & Faul (2003), who introduce the quantity $\Phi^{\mathrm{T}} C^{-1}\mathbf{y}$ in the context of updating the hyperparameters $\boldsymbol{\alpha}$ and works surprisingly well in practise. While the naive RVM uses the observed covariate values x_i in the design matrix Φ_{η_m}, SIMEX uses the artificially simulated covariate observations. For simplicity, basis calibration and structural quasi likelihood simply copy the optimal η found by the naive approach.
All methods use the analytic updating scheme of the precisions $\boldsymbol{\alpha}$ (cf. paragraph '$\boldsymbol{\alpha}$-Rule 3' in Section 2.1.2) and for the modified version in the structural quasi likelihood approach in Section 3.1.2. That is, starting from a model with only the intercept included, in each iteration step a basis function can be either deleted, updated or newly introduced into the model, according to what gives the highest improvement in the marginal likelihood.

All methods developed here are also compared to the state-of-the-art error correction approach by Berry et al. (2002). They describe a Bayesian P-spline approach where a set of 30 second order truncated power series basis are fit to the data. They construct a MCMC sampling scheme regarding the smoothing parameter, the coefficients and all parameters associated with the distributions of Y, ξ, δ as random variables. The main idea is here is to regard the latent $\xi_i, i = 1, \ldots, N$ as additional unknown parameters in the spirit of data augmentation (cf. Section 2.2.1) and benefit from the fact that all full conditional distributions take on the figure of standard distributions given the ξ_i. The final parameter estimates are based on 2000 MCMC samples (after a burn-in period of 2000 runs) from the full conditionals. These samples ideally represent an empirical version of the joint posterior. The general reasoning behind that methodology has already been described in Section 2.2.1 and Section 2.3.2 of chapter 2.
Berry et al. (2002) accompanied their article with a MATLAB implementation of the described methods which is downloadable from the homepage `http://www.stat.tamu.edu/~carroll/matlab_programs/software.php`. The results from this implementation are used here as reference.

Table 3.1 contrasts all compared methods in some essential respects.

Method	$\mathbf{RVM_{naive}}$	$\mathbf{RVM_{BC}}$ [basis calibration]	$\mathbf{RVM_{SQL}}$ [structural quasi likelihood]	$\mathbf{RVM_{SIMEX}}$	**BRS** [bayesian regression splines]
type of basis function	RBF	RBF	RBF	RBF	2nd order truncated power series
potential / effective number of basis funct. w/o intercept	100/usually very few, see results	100/usually very few, see results	100/usually very few, see results	100/usually very few	30/30
knot selection	quantiles of error-prone data	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$	quantiles of generated SIMEX-observations	same as $\mathbf{RVM_{naive}}$
error correction	none	Correction of observed mean, by using calibrated basis functions $\mathbb{E}(\Phi(\xi)\|X)$ instead of $\Phi(X)$.	Correction of observed mean and variance. Approximation for the observed marginal likelihood for hyperparameter estimation	The effect of additive error is studied in a simulation study and then corrected	The true ξ_i are regarded as unknown parameters and sampled in an MCMC approach
unknown parameters in the response model (and their estimation scheme)	the fundamental model parameters $\boldsymbol{\omega}$ (posterior mean), hyperparameters $\boldsymbol{\alpha}$ (marginal likelihood optimization), σ^2(marginal likelihood optimization), η (grid search)	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$	fundamental model parameters $\boldsymbol{\omega}$ (sampled in Gibbs-step), true covariate values ξ_i (sampled in MH-step), $\gamma := \frac{\alpha}{\sigma^2}$, with α :=smoothing parameter (sampled in Gibbs-step), σ^2 (sampled in Gibbs-step)
unknown parameters in the error model (and their estimation scheme)	σ_δ^2 (from components of variance analysis using replicates), μ_ξ (analysis of variance formula), σ_ξ^2 (analysis of variance formula)	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$	$\lambda := \frac{\sigma_\xi^2}{\sigma_\xi^2+\sigma_\delta^2/m}$ (discrete search), μ_ξ (sampled in Gibbs-step), σ_ξ^2(sampled in Gibbs-step)

Table 3.1: Overview of methods

3.2.3 The results

The quality of the correction methods is investigated by means of various criteria including mean squared error and pointwise bias. Other secondary properties like the quality of the σ^2-estimation and the effective number of kernels are also considered.

MSE:

The mean squared error is computed over a grid of 101 equidistant values in the interval $[a, b]$ as given earlier (cf. Section 3.2.1)

$$\text{MSE} = \frac{1}{101} \sum_{k=1}^{101} \Big(m(\xi_k) - \hat{f}(\xi_k) \Big)^2 ,$$

where $m(\xi)$ is the true mean function from above (cf. Section 3.2.1). Here, $\hat{f}(\xi_k)$ is an estimate for $m(\xi_k)$.

Table 3.2 presents summary results for the MSE from the 200 simulations for each data scenario. Since the results of the MCMC approach as stated in the original article by Berry et al. (2002) for some cases clearly differ from the results based on the present simulation study, they are additionally given here for reasons of completeness. These differences are partly explainable by the use of smoothing splines and (slightly) different priors in the original article compared to the P-splines and prior specifications used here (cf. Table 3.1). But beyond that, the differences are surprisingly articulate though Berry et al. (2002) state that „smoothing splines and P-splines with 30 knots give much the same result“.
The smallest mean MSE value in each scenario is in boldface.
Except in cases 7 and 8 the naive estimation is always dominated by **all** other correction methods. Usually, $\mathbf{RVM_{SIMEX}}$ yields the minimum gain while $\mathbf{RVM_{SQL}}$ (with exception of cases 2, 3 and 8) dominates all other implemented methods though it seems not to be dramatically superior to $\mathbf{RVM_{BC}}$. The large MSE values in case 4 and case 5 are not indicating a miserable fit, but are due to the scale of the true function. Against one's

intuition, in case 6 and case 7, where ξ is not normally distributed, the robustness of the functional $\mathbf{RVM_{SIMEX}}$ approach (cf. 2.3.2) does not become manifest in the MSE results. Instead $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{SQL}}$ dominate the scene. The Bayesian P-spline seems to profit the most from a larger sample size (case 2 and case 3). It can be seen from the accompanying median values that the distributions of the MSE values is typically right skewed for all correction methods.

Mean squared error Mean (SE) / Median (all $\times 10^2$)				
Method	Case 1	Case 2	Case 3	Case 4
RVM	.80 (.04) / .66	.50 (.02) / .43	.28 (.01) / .26	4.30 (1.05) / .00
$\mathbf{RVM_{naive}}$	5.42 (.14) / 5.13	4.79 (.09) / 4.73	4.52 (.05) / 4.52	434.40 (9.76) / 431.97
$\mathbf{RVM_{BC}}$	4.08 (.23) / 3.38	2.32 (.14) / 1.83	1.10 (.06) / .92	53.68 (3.61) / 40.52
$\mathbf{RVM_{SQL}}$	**3.76** (.20) / 3.23	2.23 (.12) / 1.77	1.03 (.04) / .89	**51.98** (2.64) / 38.92
$\mathbf{RVM_{SIMEX}}$	4.48 (.23) / 3.68	3.25 (.14) / 2.80	2.39 (.08) / 2.13	145.88 (7.89) / 112.22
BRS	6.05 (.38) / 4.48	**2.15** (.09) / 1.87	**.69** (.03) / .57	56.02 (3.61) / 39.94
BRS^*	2.84 (-) / -	1.56 (-) / -	1.47 (-) / -	195 (-) / -

Method	Case 5	Case 6	Case 7	Case 8
RVM	.13 (.00) / .05	.55 (.03) / .45	1.05 (.06) / .82	3.30 (.09) / 3.02
$\mathbf{RVM_{naive}}$	1885.96 (18.36) / 1888.04	4.59 (.11) / 4.45	5.18 (.12) / 5.13	7.90 (.10) / 7.82
$\mathbf{RVM_{BC}}$	288.44 (13.22) / 257.09	3.02 (.11) / 2.67	4.15 (.14) / 3.71	7.84 (.16) / 7.40
$\mathbf{RVM_{SQL}}$	**275.75** (13.07) / 234.03	**2.87** (.10) / 2.50	**4.04** (.14) / 3.63	**7.76** (.13) / 7.40
$\mathbf{RVM_{SIMEX}}$	510.38 (17.46) / 466.27	4.27 (.21) / 3.59	5.41 (.28) / 4.44	9.32 (1.27) / 7.47
BRS	580.06 (20.68) / 535.04	4.48 (.09) / 4.58	5.72 (.22) / 5.10	9.75 (.22) / 9.21
BRS^*	1031 (-) / -	2.69 (-) / -	2.49 (-) / -	7.41 (-) / -

$*$ MSE results from Berry et al. (2002) using smoothing splines and a slightly modified prior specification.

Table 3.2: The mean squared error results for the simulation. In each column, the smallest mean value among the implemented correction methods is in boldface.

Variance estimation:

Table 3.3 displays the estimation for the variance parameter σ^2. This clearly indicates the worthiness of the additional expense in the structural quasi likelihood $\mathbf{RVM_{SQL}}$ compared to the basis calibration $\mathbf{RVM_{BC}}$. However, in case 4 and case 5 the $\mathbf{RVM_{SQL}}$ mean estimate is far away from the true value, which is partly due to the scale of the respective true function.

Model complexity:

From Table 3.4 it is seen that the average number of utilized kernels for function estimation is dramatically lower than for the MCMC approach of Berry et al. (2002). While the MCMC method pre-specifies a set of 30 basis functions and retains that number of functions throughout the complete algorithm, the RVM-methods select relevant basis function for estimation from an arsenal of 100 radial basis functions and one intercept.
The optimal η is usually smaller for the naive approach, i.e. the utilized basis functions of the $\mathbf{RVM_{naive}}$ and consequently for the calibration methods are less peaked than for the error free RVM.
Taking measurement error into account might lead to a more sensible η selection for $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{SQL}}$ and possibly boost the performance of these methods. Therefore, cases 1-8 are re-run for $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{SQL}}$ with η-selection realized by a preceding RVM, however now, utilizing the (standard) regression calibrated observations $\mu_{\xi|x_i} := \mathbb{E}(\xi|x_i)$ (cf. 2.58) instead x_i. This is clearly an ad-hoc approach and more sophisticated methods are desirable. However, the new MSE results in Table 3.5 already underline the high potential that lies in an sensible kernel selection. Particularly cases 1-2 are improved by this ad-hoc amendment.

Variance estimation Mean				
Method	Case 1 (σ^2=.09)	Case 2 (σ^2=.09)	Case 3 (σ^2=.09)	Case4 (σ^2=2.25e-6)
RVM	.0907	.0928	.0918	0.0542
$\mathbf{RVM_{naive}}$	.2113	.2147	.2174	12.0153
$\mathbf{RVM_{BC}}$	.2142	.2153	.2168	11.9147
$\mathbf{RVM_{SQL}}$	**.1346**	**.1202**	.1081	4.0792
BRS	.1785	.1317	**.1038**	**.1102**

Method	Case 5 (σ^2=2.5e-3)	Case 6 (σ^2=.09)	Case 7 (σ^2=.09)	Case 8 (σ^2=.09)
RVM	.0642/	.0881	.1766	.1179
$\mathbf{RVM_{naive}}$	36.8150	.2075	.2927	.2168
$\mathbf{RVM_{BC}}$	36.8923	.2072	.2929	.2203
$\mathbf{RVM_{SQL}}$	5.2130	**.1218**	**.2131**	**.1916**
BRS	**.1785**	.1761	.2603	.2454

Table 3.3: The mean estimate of σ^2. In each column the value closest to the respective true value among the implemented correction methods is in boldface. The SIMEX method was not designed to return an estimate of the variance and thus is left out here.

Kernel parameter and number of effective kernels Mean / Mean				
Method	Case 1	Case 2	Case 3	Case 4
RVM	1.36/4.3600	1.32/4.4200	1.30/5.5000	21.01/17.7750
$\mathbf{RVM_{naive}}$	.82/2.6850	.80/2.6200	.76/2.9800	15.53/4.4950
$\mathbf{RVM_{BC}}$	.82/3.6350	.80/4.3350	.76/4.7050	15.53/1.5250
$\mathbf{RVM_{SQL}}$	.82/3.7150	.80/4.4350	.76/4.8750	15.53/1.3800
BRS	-/30	-/30	-/30	-/30

Method	Case 5	Case 6	Case 7	Case 8
RVM	69.11/23.7800	1.73/3.3700	1.71/3.1950	3.63/7.6950
$\mathbf{RVM_{naive}}$	50.95/6.0550	.84/2.6750	.84/2.5050	1.49/2.3450
$\mathbf{RVM_{BC}}$	50.95/8.5600	.84/2.2950	.84/2.2050	1.49/2.2950
$\mathbf{RVM_{SQL}}$	50.95/8.4900	.84/2.6250	.84/2.4100	1.49/2.3850
BRS	-/30	-/30	-/30	-/30

Table 3.4: The average number of utilized kernels. $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{SQL}}$ copy the optimal η-value from $\mathbf{RVM_{naive}}$. The $\mathbf{RVM_{SIMEX}}$ method utilizes different optimal values for η and different numbers of kernels during its simulation phase. Thus it is not considered here. **BRS** utilizes 30 second order regression splines that do not contain additional kernel parameters.

Mean squared error under refined η-selection Mean (previous value) (all $\times 10^2$)				
Method	Case 1	Case 2	Case 3	Case 4
$\mathbf{RVM_{BC}}$	3.73 (4.08)	2.18 (2.32)	1.13 (1.10)	175.12 (53.68)
$\mathbf{RVM_{SQL}}$	3.13 (3.76)	1.91 (2.23)	1.01 (1.03)	190.42 (51.98)

Method	Case 5	Case 6	Case 7	Case 8
$\mathbf{RVM_{BC}}$	328.88 (288.44)	3.00 (3.02)	4.28 (4.15)	7.19 (7.84)
$\mathbf{RVM_{SQL}}$	349.84 (275.75)	2.85 (2.87)	4.18 (4.04)	7.02 (7.76)

Table 3.5: The mean squared error for the $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{SQL}}$ under η-selection, now accounting for the covariate measurement error. The previous values from Table 3.2 are given here in parentheses.

Pointwise bias:

The bias of a method is its expected deviation from the true curve. The pointwise bias of the methods under investigation can be seen from the visualization of the mean functions over the 200 simulations in Figure 3.5 (for cases 1-4) and Figure 3.6 (for cases 5-8).
Already with small sample size in case 1, the correction power for the RVM-methods becomes clear, especially when regarding the fit at the positions of minimum and maximum of the true function.
The bias is again drastically reduced for higher sample sizes (case 2 and 3).
Case 4 shows a more or less consistent correction capacity of all correction methods and even a very slight overestimation of $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{SQL}}$, which coincide here.
Case 5 is satisfactorily covered by all correction methods with slight deficiencies for $\mathbf{RVM_{SIMEX}}$.
In case 6 and case 7, the latent covariate ξ is no longer normally distributed. Here, the expected superiority of the functional $\mathbf{RVM_{SIMEX}}$ compared to the structural $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{SQL}}$, which assume ξ being normal, becomes, however, not manifest.
The plateau function in case 8 is poorly fit by all methods. A potential boost of the RVM methods might result here from allowing basis functions having locally different kernel parameters instead a single global η.

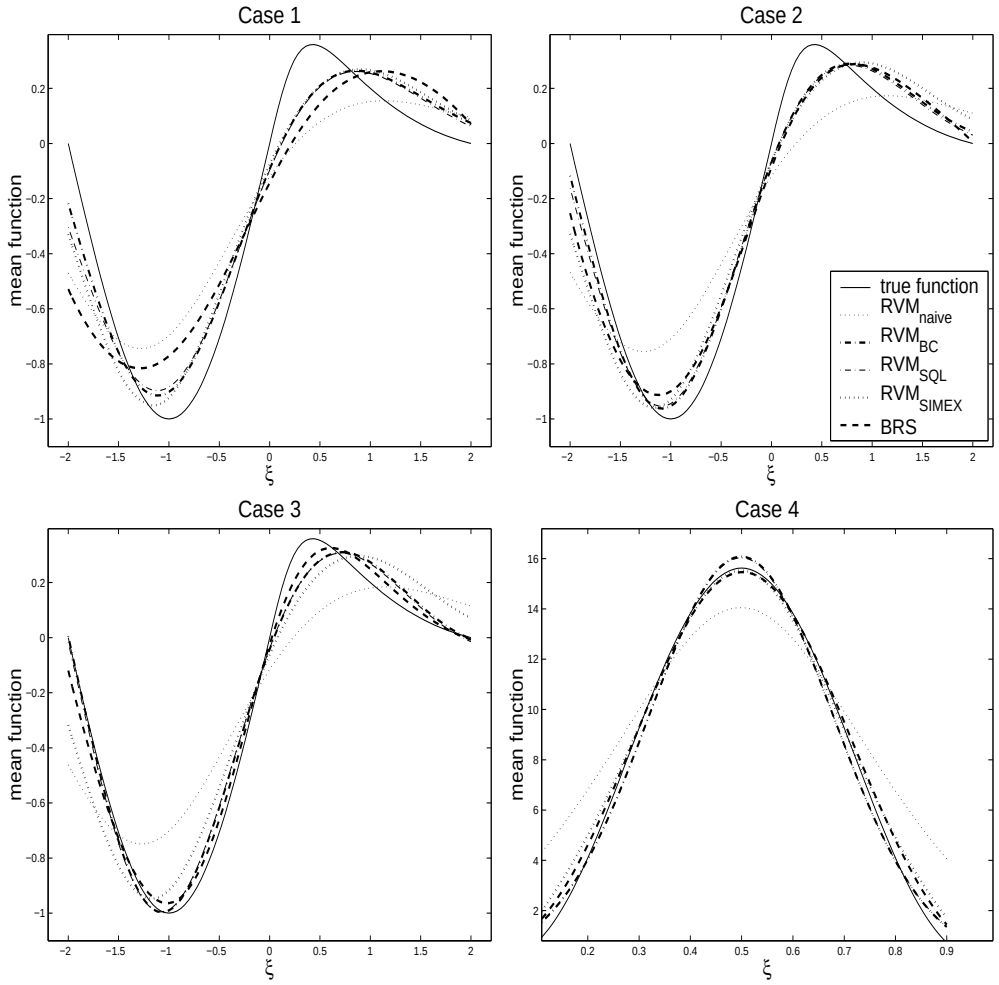

Figure 3.5: The mean functions over 200 simulations for cases 1-4. Cases 1-3 solely differ in the number of observation, which are $N = 100$, $N = 200$, $N = 500$, respectively. Only case 1 reveals accentuated differences between the correction methods. Increasing the sample size obviously boosts all methods except $\mathbf{RVM_{SIMEX}}$. *The RVM without measurement error is left out here for the sake of visibility. Particularly case 4, which has very small σ^2, but pronounced measurement error is well fit by all correction methods.* $\mathbf{RVM_{BC}}$ *and* $\mathbf{RVM_{SQL}}$ *coincide in most cases.*

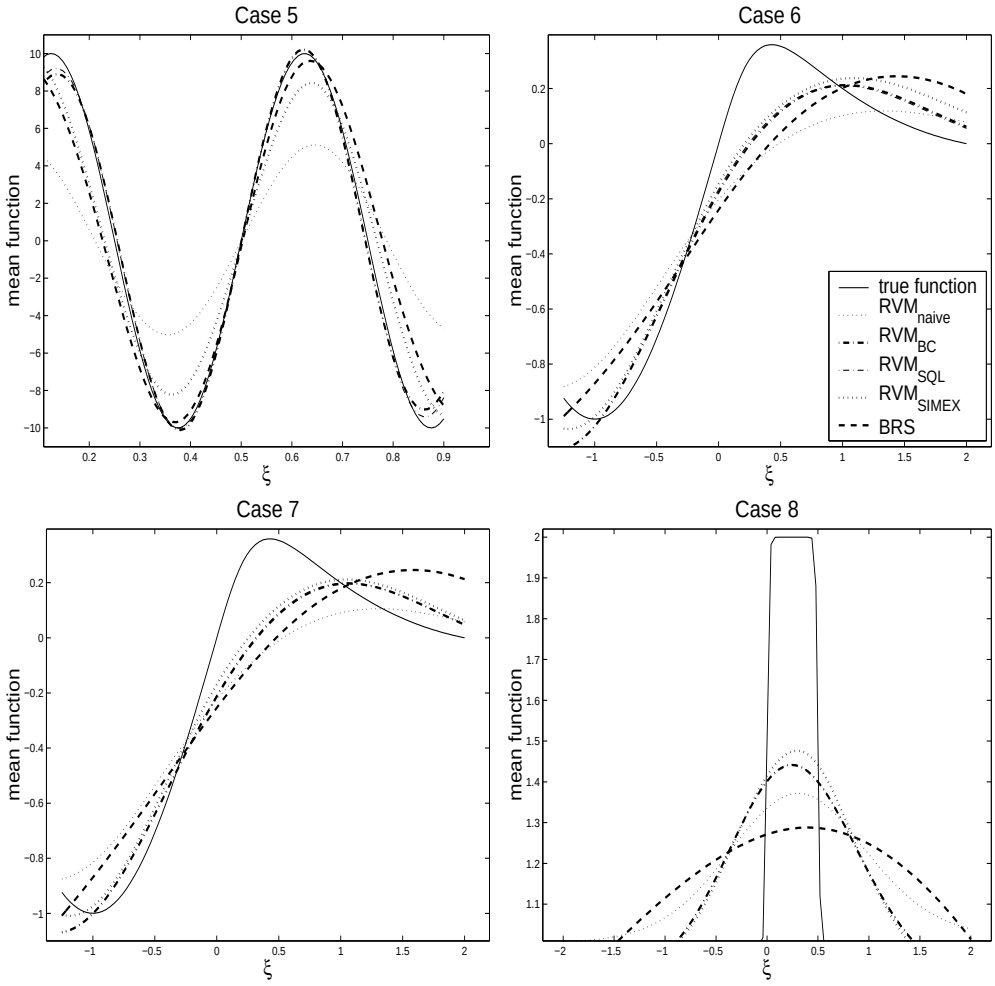

Figure 3.6: The mean functions over 200 simulations for cases 5-8. Case 5 is an oscillating function and data generated under ξ and ϵ being normally distributed, while in case 6 the data is generated under ξ being a standardized $\chi^2(4)$ random variable and in case 7 additionally ϵ being a Laplace random variable. Case 8 is a plateau function which is difficult to fit with RBF kernels and 2nd order truncated power series, respectively. A pronounced gain of using error correction methods is attested in case 5. Under distributional deviations from the model (case 6, 7) and under model misspecification the RVM methods seem to be slightly superior. $\mathbf{RVM_{BC}}$ *and* $\mathbf{RVM_{SQL}}$ *coincide in most cases.*

Chapter 4

Covariate measurement error in flexible binary regression

This chapter is concerned with covariate measurement error in a flexible binary regression model. Most of the ideas underlying the forthcoming methods have been sketched in chapter 2. Now, it is focused on putting flexible regression and error correction sensibly together and giving relevant computational details.

Firstly, the binary relevance vector machine (RVM) regression model, which is the basis for all strategies described here, is shortly recalled and then Section 4.1 contains the development of the error correction methods for the binary case. The respective underlying motives and additional literature head each subsection. A simulation study is conducted in order to investigate the strength of the developed methods.

Real data on nutritional habits and mortality from the German panel of the WHO MONICA project (MONItoring of trends and determinants in CArdiovascular disease, cf. Döring & Kußmaul (1997), Keil (2000)) that have previously analyzed by Augustin (2002) will be re-analyzed by a part of the developed methods.

Finally, a MCMC sampling scheme is developed allowing the estimation of parameters from a flexible binary regression model for longitudinal data.

While this is a highly relevant case in real data situations, here only a few data examples are discussed.

Recalling from chapter 2, the binary RVM is of the form

$$Y = G(\Phi(\xi)\boldsymbol{\omega}) + \epsilon,$$

where $Y \in \{0, 1\}$ and $\mathbb{E}(\epsilon) = 0$. The response function G is typically chosen as $G(z) = (1+\exp(-z))^{-1}$ (logit regression case) or $G(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$ (probit regression case). Beside this specific weighted basis functions form, all methods incorporate a strategy to select relevant basis functions from a large set of potential basis functions. Therefore, those RVM approaches not relying on Markov Chain Monte Carlo (MCMC) techniques employ the following prior distribution over the parameters of the mean model

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{j=0}^{J} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{\alpha_j}{2}\omega_j^2\right). \tag{4.1}$$

Further, Gamma hyperpriors are specified over the hyperparameters collected in $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_J)^{\mathrm{T}}$ and σ^2 if dispersion is included in the model (cf. Section 2.1.1).
Thus the 'ideal mean model' and the 'ideal variance model', motivated from Section 2.3, are

$$\mathbb{E}(Y|\xi) = G\left(\Phi(\xi)\boldsymbol{\omega}\right) \tag{4.2}$$

$$\mathbb{V}(Y|\xi) = \sigma^2 g^2(\xi, \boldsymbol{\omega}, \theta), \tag{4.3}$$

with dispersion parameter σ^2 and $g^2(\cdot) = G\left(\Phi(\xi)\boldsymbol{\omega}\right) \cdot \left(1 - G\left(\Phi(\xi)\boldsymbol{\omega}\right)\right)$ accounting for the inherent heteroscedasticity. The dispersion parameter is, however, not considered in this work, i.e. $\sigma^2 \stackrel{!}{=} 1$, unless stated otherwise.
Though the coefficients $\boldsymbol{\omega}$ are taken to be random parameters, the conditioning on this random vector is (and will be) suppressed in the conditional expectation and variance in (4.2) and (4.3) for the sake of clarity.

4.1 The arsenal of correction methods

Correcting for covariate measurement error in binary regression is generally a more demanding task than in the Gaussian case. This fact is particularly true for flexible binary regression, since the correction methods must now cope with the additional nonlinearity of $\mathbb{E}(Y|\xi)$ in terms of the basis functions. The main aspects of the methods, which are presented here in full detail have already been roughly discussed in Section 2.3.2.

For the calibration methods, error correction again focuses exclusively on the fundamental model parameters of the mean model, i.e. those parameters collected in the vector $\boldsymbol{\omega}$.

Basis calibration as described in Section 3.1.1 is employed again, but in contrast to the Gaussian case it is merely an approximation to the observed mean model here.

The expanded basis function calibration is in the spirit of the former structural quasi likelihood method and aims at a refined approximation of both observed moments with the aid of Taylor series expansion. However, as will be shown the necessary modifications of the estimation routine of $\boldsymbol{\omega}$ must be harmonized with the hyperparameter estimation, which is a creativity challenge here.

The simulation based SIMEX method, as described in full detail earlier in Section 3.1.3, is slightly modified to suit the binary case.

Finally, a MCMC approach to measurement error correction and a MCMC version of the RVM are combined circumventing the approximations that are indispensable in the former methods.

The basis function calibration and SIMEX approach require only minor modification compared to their Gaussian regression settings and thus their description is kept rather compact in the following. All methods are again compared to the naive RVM and a competing MCMC version of P-splines in a concluding simulation study.

4.1.1 Basis function calibration

This section presents the basis function calibration in the binary regression case. The core of this method is the replacement of the design matrix Φ, which is formulated in terms of the latent observations $\xi_i, i = 1, \ldots, N$ and thus is latent itself, by its calibrated version Φ_c (cf. (3.21)). In the Gaussian regression context this has already been described in Section 3.1.1. But here, unlike in the Gaussian regression case, this replacement does no longer yield the exact representation of the observed mean $\mathbb{E}(Y|X)$ in terms of the fundamental model parameters. This can be seen by application of the law of iterated expectations giving

$$\begin{aligned}
\mathbb{E}(Y|X) &= \mathbb{E}(\mathbb{E}(Y|X,\xi)|X) \\
&= \mathbb{E}(\mathbb{E}(Y|\xi)|X) \\
&= \mathbb{E}\left(\left(G\left(\textstyle\sum_{j=1}^{M}\omega_j\phi_j(\xi)+\omega_0\right)\right)|X\right) \\
&\neq G\left(\mu_{\Phi(\xi)|X}\boldsymbol{\omega}\right).
\end{aligned}$$

Here, $\mu_{\Phi(\xi)|X} := \mathbb{E}(\Phi(\xi)|X)$ denotes the row vector of calibrated basis functions given the observed X. The second line is again justified by the non-differentiability of the measurement error, i.e. the assumption that the error is independent of the response.
Thus in binary regression, replacing the latent Φ by the calibrated design matrix Φ_c, which is constructed from the row vectors $\mu_{\Phi(\xi)|x_i}, i = 1, \ldots, N$ (cf. (3.10) in Section 3.1.1), merely yields a *working* model for the observed mean.
The parameter estimation proceeds as described for the non-Gaussian case in Section 2.1.2 via Fisher scoring. The Laplace approximation (cf. e.g. Tierney & Kadane (1986), MacKay (2003)) for the posterior distribution of $\boldsymbol{\omega}$ yields a Gaussian with moments given by

$$\Sigma = (\Phi_c^{\mathrm{T}} B \Phi_c + A)^{-1}, \tag{4.4}$$

$$\boldsymbol{\mu} = \boldsymbol{\omega}_{MP} = \Sigma \Phi_c^{\mathrm{T}} B \mathbf{y}^*, \tag{4.5}$$

now containing the working observations

$$\mathbf{y}^* \quad = \quad \Phi_c \boldsymbol{\omega} + D^{-1}(\mathbf{y} - G(\Phi_c \boldsymbol{\omega})). \tag{4.6}$$

The covariance matrix Σ is here given as the inverse expected Fisher matrix (cf. Lin & Zhang (1999), Fahrmeir & Tutz (2001)). The diagonal matrix B contains the elements

$$B_{ii} = \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})}{\partial (\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})} \right)^2 / \, \mathbb{V}(y_i|x_i)$$

involving the first derivative of the response function with respect to the linear predictor and the following approximation to the observed variance

$$\mathbb{V}(y_i|x_i) \approx G\left(\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega}\right) \cdot \left(1 - G\left(\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega}\right)\right).$$

The diagonal matrix D in (4.6) consists of elements

$$D_{ii} = \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})}{\partial (\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})} \right).$$

Intuitively, the use of the calibrated basis functions, visualized in Figure 3.1 of the previous chapter, leads to larger estimates for the coefficients as explained earlier in Section 2.3.2.

Tipping (2001) assumes for convenience that the marginal likelihood of the working observations $p(\mathbf{y}^*|\boldsymbol{\alpha})$ is approximately Gaussian (cf. Section 2.1.2) yielding the following objective function

$$\begin{aligned} \mathcal{L} \quad &= \quad -\frac{1}{2}\left[\log \mid C \mid + {\mathbf{y}^*}^{\mathrm{T}} C^{-1} \mathbf{y}^*\right] \\ &\text{where} \quad C = B + \Phi_c A^{-1} \Phi_c^{\mathrm{T}}. \end{aligned} \tag{4.7}$$

Since the calibrated basis functions are used in approximating $\mathbb{E}(Y|X)$ it is sensible to formulate the marginal likelihood again in terms of the calibrated design matrix Φ_c. The optimal hyperparameters are then found via the one step maximization of the marginal likelihood as described in Section 2.1.2 utilizing the posterior variance (4.4) and the corrected posterior mean (4.5). In case there is a dispersion parameter included in the model, σ^2 can also be estimated via the objective function (4.7), cf. (2.28) in Section 2.1.2.

4.1.2 Expanded basis function calibration

The expanded basis function calibration presented in this section is a derivative of the approximate quasi likelihood method presented by Carroll & Stefanski (1990). Carroll et al. (1995) describe that approach, which they term as 'expanded regression calibration', for classical linear regression. The basic idea of expanded regression calibration is to find a decent approximation to the mean and variance models of the observed data under retainment of the fundamental mean and variance model parameters.
The two major aspects, involved there, are standard regression calibration and the approximation of the observed moments using Taylor series expansion. In the flexible RVM regression, where each covariate is expanded in a set of radial basis functions, a more appropriate correction method, compared to standard regression calibration, is basis function calibration, discussed in detail in Section 3.1.1.
Now, combining basis function calibration and Taylor series expansion of the observed moments, yields a new method, which is termed expanded basis function calibration and has, to the author's knowledge, not even yet attempted in the flexible regression context.

The ideal RVM mean and variance model from (4.3) and (4.2), based on the true but latent covariate ξ, is recasted here for notational reasons as

$$\begin{aligned} \mathbb{E}(Y|\xi) &= f(\Phi(\xi), \boldsymbol{\omega}) && (4.8) \\ \mathbb{V}(Y|\xi) &= \sigma^2 g^2(\Phi(\xi), \boldsymbol{\omega}, \theta) = \sigma^2 f(\Phi(\xi), \boldsymbol{\omega})(1 - f(\Phi(\xi), \boldsymbol{\omega})). && (4.9) \end{aligned}$$

Here, $f(\Phi(\xi), \boldsymbol{\omega}) : \mathbb{R}^{(J+1)} \to \mathbb{R}$ and $g^2(\Phi(\xi), \boldsymbol{\omega}, \theta) : \mathbb{R}^{(J+1)} \to \mathbb{R}$ are now viewed as working on the domain of the individual basis functions, i.e. mapping the row vector $\Phi(\xi)$ containing the values of all basis functions at position ξ to a scalar. Here, the fundamental mean model parameters are again the weights $\boldsymbol{\omega}$. However, the variance function (4.9) also includes the parameters $\boldsymbol{\omega}$, a dispersion parameter σ^2 and possibly nuisance parameters collected in θ. The fact that $\boldsymbol{\omega}$ is a random parameter vector is again notationally ignored. A correct notation would demand conditioning on $\boldsymbol{\omega}$ in (4.8)-(4.9),

but this will be suppressed in the following for the sake of clarity.

Firstly, it is important to understand that the idea of expanded regression calibration is in general not restricted to the simple approximation utilizing $\mu_{\xi|X}$ in place of X, as in the standard regression calibration. Since the basis function calibration utilizing $\mathbb{E}(\Phi(\xi)|X)$ in place of $\Phi(X)$ turned out to be a clever strategy in the Gaussian regression case, it may be fruitful to apply the basic idea of expanded regression calibration to basis function calibration. The necessary amendments for this new 'expanded basis function calibration' approach are described in the following.

In standard regression calibration, the approximate models for the observed mean and variance function are based on the conditional mean of ξ given X (X may here again denote the average over available replicates)

$$\mathbb{E}(\xi|X) \;=\; \mu_{\xi|X}.$$

However, the basis function calibration (cf. Section 3.1.1) successfully extended this for the RVM and applied the conditional mean of the row vector $\Phi(\xi)$ given X

$$\mathbb{E}(\Phi(\xi)|X) \;=\; \mu_{\Phi(\xi)|X}. \tag{4.10}$$

Calculating $\mu_{\Phi(\xi)|X}$ requires here distributional assumptions over ξ. Via the specification of $\sigma_{\xi|X}$, the calibrated row vector $\mu_{\Phi(\xi)|X}$ may account for replicate measurements and heteroscedasticity in the measurement process. The specific figure of (4.10) depends of course on the type of error model - classical or Berkson.
Basis function calibration uses $\mu_{\Phi(\xi)|X}$ instead of the error-prone $\Phi(X)$, yielding the following approximations to the observed models

$$\mathbb{E}(Y|X) \;\approx\; f(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) \tag{4.11}$$

$$\mathbb{V}(Y|X) \;\approx\; \sigma^2 g^2(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}, \theta). \tag{4.12}$$

Here, it is again assumed that, under small measurement error, $\Phi(\xi)$ will be close to its expectation $\mu_{\Phi(\xi)|X}$ whereas $\Phi(\xi)$ may not be close to $\Phi(X)$. So,

naively replacing $\Phi(\xi)$ by $\Phi(X)$ may introduce a large bias into the analysis and hence the need for calibration.
More formally, the true vector $\Phi(\xi)$ is decomposable into the vector $\mu_{\Phi(\xi)|X}$ containing the calibrated basis functions and a random vector V

$$\begin{aligned} \Phi(\xi) &= \mu_{\Phi(\xi)|X} + V \\ &\quad \text{where } \mathbb{E}(V|X) = 0, \mathbb{V}(V|X) = \Sigma_{\Phi(\xi)|X}. \end{aligned} \tag{4.13}$$

The assumption of $\Sigma_{\Phi(\xi)|X}$ being small justifies the approximations (4.11) and (4.12). The following insertion defines the notational shortcuts, which are used throughout this section.

Insertion: Notational details

The basis function calibration is not least because of the notational details demanding. The following shortcuts are defined to make the formulas involved in this approach more lucid:

$$\begin{aligned} f &:= f(\Phi(\xi), \boldsymbol{\omega}) = G(\Phi(\xi)\boldsymbol{\omega}) \\ f_\mu &:= f(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) = G(\mu_{\Phi(\xi)|X}\boldsymbol{\omega}) \\ f' &:= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})}{\partial \Phi(\xi)} \\ f'_\mu &:= f'(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) \\ f'' &:= \frac{\partial^2 G(\Phi(\xi)\boldsymbol{\omega})}{\partial \Phi(\xi)^{\mathrm{T}} \Phi(\xi)} \\ f''_\mu &:= f''(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) \\ g &:= g^2(\Phi(\xi), \boldsymbol{\omega}, \theta) = G(\Phi(\xi)\boldsymbol{\omega})(1 - G(\Phi(\xi)\boldsymbol{\omega})) \\ g_\mu &:= g^2(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}, \theta) = G(\mu_{\Phi(\xi)|X}\boldsymbol{\omega})(1 - G(\mu_{\Phi(\xi)|X}\boldsymbol{\omega})) \\ g' &:= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})(1 - G(\Phi(\xi)\boldsymbol{\omega}))}{\partial \Phi(\xi)} \\ g'_\mu &:= g'(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) \\ g'' &:= \frac{\partial^2 G(\Phi(\xi)\boldsymbol{\omega})(1 - G(\Phi(\xi)\boldsymbol{\omega}))}{\partial \Phi(\xi)^{\mathrm{T}} \Phi(\xi)} \\ g''_\mu &:= g''(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}), \end{aligned}$$

with all first derivatives being row vectors and all second derivatives being matrices. Recall that in binary regression the response function $G(z)$ is either the inverse probit or inverse logit function, i.e. $G(z) = \int_{\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$ or $G(z) = (1 + \exp(-z))^{-1}$. The general form of these derivatives and the specific form for probit and logistic regression will be given later.

Expanded basis function calibration now refines the approximation for the observed mean (4.11) and the observed variance (4.12).
This is achieved by utilizing second order Taylor series expansion of the mean model f and of the variance model g^2 around the calibrated row vector $\mu_{\Phi(\xi)|X}$ from (4.10). Here it is displayed only for the mean model f:

$$f \approx f_\mu + f'_\mu V^{\mathrm{T}} + \frac{1}{2} V f''_\mu V^{\mathrm{T}}. \tag{4.14}$$

For the variance model g^2 this is in complete analogy to (4.14) with g substituting f where it occurs. Under the assumption of $\Sigma_{\Phi(\xi)|X}$ being small and thus the deviation V (cf. (4.13)) being small with high probability the second order Taylor series expansion should work quite well.
Finally, the expansion (4.14) is used to derive an approximation for the mean function of the observed data under application of the theorem of iterated expectations

$$\begin{aligned} \mathbb{E}(Y|X) &= \mathbb{E}(\mathbb{E}(Y|X,\xi)|X) \\ &= \mathbb{E}(f|X) \\ &\approx \mathbb{E}\left\{ \left(f_\mu + f'_\mu V^{\mathrm{T}} + \frac{1}{2} V f''_\mu V^{\mathrm{T}} \right) |X \right\} \\ &= f_\mu + \frac{1}{2} \mathrm{tr}\left(\Sigma_{\Phi(\xi)|X} f''_\mu \right), \end{aligned} \tag{4.15}$$

where 'tr' denotes the trace function, here applied to the product of matrices $\Sigma_{\Phi(\xi)|X}$ and f''_μ. As a consequence of $\mathbb{E}(V|X) = 0$ the first derivative f'_μ is irrelevant in the representation (4.15).
Furthermore, a refined approximation of the observed variance is now found

by utilizing the variance decomposition formula

$$\begin{aligned}\mathbb{V}(Y|X) &= \mathbb{V}(\mathbb{E}(Y|\xi, X)|X) + \mathbb{E}(\mathbb{V}(Y|\xi, X)|X) \\ &= \mathbb{V}(\mathbb{E}(Y|\xi)|X) + \mathbb{E}(\mathbb{V}(Y|\xi)|X). \end{aligned} \tag{4.16}$$

The first term on the right hand side in (4.16) is approximated by using the Taylor expansion (4.14) and additionally assuming that $\mathbb{V}(V^2|X) \approx 0$:

$$\begin{aligned}\mathbb{V}(\mathbb{E}(Y|\xi)|X) &= \mathbb{V}(f|X) \\ &\approx \mathbb{V}(f'_\mu V^{\mathrm{T}}|X) \\ &= f'^{\mathrm{T}}_\mu \Sigma_{\Phi(\xi)|X} f'_\mu.\end{aligned}$$

This expression represents variability in Y due to measurement error. The presence of $\Sigma_{\Phi(\xi)|X}$ makes calculation of this expression computationally heavy, since every individual observation in the data generates a matrix, i.e. the computation of N matrices $\Sigma_{\Phi(\xi)|x_i}, i = 1, \ldots, N$ is required. The formula for computing $\Sigma_{\Phi(\xi)|x_i}$ and an elegant way to control the required time and space resources is presented in the structural likelihood approach, which also requires these matrices (cf. (3.18) in chapter 3).
The second term on the right hand side of (4.16) is

$$\begin{aligned}\mathbb{E}(\mathbb{V}(Y|\xi)|X) &= \mathbb{E}(g|X) \\ &\approx \sigma^2 g_\mu + \sigma^2 \frac{1}{2} \mathrm{tr}\left(\Sigma_{\Phi(\xi)|X} g''_\mu\right),\end{aligned}$$

where Taylor expansion of g around $\mu_{\Phi(\xi)|X}$ (cf. (4.14)) is used to approximate the variance function. Now, putting the pieces together yields the approximation of the observed variance

$$\mathbb{V}(Y|X) \approx f'^{\mathrm{T}}_\mu \Sigma_{\Phi(\xi)|X} f'_\mu + \sigma^2 g_\mu + \sigma^2 \frac{1}{2} \mathrm{tr}(\Sigma_{\Phi(\xi)|X} g''_\mu). \tag{4.17}$$

Most appealing, these approximations (4.15) and (4.17) can be used for general response functions $G(z)$. The first and second derivatives of f and g for general response functions are given in the following.

The first derivative f' is

$$\begin{aligned} f' &= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})}{\partial \Phi(\xi)} \\ &= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})}{\partial \left(\Phi(\xi)\boldsymbol{\omega}\right)} \frac{\partial \Phi(\xi)\boldsymbol{\omega}}{\partial \Phi(\xi)} \\ &= G'(\Phi(\xi)\boldsymbol{\omega})\boldsymbol{\omega}^{\mathrm{T}}, \end{aligned} \tag{4.18}$$

where $G'(\Phi(\xi)\boldsymbol{\omega})$ is scalar and denotes the first derivative of the response function with respect to the linear predictor $\Phi(\xi)\boldsymbol{\omega}$.
Further, the second derivative f'' is

$$\begin{aligned} f'' &= \frac{\partial G'(\Phi(\xi)\boldsymbol{\omega})\boldsymbol{\omega}^{\mathrm{T}}}{\partial \Phi(\xi)^{\mathrm{T}}} \\ &= \frac{\partial G'(\Phi(\xi)\boldsymbol{\omega})}{(\Phi(\xi)\boldsymbol{\omega})} \frac{\partial \Phi(\xi)\boldsymbol{\omega}}{\partial \Phi(\xi)^{\mathrm{T}}} \boldsymbol{\omega}^{\mathrm{T}} \\ &= G''(\Phi(\xi)\boldsymbol{\omega})\boldsymbol{\omega}\boldsymbol{\omega}^{\mathrm{T}}, \end{aligned} \tag{4.19}$$

where $G''(\Phi(\xi)\boldsymbol{\omega})$ is scalar and denotes the second derivative of the response function with respect to the linear predictor $\Phi(\xi)\boldsymbol{\omega}$ and $\boldsymbol{\omega}\boldsymbol{\omega}^{\mathrm{T}}$ is a matrix.

The first and second derivative of the variance function g^2 for general response functions $G(z)$ in binary regression are given as

$$\begin{aligned} g' &= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})(1 - G(\Phi(\xi)\boldsymbol{\omega}))}{\partial \Phi(\xi)} \\ &= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})}{\partial \left(\Phi(\xi)\right)} - \frac{\partial [G(\Phi(\xi)\boldsymbol{\omega})]^2}{\partial \Phi(\xi)} \\ &= \left[G'(\Phi(\xi)\boldsymbol{\omega}) - 2G(\Phi(\xi)\boldsymbol{\omega})G'(\Phi(\xi)\boldsymbol{\omega})\right]\boldsymbol{\omega}^{\mathrm{T}} \end{aligned} \tag{4.20}$$

$$\begin{aligned} g'' &= \frac{\partial (G'(\Phi(\xi)\boldsymbol{\omega}) - 2G(\Phi(\xi)\boldsymbol{\omega})G'(\Phi(\xi)\boldsymbol{\omega}))}{\partial \Phi(\xi)^{\mathrm{T}}} \\ &= [G''(\Phi(\xi)\boldsymbol{\omega})(1 - 2G(\Phi(\xi)\boldsymbol{\omega})) \\ &\quad - 2G'(\Phi(\xi)\boldsymbol{\omega})G'(\Phi(\xi)\boldsymbol{\omega})]\,\boldsymbol{\omega}\boldsymbol{\omega}^{\mathrm{T}}, \end{aligned} \tag{4.21}$$

where g' is a vector and g'' is a matrix.
Here, f'' and g'' are of matrix form for every observation in the data set. Since

both depend also on the weight parameters, recalculation of these matrices has to be performed in every step of the optimization algorithm, making the procedure computationally extraordinarily challenging.

While the previous derivations hold for arbitrary binary regression models, the specific expressions for the popular probit and logit case are given in the following.
For the binary probit regression, where

$$f := G(\Phi(\xi)\boldsymbol{\omega}) = \int_{-\infty}^{\Phi(\xi)\boldsymbol{\omega}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$$

and with ϕ denoting the standard normal density these specific derivatives are

$$\begin{aligned} f' &= \phi\left(\Phi(\xi)\boldsymbol{\omega}\right)\boldsymbol{\omega}^{\mathrm{T}} \\ f'' &= -\boldsymbol{\omega} f' \\ g' &= f' - 2f \cdot f' \\ g'' &= f''(1-2f) - 2f'^{\mathrm{T}} f'. \end{aligned}$$

For the binary logit regression, where

$$f := G(\Phi(\xi)\boldsymbol{\omega}) = \frac{1}{1+\exp(-\Phi(\xi)\boldsymbol{\omega})},$$

the required derivatives are given by

$$\begin{aligned} f' &= f(1-f)\boldsymbol{\omega}^{\mathrm{T}} \\ f'' &= \boldsymbol{\omega}(f' - 2f \cdot f') \\ g' &= f' - 2f \cdot f' \\ g'' &= f''(1-2f) - 2f'^{\mathrm{T}} f'. \end{aligned}$$

Here, f' and subsequently $(f' - 2Gf')$ are row vectors.

One potential problem with the approximations of the observed moments (4.15) and (4.17) is that they might not be range preserving since g'' and f''

need not be positive. A range preserving alternative is presented in Carroll & Stefanski (1990) for expanded regression calibration and for the expanded basis function calibration approach developed here it is derived as

$$\begin{aligned}\mathbb{E}(Y|X) &\approx f\left(\mu_{\Phi(\xi)|X} + \frac{1}{2}\frac{f'_\mu}{\|f'_\mu\|^2}\mathrm{tr}(f''_\mu \Sigma_{\Phi(\xi)|X}), \boldsymbol{\omega}\right)\\ \mathbb{V}(Y|X) &\approx \sigma^2 g^2\left(\mu_{\Phi(\xi)|X} + \frac{1}{2}\frac{g'_\mu}{\|g'_\mu\|^2}\mathrm{tr}(g''_\mu \Sigma_{\Phi(\xi)|X}) + \sigma^{-2} f'^{\mathrm{T}}_\mu \Sigma_{\Phi(\xi)|X} f'_\mu, \boldsymbol{\omega}, \theta\right).\end{aligned}$$

Another (ad-hoc) strategy to cope with this problem, is to exit the current weight optimization in case the Taylor series approximation for $\mathbb{E}(Y|X)$ (4.15) leaves the $[0,1]$-interval and do a hyperparameter update until further $\boldsymbol{\omega}$ updating is performed. Even coming close to the boundaries of admissible values might cause numerical instabilities, which are somewhat aggravated when happening at the fringe of the data, where only sparse data is available. This was, however, not encountered in the simulations presented below. Several strategies placing lower and upper bounds on key values are conceivable to partly cure the problem. In general it is not expected that the measures to stabilize the numerical computations affect the presented method to a great extend.

Once the approximate mean and variance models for the observed data are calculated, model exploration as discussed in Carroll & Ruppert (1988) can be used in addition as guide to the construction of a final model.
However, in this work both approximate observed moments (4.15) and (4.17) are directly used in the penalized quasi score function.
The parameter estimation is then performed in an analog way to the algorithm used in the structural quasi likelihood (cf. Section 3.1.2), where one tries to find the root of the score function

$$s^X(Y, X, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\partial \mathbb{E}^*(y_i|x_i)}{\partial \boldsymbol{\omega}}\left(\frac{y_i - \mathbb{E}^*(y_i|x_i)}{\mathbb{V}^*(y_i|x_i)}\right) - \boldsymbol{\omega} A. \tag{4.22}$$

The symbol $*$ is chosen to indicate the approximative nature of these moments. Applying basic algebra, the required differentiation $\mu^*_{\Phi(\xi)|x_i} := \frac{\delta \mathbb{E}^*(y_i|x_i)}{\delta \boldsymbol{\omega}}$

in (4.22) is easily calculated as

$$\mu^*_{\Phi(\xi)|x_i} = \mu_{\Phi(\xi)|x_i} D_{1i} + \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}} \Sigma_{\Phi(\xi)|x_i} \boldsymbol{\omega} D_{3i} + \boldsymbol{\omega}^{\mathrm{T}} \Sigma_{\Phi(\xi)|x_i} D_{2i}, \tag{4.23}$$

with

$$\begin{aligned} D_{1i} &= \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})}{\partial(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})} \right) \\ D_{2i} &= \left(\frac{\partial^2 G(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})}{\partial(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})} \right) \\ D_{3i} &= \left(\frac{\partial^3 G(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})}{\partial(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})} \right). \end{aligned}$$

The score function (4.22) is then readily given.
The expected Fisher matrix is derived via the first derivative of the score function (4.22) with respect to $\boldsymbol{\omega}$ and here given as

$$F(\boldsymbol{\omega}) = \left(\Phi_c^{*\mathrm{T}} B_c^* \Phi_c^* + A \right). \tag{4.24}$$

For the sake of clarity, two quantities are introduced and defined here: differentiation of the observed mean model with respect to the weights yields the matrix Φ_c^*, constructed from the row vectors (4.23), and the observed variances for each individual are collected in the diagonal matrix B_c^*:

$$\Phi_c^* = \begin{pmatrix} \mu^*_{\Phi(\xi)|x_1} \\ \mu^*_{\Phi(\xi)|x_2} \\ \cdots \\ \mu^*_{\Phi(\xi)|x_N} \end{pmatrix}, \quad B_c^* = \begin{pmatrix} \mathbb{V}^*(y_1|x_1)^{-1} & & \\ & \ddots & \\ & & \mathbb{V}^*(y_N|x_N)^{-1}. \end{pmatrix}. \tag{4.25}$$

The subscript 'c' is chosen to indicate that both quantities are affected by calibration, while the superscript $*$ again denotes that the approximate moments (4.15) and (4.17) are involved here.
The score function (4.22) with corrected mean and variance functions can no longer be solved analytically. Again, Fisher scoring including the modified score function (4.22) and the expected Fisher matrix (4.24) has to be applied,

analogously to the non-Gaussian regression case in Section 2.1.2.
Approximating the posterior covariance matrix by the inverse expected Fisher matrix yields the posterior moments of the parameters $\boldsymbol{\omega}$

$$\Sigma = F(\boldsymbol{\omega})^{-1} = \left(\Phi_c^{*\mathrm{T}} B_c^* \Phi_c^* + A\right)^{-1}, \tag{4.26}$$

$$\boldsymbol{\mu} = \Sigma \Phi_c^{*\mathrm{T}} B_c^* \mathbf{y}^*, \quad \text{where} \quad \mathbf{y}^* = \Phi_c^* \boldsymbol{\omega} + (\mathbf{y} - \mathbb{E}^*(\mathbf{y}|\mathbf{x})). \tag{4.27}$$

Here, $\mathbb{E}^*(\mathbf{y}|\mathbf{x})$ in the expression for the working observations $\mathbf{y}^*$ is a column vector based on the approximation in (4.15) with $\mathbf{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ and $\mathbf{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$.
It is stressed here, that the standard errors based on (4.26) are not properly corrected for the inherent covariate measurement error, which is a open problem, yet.

Now, approximating the marginal likelihood for estimation of the hyperparameters is a creativity challenge. To understand this, the construction of the marginal likelihood in the basis function calibration from the previous Section 4.1.1 is recalled here. There, the specific figure of the working observations was given as

$$\mathbf{y}^* = \Phi_c \boldsymbol{\omega} + D^{-1} \left(\mathbf{y} - G(\Phi_c \boldsymbol{\omega})\right),$$

involving the diagonal matrix D consisting of elements $D_{ii} = \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})}{\partial (\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})}\right)$.
Tipping (2001) follows the so-called Laplace approximation (cf. e.g. Tierney & Kadane (1986), MacKay (2003)) and assumes the marginal likelihood of these working observations being approximately normal

$$p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\alpha}) = \mathcal{N}(0, B + \Phi_c A^{-1} \Phi_c^{\mathrm{T}}).$$

Then the hyperparameter estimation proceeded via the one step marginal likelihood maximization scheme as discussed in Section 2.1.2. In contrast to that previous specification, the working observations in the expanded calibration case (4.27) are defined as

$$\mathbf{y}^* = \Phi_c^* \boldsymbol{\omega} + (\mathbf{y} - \mathbb{E}^*(\mathbf{y}|\mathbf{x})).$$

They do no longer include the linear predictor $\Phi_c\boldsymbol{\omega}$ in terms of the calibrated design matrix Φ_c, but instead they contain Φ_c^* from (4.25), which is the derivative of the observed mean model with respect to the weights.
The reason for this lies in the fact that the matrix Φ_c^* is not decomposable into a product of the calibrated design matrix Φ_c and a second factor, which can be captured in the diagonal covariance matrix B_c^* as it has been done for basis function calibration (cf. (4.5)). A fact that makes the optimization of the hyperparameters slightly obscure. Two conceivable approaches for hyperparameter estimation via the marginal likelihood are developed in the subsequent section:

Estimating the hyperparameters

Once the posterior moments (4.26) and (4.27) are computed, a strategy for estimation of the hyperparameters $\boldsymbol{\alpha}$ is needed. This should again proceed via the marginal likelihood optimization, as described in Section 2.1.2 for the non-Gaussian case. Therefore, Tipping (2001) assumes the marginal likelihood being approximately normal, following the Laplace approximation. In this respect, a first possible specification of the marginal likelihood is here given as

$$\begin{aligned} p(\mathbf{y}^*|\mathbf{x},\boldsymbol{\alpha}) \quad &= \quad \mathcal{N}(0, C^*) \\ &\text{where} \quad C^* = B_c^* + \Phi_c^* A^{-1} \Phi_c^{*\mathrm{T}}, \end{aligned} \tag{4.28}$$

with Φ_c^* and B_c^* defined as in (4.25) and the working observations $\mathbf{y}^*$ from (4.27). Since the computational burden of calculating the posterior moments is steeply rising with the number of basis functions in the actual model, the one step optimization of the marginal likelihood by Tipping & Faul (2002), as described in Section 2.1.2, is generally favored in this work. It allows to build up the final model by introducing, updating and deleting basis functions in each step of the algorithm. Usually, there are at no time more than a minor fraction of all possible basis in the model, and thus the required computations are reasonable.

Revisiting this updating scheme, however, reveals a pronounced problem when the marginal likelihood is taken to be as suggested in (4.28). The updating rule is then given by

$$\alpha_j = \begin{cases} \frac{s_j^2}{q_j^2 - s_j} & \text{if } q_j^2 - s_j > 0 \\ \infty & \text{else} \end{cases}, \tag{4.29}$$

where for simplicity $q_j = \phi_j^{*\mathrm{T}} C_{-j}^{*-1} \mathbf{y}^*$ and $s_j = \phi_j^{*\mathrm{T}} C_{-j}^{*-1} \phi_j^*$ have been defined. Here, $C_{-j}^* = C^* - \alpha_j \phi_j^* \phi_j^{*\mathrm{T}}$ denotes the covariance matrix in (4.28) with the influence of basis vector ϕ_j^* removed and ϕ_j^* being the jth basis vector from the design matrix Φ_c^* from the expanded calibration framework (cf. (4.25)). Now, to provide updated values for all $\alpha_j, j = 0, \ldots, J$, the computation of the 'full' design matrix including all potential basis functions is necessary at each step of the algorithm. The required $(N \times J + 1)$ design matrix Φ_c^* on its part, however, involves the computation of the $(J+1) \times (J+1)$ matrices $\Sigma_{\Phi(\xi)|x_i}, i = 1, \ldots, N$ (cf. (4.23)). Theoretically, only that part of the covariance matrix $\Sigma_{\Phi(\xi)|x_i}$ corresponding to those weights currently in the model needs to be calculated. However, it is excessively expensive from a computational point of view to recalculate the 'full' Φ_c^* at each step, whenever the parameters $\boldsymbol{\omega}$ have changed! A potential remedy lies in providing updated values merely for a subset of basis, which means reducing the computational costs by computing only a few ϕ_j^*'s and their respective hyperparameter update using rule (4.29).

An alternative approach:

In order to keep the hyperparameter update computationally operable it is assumed that $\Phi_c^* = \frac{\delta \mathbb{E}^*(\mathbf{y}|\mathbf{x})}{\delta \boldsymbol{\omega}}$ is decomposable into a product of the calibrated design matrix Φ_c and a second factor, the diagonal matrix D with elements

$$D_{ii} = \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})}{\partial (\mu_{\Phi(\xi)|x_i} \boldsymbol{\omega})} \right).$$

Subsequently, the posterior moments of the weights simplify to

$$\Sigma = \left(\Phi_c^{\mathrm{T}} B_c^{**} \Phi_c + A \right)^{-1}, \tag{4.30}$$

$$\boldsymbol{\mu} = \Sigma \Phi_c^{\mathrm{T}} B_c^{**} \mathbf{y}^{**}, \quad \text{where} \quad \mathbf{y}^{**} = \Phi_c \boldsymbol{\omega} + D^{-1}(\mathbf{y} - \mathbb{E}^*(\mathbf{y}|\mathbf{x})). \tag{4.31}$$

The redefined working observations $\mathbf{y}^{**}$ are very similar to those specified earlier in the basis function calibration (cf. (4.6) in Section 4.1.1) and only differ from those with respect to $\mathbb{E}^*(y_i|x_i)$ now incorporating the Taylor series expansion (cf. (4.15)).
The matrix D has been factored out in $\mathbf{y}^{**}$ (4.31) and has been collected, together with matrix B_c^* from (4.25), into the new diagonal matrix B_c^{**} with diagonal elements given by

$$B_{c_{ii}}^{**} = \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})}{\partial(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})} \right)^2 / \, \mathbb{V}^*(y_i|x_i).$$

Matrix B_c^{**} now involves the first derivative of the response function with respect to the linear predictor and $\mathbb{V}^*(y_i|x_i)$, the approximated observed variance of the responses. Here, B_c^{**} only differs from B in the basis function calibration (cf. Section 4.1.1) with respect to $\mathbb{V}^*(y_i|x_i)$ now incorporating the Taylor series expansion (cf. 4.17).
The marginal log likelihood of the working observations $\mathbf{y}^{**}$, based on the decomposability of Φ_c^*, is then

$$\begin{aligned} \mathcal{L} \quad &= \quad -\frac{1}{2}\left[\log \mid C^{**} \mid + {\mathbf{y}^{**}}^{\mathrm{T}} C^{**-1} \mathbf{y}^{**}\right] \\ &\text{where} \quad C^{**} = B_c^{**} + \Phi_c A^{-1} \Phi_c^{\mathrm{T}}. \end{aligned} \tag{4.32}$$

Maximization of this objective function yields the hyperparameter estimates. The one step maximization scheme from Section 2.1.2 therefore utilizes the covariance matrix from (4.32) and the working vector $\mathbf{y}^{**}$ as defined in (4.31). There is no need for computationally intensive recalculation of the design matrix Φ_c^* in each iteration step, since only the calibrated design matrix Φ_c is required here in order to compute the quantities s_j and q_j in (4.29). Φ_c is readily given and does not change during the algorithm. That approach is used in the simulations presented later in this section.
In case there is a dispersion parameter included in the model, σ^2 can also be estimated via this objective function (4.32), cf. (2.28) in Section 2.1.2.
Note, that the modified posterior moments (4.30), (4.31) and the respective working observations are only used here in the optimization scheme for the hyperparameters $\boldsymbol{\alpha}$ – **not** in the optimization of the weights $\boldsymbol{\omega}$.

4.1.3 SIMEX

The fundamental concept of SIMEX in the flexible regression setting lies in learning how the measurement error affects the prediction function and correction of the error based on this acquired knowledge. An indispensable requirement is the valid assumption of the classical measurement error model (possibly after transformation) $X = \xi + \delta$, where δ is independent of Y and ξ and has mean zero and variance σ_δ^2.
A very minor amendment of the Gaussian regression SIMEX, as presented in Section 3.1.3 in order to satisfy the binary case lies in studying the error effect on the linear predictor $\widehat{f}^*(\xi_k) := \Phi(\xi_k)\widehat{\boldsymbol{\omega}}$ at points of interest ξ_k instead of investigating its effect on the estimated probability $\widehat{f}(\xi_k) := G\left(\Phi(\xi_k)\widehat{\boldsymbol{\omega}}\right)$ directly. This modification guarantees the final SIMEX prediction function $\widehat{f}_{SIMEX}(\xi_k) = G\left(\widehat{f^*}_{SIMEX}(\xi_k)\right)$ to be in the scope of $[0, 1]$ as postulated for binary regression. The recipe for SIMEX is then

1a) Generate random errors $\delta_i^* \sim \mathcal{N}\left(0, \frac{\sigma_{\delta^*}^2}{m}\right)$ and add these to the observed $x_i, i = 1, \ldots, x_N$, where x_i may be the mean of m replicate measurements $x_{i1}, x_{i2} \ldots, x_{im}$. For simplicity, it is again assumed that the number of replicates are identical for all objects in the sample.

1b) Then perform a standard RVM analysis using these 'new' data containing the additional error.

Repeat these steps sufficiently often to obtain a series of estimates for the linear predictor $\widehat{f^*}_1(\xi_k)$, $\widehat{f^*}_2(\xi_k), \ldots, \widehat{f^*}_B(\xi_k), B = 50 - 200$ at ξ_k.

2) Compute the average over the B predictions $\widehat{f}^*(\xi_k) = \sum_{s=1}^{B} \widehat{f^*}_s(\xi_k)$ for every ξ_k.

Now, repeat this whole scheme for a set of error variances $\sigma_{\delta^*}^2 = c \cdot \sigma_\delta^2$, in multiples of the original error variance. Typically, the multiplier is chosen to be $c = c_1, c_2, c_3, c_4, \ldots$ with $c_1 = 0, c_2 = 0.5, c_3 = 1, c_4 = 1.5$. Here, $c_1 = 0$ corresponds to the analysis based on the originally observed data, which, of course, has to be performed only once, not B times.
This yields a series of mean estimates $\widehat{f^*}_{c_1}(\xi_k), \widehat{f^*}_{c_2}(\xi_k), \ldots$ depending on the

multiple c of the error variance. Plotting these mean estimates versus their respective inherent measurement error variance reveals a pattern of how the error affects the estimate for the linear predictor. Fitting a line to the error contaminated estimates $\widehat{f^*}_c(\xi_k)$ and extrapolating to $c = -1$ gives the desired SIMEX corrected linear predictor $\widehat{f^*}_{SIMEX}(\xi_k)$. The SIMEX estimate for the true probability is then simply $\widehat{f}_{SIMEX}(\xi_k) = G\left(\widehat{f^*}_{SIMEX}(\xi_k)\right)$.

The pronounced crux of the method is the extrapolation part as this is always an adventurous task. A popular choice is the quadratic extrapolant, which sufficiently models the error pattern in many practical cases. The true error variance σ^2_δ must either be known or estimated from replication/validation data. In the following simulation study a quadratic extrapolation based on the naive analysis (c=0) and the mean estimates over $B = 50$ repetitions for each $c \in \{0.5, 1, 1.5, 2\}$ is used to attain the SIMEX estimates. The computational heaviness of SIMEX is even increased here by the Fisher scoring algorithm inherent in the standard RVM for binary regression (cf. Section 2.1.2).

4.1.4 MCMC error correction in flexible binary regression

The Markov Chain Monte Carlo (MCMC) version of the RVM (cf. Section 2.2.2) will be enriched in order to account for covariate measurement error. Thus, it circumvents the approximations that are indispensable in the former methods described in this section.
This approach is mainly inspired by three (different) aspects discussed in the existent literature: Chakraborty et al. (2005) develop a MCMC version of the RVM for problems where the sample size is substantially smaller as the number of available covariates, known as large p small n problems. However, they perform no selection of basis functions in their approach. This gap can be filled by adopting Bayesian averaging allowing for model selection in each step of the MCMC algorithm and finally averaging over all visited models (cf.

Denison et al. (2002) and the respective paragraph in Section 2.2.1). Berry et al. (2002) most notably introduce error correction for P-splines, where the latent observations $\xi_i, i = 1, \ldots, N$ are introduced as further unknown parameters into the MCMC scheme following the concept of data augmentation (cf. the respective paragraph in Section 2.2.1). This core idea of filling in the latent ξ_i's goes back to Richardson & Gilks (1993a) and Richardson & Gilks (1993b). A detailed overview of the field of Bayesian measurement error correction is given by Richardson (1996).

An outstanding property of MCMC approaches is that these techniques resemble a construction kit, where the individual building blocks remain practically unchanged regardless of the 'monument' one aims to construct. These building blocks are the unknown model parameters and their respective sampling schemes. The monument one hopes to complete is to model the data adequately and to estimate the unknown parameters. More complex models naturally require more building blocks. Typically more building blocks make the construction more susceptible to collapse and consequently the number of parameters to set up an adequate model should be limited. An exception is the introduction of parameters for data augmentation reasons. This technique is intended to allow for the use of more manageable building blocks, e.g. the use of Gibbs-sampling instead of difficult Metropolis Hastings (MH) sampling, to realize the aspired model estimation.
This construction kit character is a tremendous advantage when it comes to writing about these strategies since the flexible Bayesian probit regression model (cf. Section 2.2.2) and the Bayesian correction for measurement error (cf. Section 2.3.2) have been discussed at full length earlier and now the remaining task is to put the modules together.
Firstly, the flexible Bayesian probit model is recalled, together with model selection and then the specification of the measurement error and covariate model are given. Finally the complete 'ready to implement' sampling scheme is presented.

Model setup and prior specifications

The flexible Bayesian probit regression model adopts a data augmentation step in order to make the sampling scheme more handy by essentially imitating the Bayesian Gaussian regression model. Therefore the latent (dependent) variable Z is artificially introduced in the model - a variable that is in an economic context often termed as latent utility. The Bayesian measurement error correction also adopts a data augmentation step, which treats the true but unobservable $\xi_i, i = 1, \ldots, N$ as additionally unknown parameters. Thus, and most conveniently, the flexible Bayesian probit model under Bayesian error correction may now be formulated directly in terms of the latent ξ_i's.
Then the Bayesian probit regression model accounting for covariate measurement error reads as

$$\begin{aligned} y_i &= \begin{cases} 1 & : \text{ if } z_i > 0 \\ 0 & : \text{ otherwise} \end{cases} \\ z_i &= \Phi_{\boldsymbol{\gamma}}(\xi_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}} + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, 1). \end{aligned} \tag{4.33}$$

While the original sample consists of binary random outcomes $y_i \in \{0, 1\}$, these y_i's are no longer random given the z_i's. Here, the linear predictor $\Phi_{\boldsymbol{\gamma}}(\xi_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}}$ is made up of the weighted basis functions, as in the previous RVM methods, but now – most remarkably – constructed from the latent covariate observations $\xi_i, i = 1, \ldots, N$. The error-prone covariate observations $x_i, i = 1, \ldots, N$ are now "hidden" behind the newly introduced unknown parameters $\xi_i, i = 1, \ldots, N$. So, given values for the ξ_i's this is a standard, however flexible, Bayesian probit regression model as promoted by e.g. Albert & Chib (1993) and Holmes & Held (2006). The parameter $\boldsymbol{\gamma}$ indicates, as explained earlier, the model complexity.

In analogy to the RVM methods a multivariate Gaussian prior distribution is defined over the weights. However, here with the slight modification of now selecting a single fixed prior variance v in advance, which should be large

enough to give reasonable probability even to relatively large weight values

$$p(\boldsymbol{\omega}|v) = \prod_{j=0}^{J} \sqrt{\frac{1}{2\pi v}} \exp\left(-\frac{\omega_j^2}{2v}\right). \tag{4.34}$$

While all presented RVM methods select relevant basis functions from a large set of $(J+1)$ potential basis functions (usually $(J+1) = 101$, i.e. 1 intercept + 100 radial basis functions) this sparsity concern should somehow translate to the MCMC approach. This is particularly important here since the sampling of $\boldsymbol{\omega}$ would otherwise include all $(J+1)$ parameters – a monstrous since computer intensive task.

Two ways to realize sparsity have been described in Section 2.2.3, and finally the Bayesian averaging method, discussed there at length is implemented here: an additional parameter vector $\boldsymbol{\gamma}$ is introduced into the model naming the relevant basis function, i.e. $\boldsymbol{\gamma} = \{1, 3, 89\}$ refers to the model only including the first, third and 89th basis function. Typically the dimension of $\boldsymbol{\gamma}$ is varying and thus the reversible jump algorithm, as allowing for variable parameter dimensions, is employed.

The discrete uniform prior distribution over $\boldsymbol{\gamma}$ is adopted from Denison et al. (2002) that takes

$$p(\boldsymbol{\gamma}) = \binom{J+1}{dim(\boldsymbol{\gamma})}^{-1} \times \frac{1}{T+1}, \tag{4.35}$$

with $dim(\boldsymbol{\gamma})$ denoting the number of elements in $\boldsymbol{\gamma}$, $J+1$, the number of the candidate basis functions, and T the maximum number of basis functions allowed in a model, so $T \leq (J+1)$. This is in concordance with all previous methods, which select basis functions from an arsenal of J radial basis functions and 1 intercept. The maximum number of basis functions T should be chosen large enough not to affect the posterior, which is, however, not checked here, instead this number is set to $T = 20$.

Then, the covariate and the measurement error model need to be characterized. Both are taken from Berry et al. (2002) and are briefly recalled here from Section 2.3.2.

A multivariate normal prior distribution is applied over $\boldsymbol{\xi}$, with its elements

being independently

$$\xi_i \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2), \tag{4.36}$$

with respective normal and inverse Gamma hyperpriors over its moments

$$\mu_\xi \sim \mathcal{N}(f, g^2), \tag{4.37}$$
$$\sigma_\xi^2 \sim IG(A_\xi, B_\xi). \tag{4.38}$$

The inverse Gamma distribution is defined as

$$f(x|A,B) = \frac{1}{\Gamma(A)B^A x^{A+1}} \exp\left(-\frac{1}{Bx}\right) I(0 \leq x < \infty).$$

Assuming normally distributed measurement error, yields the error model

$$x_{ij} = \xi_i + \delta_{ij}, \quad (\delta_{ij}, \xi_i) \sim \text{indep.}, \quad \delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2), \tag{4.39}$$

where $j = 1, \ldots, m_i$ indexes the replicate measurements for person i.
Following Berry et al. (2002), the measurement error variance σ_δ^2 is reparametrized as

$$\sigma_\delta^2 = \frac{1-\lambda}{\lambda} \cdot \sigma_\xi^2, \quad \text{with } \lambda = \frac{\sigma_\xi^2}{\sigma_X^2} \tag{4.40}$$

in order to account for the inherent additive relationship between the variances: $\sigma_x^2 = \sigma_\xi^2 + \sigma_\delta^2$. Thus, σ_δ^2 itself does not appear in the sampling scheme and instead draws for the so-called attenuation factor λ will be generated. Therefore, following again the implementation of Berry et al. (2002) a uniform distribution on the interval $[\lambda_L, \lambda_H]$ is specified here as prior distribution.
In the forthcoming simulations the following prior specifications are used: $v = 100$, $T = 20$, $f = 0$, $g^2 = 100$, $A_\xi = 1$, $B_\xi = 1$, $\lambda_L = 0.7$, $\lambda_H = 0.99$.
The robustness to prior specification will not be investigated here. However, the MCMC-RVM uses the same prior distributions as adopted in Berry et al. (2002) for those parameters appearing in both methods, and Berry et al. (2002) state that their method shows only minimal changes when using different priors.

Inference

Samples of $\boldsymbol{\omega}_{\boldsymbol{\gamma}}, \mu_\xi, \sigma_\xi^2$ and the latent variable $\mathbf{z}$ are obtained from their respective full conditional densities based on the current model as defined by $\boldsymbol{\gamma}$. Most notably, the design matrix $\Phi_{\boldsymbol{\gamma}}$, wherever it appears in this sampling scheme, is constructed from the sampled values $\xi_i, i = 1, \ldots, N$ for the latent covariate ξ.
The full conditionals for $\boldsymbol{\omega}_{\boldsymbol{\gamma}}, \mathbf{z}, \mu_\xi$ and σ_ξ^2 are recognized as standard distributions:

$$\begin{aligned} p(\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\mathbf{z}, \boldsymbol{\xi}) &= \mathcal{N}\left(\mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}, \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}\right) \\ \text{where } & \mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} = \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}\Phi_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z} \\ & \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} = \left(v^{-1}\mathbf{I} + \Phi_{\boldsymbol{\gamma}}^{\mathrm{T}}\Phi_{\boldsymbol{\gamma}}\right)^{-1}, \end{aligned} \tag{4.41}$$

$$\begin{aligned} p(z_i|\mathbf{z}_{-i}, y_i, \boldsymbol{\xi}, \boldsymbol{\gamma}) &\propto \begin{cases} \mathrm{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i > 0) & \text{if } y_i = 1 \\ \mathrm{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i < 0) & \text{otherwise} \end{cases} \\ \text{where } & \mu_{z_i|\cdot} = \Sigma_{z_i|\cdot}\Phi_{\boldsymbol{\gamma}}(\xi_i)\mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} - w_i z_i \\ & \Sigma_{z_i|\cdot} = 1 + w_i \\ & w_i = h_i/(1 - h_i), \end{aligned} \tag{4.42}$$

$$p(\mu_\xi|\boldsymbol{\xi}, \sigma_\xi^2) = \mathcal{N}\left(\frac{\left(\sum_{i=1}^N \xi_i\right) g^2 + f\sigma_\xi^2}{Ng^2 + \sigma_\xi^2}, \frac{\sigma_\xi^2 g^2}{Ng^2 + \sigma_\xi^2}\right)$$

$$\begin{aligned} p(\sigma_\xi^2|\mathbf{x}, \boldsymbol{\xi}, \mu_\xi, \lambda) &= IG\left(A_{\xi|\cdot}, \frac{1}{B_{\xi|\cdot}}\right) \\ \text{where } & A_{\xi|\cdot} = A_\xi + \frac{1}{2}\sum_{i=1}^N m_i + \frac{N}{2}, \\ & B_{\xi|\cdot} = B_\xi^{-1} + \frac{\lambda}{2(1-\lambda)}\sum_{i=1}^N\sum_{j=1}^{m_i}(x_{ij} - \xi_i)^2 \\ & \qquad + \frac{1}{2}\sum_{i=1}^N(\xi_i - \mu_\xi)^2. \end{aligned}$$

Here, h_i is the ith diagonal element of the Bayesian hat matrix, $h_i = H_{ii}, H = \Phi_{\boldsymbol{\gamma}}\Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}}\Phi_{\boldsymbol{\gamma}}^{\mathrm{T}}$. Most interestingly, since model selection is supposed to occur here in each cycle of the sampling scheme, the necessary calculation of $\Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}|\cdot}}$ in (4.41) is usually not computer intensive since typically only few basis functions are contained in the model. The details of how to sample the model dimension parameter $\boldsymbol{\gamma}$ follow now.

As presented in Section 2.2.3, sampling the model selection parameter $\boldsymbol{\gamma}$ is more complex than sampling the previously discussed parameters $\boldsymbol{\omega}_{\boldsymbol{\gamma}}, \mu_\xi, \sigma_\xi^2$ and the latent response observations $\mathbf{z}$. The reversible jump algorithm needs to be adopted here. This is based on the Metropolis-Hastings (MH) sampler allowing for two move types, either in higher or lower model dimension. The current model dimension may be denoted as $t := dim(\boldsymbol{\gamma})$. Then these moves are characterized by:

BIRTH. Proposal of adding a randomly chosen basis (including intercept) from those that are not present in the current model with probability b_t.

DEATH. Proposal of removing a randomly chosen basis (including intercept) from those that are present in the current model with probability d_t.

Then, the proposal probabilities for BIRTH and DEATH are chosen to be $b_t = d_t = 0.5$ for $0 < t < 20$ and $b_0, d_20 = 1$ and $b_20, d_0 = 0$, a choice that allows a maximum number of $T = 20$ basis functions being in the model at the same time. Depending on whether BIRTH or DEATH is proposed, the specific basis to be contained or excluded is randomly selected with uniform probability from the basis functions currently being excluded or contained, respectively.
The acceptance probability of a move from a t basis function model $\boldsymbol{\gamma}$ to a t' basis functions model $\boldsymbol{\gamma}'$, where the prime indicates the 'proposal', is

$$\alpha = \min\left\{1, \frac{p(\mathbf{z}|\boldsymbol{\gamma}', \boldsymbol{\xi})}{p(\mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\xi})} \times R\right\}, \tag{4.43}$$

the ratio of the marginal likelihoods multiplied by $R = d'_t/b_t$ for a BIRTH and $R = b'_t/d_t$ for a DEATH proposal.
The proposed jump from t to t' dimensions is accepted if an uniformly generated random number is smaller than the acceptance probability (4.43). Sparse models are favored here through the use of the specific zero mean normal prior over the weights (4.34) – a fact, which is explained by Ockham's razor (cf. Section 2.2.1).
The fact that the weights have been marginalized (i.e. integrated out) of the marginal likelihood in (4.43), allows for a block update of $\boldsymbol{\gamma}$ together with $\boldsymbol{\omega}_{\boldsymbol{\gamma}}$, which is „extremely important" (cf. Holmes & Held (2006)), as typically, when the basis functions are non-orthogonal, there is strong linear dependence between the weights. The block update proceeds by firstly drawing a new sample $\boldsymbol{\gamma}$ and then generating the respective coefficients $\boldsymbol{\omega}_{\boldsymbol{\gamma}}$ for this model. Since, however, the moments of the full conditional distribution of the weights (4.41) are needed for calculating the acceptance probability, it was decided here, to firstly present the full conditional of the weights and then the sampling scheme for $\boldsymbol{\gamma}$. However, in an computer implementation this order would be vice versa.
Since accounting for measurement error here, the marginal likelihood in (4.43) is now conditioned on the latent ξ_i's, which means using the design matrix Φ based on samples $\xi_i, i = 1, \ldots, N$ and is given by

$$p(\mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\xi}) = (2\pi)^{-\frac{N}{2}} \frac{\mid \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \mid^{\frac{1}{2}}}{\mid v_{\boldsymbol{\gamma}} \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\mathbf{z}^{\mathrm{T}}\mathbf{z} - \mu^{\mathrm{T}}_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \Sigma^{-1}_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}\right)\right),$$

where the exponential term $\mathbf{z}^{\mathrm{T}}\mathbf{z}$ cancels out in the acceptance probability (4.43) since it is independent of the selected model. Here, $v_{\boldsymbol{\gamma}}$ denotes the prior covariance matrix over the weights for a model as defined by $\boldsymbol{\gamma}$.

Samples for the attenuation parameter λ relating σ^2_δ to σ^2_ξ, cf. (4.40), are generated with the aid of a gridded Gibbs estimator. The full conditional of

this parameter is given by

$$\begin{aligned} p(\lambda|\boldsymbol{\xi}, \sigma_\xi^2) &\propto I(\lambda_L < \lambda < \lambda_H) \left(\frac{\lambda}{1-\lambda}\right)^a \exp\left(-\frac{\lambda \cdot b}{2(1-\lambda)\sigma_\xi^2}\right) \\ a &= \sum_i m_i/2 \\ b &= \sum_i m_i(\overline{x}_i - \xi_i)^2 + \widehat{\sigma_\delta^2} \sum_i (m_i - 1), \end{aligned} \tag{4.44}$$

with $\overline{x}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}$ denoting the average over a subjects's replicates. Here, $\widehat{\sigma_\delta^2}$ is the methods of moments estimate for the measurement error variance (cf. (2.52)). In the forthcoming simulations, the set $\lambda \in [\lambda_L, \lambda_H]$ is discretized into 40 different values, then (4.44) is computed for these values, a discrete distribution function is constructed from the results and λ is sampled from this distribution function. This approach has also been used by Berry et al. (2002) and Carroll et al. (2004) in this situation and usually provides good mixing, though it is not strictly correct. Alternatively, one can also implement a MH step based on (4.44).

Finally, a sampling scheme for the latent covariate observations is required. The conditional density of a single true covariate observation ξ_i is recalled from Section 2.3.2 as

$$\begin{aligned} p(\xi_i|z_i, \mathbf{x}_i, \mu_\xi, \sigma_\xi^2, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \lambda) &\propto \exp\left(-\frac{1}{2}(z_i - \Phi_{\boldsymbol{\gamma}}(\xi_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}})^2\right) \\ &\times \exp\left(-\frac{1}{2\frac{1-\lambda}{\lambda} \cdot \sigma_\xi^2} \sum_{j=1}^{m_i} (x_{ij} - \xi_i)^2\right) \\ &\times \exp\left(-\frac{1}{2\sigma_\xi^2}(\xi_i - \mu_\xi)^2\right). \end{aligned} \tag{4.45}$$

The ξ_i's are independent a posteriori and $\mathbf{x}_i := (x_{i1}, \ldots, x_{im_i})$ denotes here the vector of replicates for person i: only the complete series of measurements taken for person i are required in order to compute the full conditional of ξ_i. However, (4.45) is not a standard density and a Metropolis Hastings (MH) step needs to be implemented. Therefore a symmetric random walk proposal

is specified as

$$p\left(\xi_i'\right) = \mathcal{N}\left(\xi_i, \frac{1}{2}\sigma^2_{\xi|x_i}\right), \tag{4.46}$$

with the prime symbol indicating the proposal, so this must not be confused with the prior over ξ_i derived from the covariate model (4.36). This proposal is a Gaussian with the old value ξ_i as its mean and with variance being half the conditional variance of ξ given the observed value x_i. Here $\sigma^2_{\xi|x_i}$ can be calculated from the general formula (2.51), but uses the current samples of σ^2_ξ and λ. A generated proposal ξ_i' is accepted with probability

$$\alpha = \min\left\{1, \frac{p(\xi_i'|z_i, \mathbf{x}_i, \mu_\xi, \sigma^2_\xi, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \lambda)}{p(\xi_i|z_i, \mathbf{x}_i, \mu_\xi, \sigma^2_\xi, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \lambda)}\right\}.$$

The conditional density (4.45) is easily evaluated at the proposed value ξ_i' and the old value ξ_i to perform MH sampling. If the candidate value ξ_i' is rejected, the current value of ξ_i is repeated as the next value of the MCMC sample. So, according to the new samples $\xi_i, i = 1, \ldots, N$ the design matrix Φ has to be recalculated in every step of the MCMC algorithm.

Given starting values for the unknown parameters, the sampling scheme is executed and after a preceding burn-in period of 2000 runs, another 2000 samples are collected for parameter estimation. The choice of using 2000 samples as a burn-in period and another 2000 for inference is inspired by the MCMC implementation of Berry et al. (2002) and no investigation of convergence, as has been advised in Section 2.2.1, is done here.

Finally, predicting the probability of $y^* = 1$ at an previously unseen ξ^* is accomplished in two steps. Firstly, by using the histogram estimator (cf. (2.30)) as an consistent estimator for the posterior mean of the linear predictor $z^* := \Phi(\xi^*)\boldsymbol{\omega}$:

$$\widehat{\mathbb{E}}(z^*|\mathbf{y}) = \frac{1}{2000}\sum_{j=1}^{2000}(\Phi(\xi^*)\boldsymbol{\omega}^{[j]}|\mathbf{y}). \tag{4.47}$$

Here, the implicit integration is done over both, the parameter space $\mathbf{z}$, $\boldsymbol{\omega}$, $\boldsymbol{\xi}$, μ_ξ, σ^2_ξ, λ and the model space $\boldsymbol{\gamma}$.

Then, in a second step, this posterior mean estimate (4.47) is linked to the desired probability by multiplying this by '1.7' and applying the inverse logit function to the result, which gives

$$\widehat{P}(y^* = 1|\xi^*, \mathbf{y}) = \left(1 + \exp\left(1.7 \cdot \widehat{\mathbb{E}}(z^*|\mathbf{y})\right)\right)^{-1}.$$

Multiplying the linear predictor of a probit link model with the factor '1.7', yields approximately the linear predictor of a logit link model. This circuitous way of deriving the final estimator is due to the fact that in a previous setup of the simulation study the estimates of the linear predictor assuming a logit link model have been compared across all methods.
This estimator adopted here is not equivalent to the histogram estimator using the transformed samples directly:

$$\widehat{P}(y^* = 1|\xi^*, \mathbf{y}) = \frac{1}{2000}\sum_{j=1}^{2000}\left(\left(1 + \exp\left(1.7 \cdot \Phi(\xi^*)\boldsymbol{\omega}^{[j]}\right)\right)^{-1}|\mathbf{y}\right).$$

Here, the transformed samples are restricted to the interval $[0, 1]$ and are expected to follow a skew distribution for many ξ^*'s. Then, the mean is a rather poor estimate for the posterior mode. Thus, utilizing (4.47) is favored here.
It may be stressed here again, that a major drawback of the MCMC approach, which is generally inherent in sampling approaches, is the vast computational effort to accomplish the desired prediction.

4.1.5 A byproduct - data augmentation and calibration

Theoretical considerations for the Gaussian case in Section 3.1.1 showed that basis function calibration leads to an exact representation of the observed mean $\mathbb{E}(Y|X)$ when plugging the calibrated row vectors $\mu_{\Phi(\xi)|X}$ into the model (cf. (3.9)). The data augmentation approach described in the previous Section 4.1.4 introduces the latent response variable Z and so converts binary regression into Gaussian regression and thus simplifies the inference for the MCMC approach. There may be a gain in bringing both strategies together...

Most interestingly, when adopting data augmentation in the probit model

$$\begin{aligned} y_i &= \begin{cases} 1 & : \text{ if } z_i > 0 \\ 0 & : \text{ otherwise} \end{cases} \\ z_i &= \Phi(\xi_i)\boldsymbol{\omega} + \epsilon_i \qquad (4.48) \\ \epsilon_i &\sim \mathcal{N}(0,1) \qquad (4.49) \end{aligned}$$

the observed mean function of the artificial responses $z_i, i = 1, \ldots, N$ can be given exactly as

$$\mathbb{E}(z_i|x_i) = \mu_{\Phi(\xi)|x_i}\boldsymbol{\omega}, \qquad (4.50)$$

with the elements in the row vector $\mu_{\Phi(\xi)|X}$ as calculated in (3.8). The parameter vector $\boldsymbol{\omega}$ is again viewed as being fixed as usual in the basis function calibration context (cf. Section 3.1.1). This relation (4.50) has an important meaning: **if** z_i is available, then using basis function calibration in the binary context is equivalent to using it in the Gaussian context. Of course these latent quantities are **not** available and thus the need for coping with the binary responses y_i.
However, conditioning the latent model (4.48) on the error-prone observations $\mathbf{x}$ and additionally on the observed dichotomous responses $\mathbf{y}$, yields an alternative observed model to (4.50), which can be used for inference. This

alternative observed model is, most notably, formulated in the parameters $\boldsymbol{\omega}$ of the ideal model (4.48) and is given by

$$\begin{aligned}\mathbb{E}(z_i|y_i,x_i) &= \mathbb{E}(\Phi(\xi_i)\boldsymbol{\omega}|y_i,x_i)+\mathbb{E}(\epsilon_i|y_i,x_i)\\ &= \mu_{\Phi(\xi)|x_i}\boldsymbol{\omega}+\mathbb{E}\left(-\frac{\phi_{(0,1)}}{\Phi_{(0,1)}(-f_i^*)-y_i}|x_i\right)\\ &= \mu_{\Phi(\xi)|x_i}\boldsymbol{\omega}+\epsilon_i^*, \end{aligned} \tag{4.51}$$

where $f_i^* := \Phi(\xi)\boldsymbol{\omega}$ denotes the linear predictor from the ideal latent model (4.48). Here, $\phi_{(a,b)}$ and $\Phi_{(a,b)}$ denote the Gaussian density and cumulative distribution function with moments a and b. In the second line of (4.51) it is assumed that y_i has no additional information about $\Phi(\xi_i)$ given x_i, i.e. $\mathbb{E}(\Phi(\xi_i)|y_i,x_i) = \mathbb{E}(\Phi(\xi_i)|x_i) =: \mu_{\Phi(\xi)|x_i}$. The expansion of $\mathbb{E}(\epsilon_i|y_i,x_i)$ in the second line comes from the application of the law of iterated expectations and uses the non-differentiability of the measurement error. It involves the usual formula for computing the mean of a truncated normal distribution. For convenience, it is assumed that ϵ_i^* is normally distributed with zero mean and heteroscedastic variance $\sigma_i^{*^2} = \mathbb{V}\left(\mathbb{E}\left(\epsilon_i|y_i,x_i\right)\right)$. Thus, $\sigma_i^{*^2}$ is no longer equal to one as before in the original probit model (4.49).
Then, if an estimator for $\mathbb{E}(z_i|y_i,x_i)$ exists, the 'observed' model (4.51) can be used to find the parameter estimates for $\boldsymbol{\omega}$. This can be accomplished via some least squares estimator like in the Gaussian case but now using the estimates $\widehat{\mathbb{E}}(z_i|y_i,x_i)$ in place of the observed responses y_i.
The residuals ϵ_i^* from (4.51) can not be computed offhand and are here very naively approximated as

$$\widehat{\epsilon_i^*} = -\frac{\phi_{(0,1)}}{\Phi_{(0,1)}(-\widehat{f_i^*})-y_i}, \tag{4.52}$$

where $\widehat{f_i^*} := \mu_{\Phi(\xi)|x_i}\widehat{\boldsymbol{\omega}}$ denotes the linear predictor, however, now using the calibrated row vector $\mu_{\Phi(\xi)|x_i}$ and the current estimate for the weights $\widehat{\boldsymbol{\omega}}$.
The required estimation for $\mathbb{E}(z_i|y_i,x_i)$ is then given by

$$\widehat{\mu}_{z_i|\cdot} := \widehat{\mathbb{E}}(z_i|y_i,x_i) = \widehat{f_i^*} - \frac{\phi_{(0,1)}}{\Phi_{(0,1)}(-\widehat{f_i^*})-y_i}. \tag{4.53}$$

From model (4.51), the following penalized quasi score function can be constructed

$$\sum_{i=1}^{N} \frac{\partial \mathbb{E}\left(\mathbb{E}(z_i|y_i,x_i)\right)}{\partial \boldsymbol{\omega}} \frac{\mathbb{E}(z_i|y_i,x_i) - \mathbb{E}\left(\mathbb{E}(z_i|y_i,x_i)\right)}{\mathbb{V}\left(\mathbb{E}(z_i|y_i,x_i)\right)} - \boldsymbol{\omega} A,$$

where the original responses $y_i, i = 1\ldots,N$ have been replaced by the new responses $\mathbb{E}(z_i|y_i,x_i)$.
The observed mean model derived from (4.51), under the assumption $\mathbb{E}(\epsilon_i^*) = 0$, is given by

$$\mathbb{E}\left(\mathbb{E}(z_i|y_i,x_i)\right) = \mu_{\Phi(\xi)|x_i}\boldsymbol{\omega}. \tag{4.54}$$

The observed variance $\mathbb{V}\left(\mathbb{E}(z_i|y_i,x_i)\right) = \mathbb{V}(\epsilon_i^*)$ is approximated by

$$\mathbb{V}\left(\mathbb{E}(z_i|y_i,x_i)\right) \approx \mathbb{V}\left(\mathbb{E}(z_i|y_i)|x_i\right) = \boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\Phi(\xi)|x_i}\boldsymbol{\omega} + \sigma^{**^2}, \tag{4.55}$$

with $\sigma^{**^2} = \mathbb{V}(z_i|y_i)$. The fact that σ^{**^2} depends on $\Phi(\xi_i)\boldsymbol{\omega}$ and thus is heteroscedastic is ignored here.
The observed moments (4.54) and (4.55) together with the estimated responses (4.53) yield the 'practical penalized quasi score function'

$$\sum_{i=1}^{N} \mu_{\Phi(\xi)|x_i} \frac{\widehat{\mu}_{z_i|\cdot} - \mu_{\Phi(\xi)|x_i}\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\Phi(\xi)|x_i}\boldsymbol{\omega} + \sigma^{**^2}} - \boldsymbol{\omega} A, \tag{4.56}$$

where $\Sigma_{\Phi(\xi)|x_i}$ is the conditional covariance matrix of the basis functions as derived in Section 3.1.2.
The main difference to the quasi score function employed in the true Gaussian context is that the observable responses y_i enter conditionally in the moments and that (4.56) is based on some artificial responses, which have to be estimated and updated in turn with the other unknown parameters.
However, by equating (4.56) to zero one should obtain parameter estimates $\widehat{\boldsymbol{\omega}}$ that possibly benefit from having transferred the problem into a Gaussian context (at least if the estimation for the latent $\mu_{z_i|\cdot}$ is not too bad). But, on the other hand, the estimation can be impaired by the misspecification of the error distribution of ϵ_i^* as zero mean Gaussian random variable and the

approximation of its variance.
Care must be taken since the so estimated model parameters correspond to a probit model. An approximation for the logit model is however easily derived by multiplying $\widehat{\boldsymbol{\omega}}$ with the factor 1.7.
The hyperparameters $\boldsymbol{\alpha}$ and σ^{**^2} are estimated via the optimization scheme as described in Section 3.1.2 for the structural quasi likelihood approach, where now the estimates $\widehat{\mu}_{z_i|\cdot}$ (4.53) replace the former responses y_i.

This 'latent variable'-approach discussed here is a genuine ad-hoc method, which can possibly be refined after further consideration.

4.2 Simulation study

The presented correction methods, basis function calibration, expanded basis function calibration, SIMEX, the MCMC approach and the rather ad-hoc structural quasi likelihood, based on the latent utility, are compared in a simulation study. The MCMC implementation of Bayesian binary regression under measurement error escorting the article by Berry et al. (2002) is chosen as a reference. It is stressed here once more that the robustness of the MCMC approaches to prior specification is not investigated here.
All methods are checked in a variety of data scenarios. Firstly, these data scenarios are described, before all competing methods are contrasted in some essential respects. This section concludes with a discussion of the presented results from the simulation study.

4.2.1 The data

For each data scenario 200 data sets are simulated.
There are always two replicates ($m_i = 2$) available containing classical additive measurement error with $\mu_\delta = 0$. Thus, each surrogate $x_i, i = 1, \ldots, N$ represents the average over these two replicates. All methods described in this chapter use these replicates in order to estimate the measurement error variance σ_δ^2. If no replicates were available, a good intuition or a divine hint would do as well.
The latent $\xi_i, i = 1, \ldots, N$ are generated as independent normal random variables with mean μ_ξ and variance σ_ξ^2 for all, except one data case, where the ξ_i's are sampled from a standardized $\chi^2(4)$ variable. Each data set contains usually $N = 500$ samples, with exception of case 6, where $N = 1000$ samples are available. The level of measurement error variance is different for the data scenarios. As a consequence of having two replicates, the measurement error variance of the surrogates $x_i = \frac{x_{i1}+x_{i2}}{2}$ is only half the error variance that is stated below in the respective cases.
For the purpose of mean squared error calculations, a prediction function is

calculated for each of the methods. Therefore, predictions were computed on a grid of 101 points in the interval $[a, b]$, which is expected to contain most of the distribution for ξ.
The series of simulation includes eight data cases, where the binary responses are generated from the underlying true probability $P(Y = 1|\xi) = \mathbb{E}(Y|\xi) = (1 + \exp(-m(\xi)))^{-1}$ with functional argument $m(\xi)$. No under-/overdispersion is specified here.

Case 1: A quadratic function of the covariate with $m(\xi) = -0.2 + 0.25\xi + 0.1\xi^2$, with $N = 500$, $a = -2.0$, $b = 2.0$, $\sigma_\delta^2 = 0.8^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5$

Case 2: An oscillating function of the covariate with

$$m(\xi) = \begin{cases} -0.9 + 3\sin(5\xi)/(5\xi) & \xi \neq 0 \\ 2.1 & \xi = 0 \end{cases},$$

$N = 500$, $a = -2.0$, $b = 2.0$, $\sigma_\delta^2 = 0.2^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5$

Case 3: Same as case 2 except $\sigma_\delta^2 = 0.5^2$

Case 4: Same as case 2 except $\sigma_\delta^2 = 0.8^2$

Case 5: Another oscillating function of the covariate with

$$m(\xi) = \begin{cases} 0.5 + 2\frac{\sqrt{(0.25\xi+0.5)(1-(0.25\xi+0.5))}\sin(2\pi(1+2^{(9-4j)/5}))}{\frac{\xi}{4}+0.5+2^{(9-4j)/5}} & -2 \leq \xi \leq 2 \\ 0.5 & \text{otherwise} \end{cases}$$

for $j = 3$, with $N = 500$, $a = -2$, $b = 2$, $\sigma_\delta^2 = 0.5^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5^2$

Case 6: Same as case 4 except N=1000

The violation of the assumption that ξ is normally distributed is studied in the following case:

Case 7: The same as case 3 above, except that ξ is a standardized $\chi^2(4)$ random variable. The MSE will be evaluated on $[-1.25, 2.00]$.

A plateau function is difficult to model with the RVM methods using RBF kernels or the MCMC approach using 2nd order truncated power series. This model misspecification is investigated here:

Case 8: The same as case 3 above except that

$$\begin{aligned} m(\xi) &= 1 + 2(-2 + H(100\xi) + H(100(\xi - 0.5))), \\ &\quad \text{where } H(\xi) = (1 + \exp(-\xi))^{-1}. \end{aligned}$$

Figure 4.1 and 4.2 display example data sets for each scenario as well as the true mean function (probability). Despite there are two replicate measurements available and usually the average is taken to perform the model estimation, here only a single measurement is displayed. These figures may give an impression of the detective work the correction methods have to fulfill.

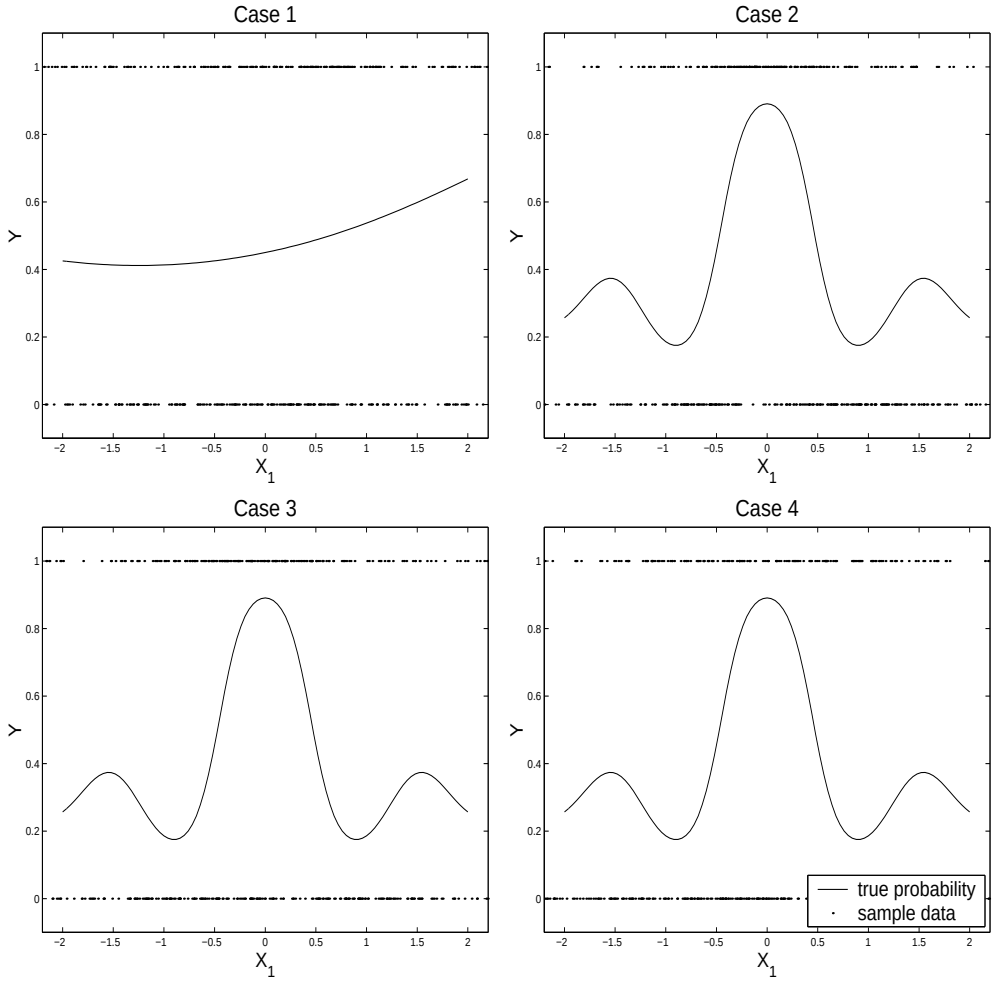

Figure 4.1: Example data sets for cases 1-4 and the respective true underlying probability function. The response is plotted versus one error-prone measurement, i.e. no averaging over the two available replicates is done here. Only the range $[a, b]$ on which the methods will be evaluated is shown.

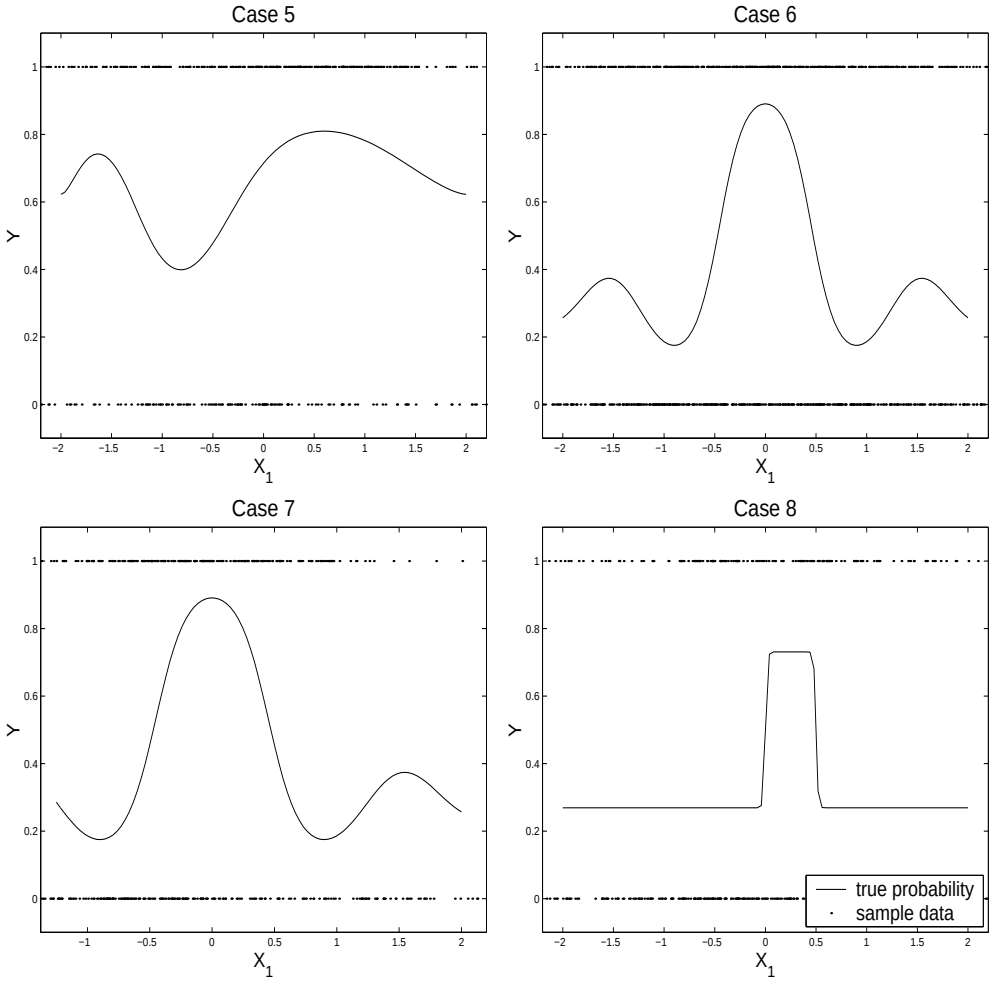

Figure 4.2: Example data sets for cases 5-8 and the respective true underlying probability function. The response is plotted versus one error-prone measurement, i.e. no averaging over the two available replicates is done here. Only the range $[a, b]$ on which the methods will be evaluated is shown.

4.2.2 Specification details of the methods

General settings of the presented methods are briefly described in the following.

Both MCMC techniques and the latent variable byproduct from Section 4.1.5 fit a probit model to the data. All other methods apply a logit model.
Basis calibration, expanded basis function calibration, SIMEX and the MCMC approach work with an arsenal of one intercept and 100 radial basis functions located at the quantiles of the observed data. For SIMEX the knots are again located at the quantiles of the artificially generated observations in each simulation step.
The kernel parameter η is selected from a set of admissible values like in the Gaussian implementations before (cf. Section 3.2.2). Basis calibration, expanded basis function calibration and the MCMC approach simply copy the optimal parameter found by the naive approach.
Basis function calibration and SIMEX use the analytic updating scheme of the precisions $\boldsymbol{\alpha}$ as described for the non-Gaussian case in Section 2.1.2. For the expanded basis function calibration approach it is given in Section 4.1.2.
The 'latent utility'-byproduct uses the updating scheme as presented in the Gaussian case in Section 2.1.2. The MCMC version of the RVM does not use this hyperparameter, but instead the model selection parameter vector $\boldsymbol{\gamma}$ naming relevant basis functions. This is sampled together with the other unknowns in the sampling scheme.

The RVM without measurement error and the naive estimation are taken as reference methods together with an MCMC implementation that came with the article by Berry et al. (2002) and is downloadable from the home page `http://www.stat.tamu.edu/~carroll/matlab_programs/software.php`.
Their implementation differs from the MCMC approach for the RVM developed here in using a set of only 30 truncated second order power series basis and not doing any selection.
Both MCMC methods obtain their respective estimates $\widehat{P}(y^* = 1|\xi^*, \mathbf{y})$ via

the posterior mean estimator for the linear predictor $z^* := \Phi(\xi^*)\boldsymbol{\omega}$, as described at the end of Section 4.1.4. This estimator is not equivalent to the histogram estimator (cf. (2.30)) directly using the transformed samples $P(y^* = 1|\xi^*, \boldsymbol{\omega}^{[j]}, \mathbf{y})$ for consistently estimating the posterior mean of this probability. However, these transformed samples are restricted to the interval $[0, 1]$ and their distribution is expected to be skewed at values ξ^* that are associated with high or low probability. Then, the posterior mean estimate is not representative for the posterior mode at these values. Thus, the histogram estimator based on the samples of the linear predictor is implemented in the MCMC methods. These samples are expected to be symmetrically distributed and thus the posterior mean is more representative for the mode of the posterior distribution.

Table 4.1 contrasts all compared methods in some essential respects.

Table 4.1: Overview of methods and their specifications

Method	Type of basis function	Number of potential/effective basis functions w/o intercept	Knot selection	Error correction
$\mathbf{RVM_{naive}}$	RBF	100/usually very few, see results	quantiles of error-prone data	none
$\mathbf{RVM_{BC}}$ (RVM + basis function calibration)	RBF	100/usually very few, see results	same as $\mathbf{RVM_{naive}}$	Approximation of the observed mean function, by using calibrated basis functions $\mathbb{E}(\Phi(\xi)\|X)$ instead of $\Phi(X)$.
$\mathbf{RVM_{EBC}}$ (RVM + expanded basis function calibration)	RBF	100/usually very few, see results	same as $\mathbf{RVM_{naive}}$	Refining the approximation of observed mean and variance function by using Taylor series. Approximation of observed marginal LH for hyperparameter estimation
$\mathbf{RVM_{SIMEX}}$	RBF	100/usually very few, see results	quantiles of generated SIMEX-observations	The effect of additive error is studied in a simulation study and then corrected
BRS (bayesian regression splines)	2nd order truncated power series	30/30	same as $\mathbf{RVM_{naive}}$	The true ξ_i are regarded as unknown parameters and sampled in an MCMC approach.
$\mathbf{RVM_{MCMC}}$ (MCMC approach imitating the original RVM)	RBF	100 (maximal 20 permitted at once)/differs in the course of updating	same as $\mathbf{RVM_{naive}}$	same as **BRS**
$\mathbf{RVM_{LatVar}}$ (combination of data augmentation and structural quasi likelihood)	RBF	100/usually very few, see results	same as $\mathbf{RVM_{naive}}$	combination of data augmentation and quasi structural likelihood

Overview of methods and their specifications (continued)

Method	Unknown parameters in the response model (and their estimation scheme)	Unknown parameters in the error model (and their estimation scheme)
$\mathbf{RVM_{naive}}$	fundamental model parameters $\boldsymbol{\omega}$ (mean of full conditional), hyperparameters $\boldsymbol{\alpha}$ (marginal likelihood optimization), σ^2(marginal likelihood optimization), η (grid search)	σ_δ^2 (from usual components of variance analysis using replicates), μ_ξ (analysis of variance formula), σ_ξ^2 (analysis of variance formula)
$\mathbf{RVM_{BC}}$ (RVM + basis function calibration)	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$
$\mathbf{RVM_{EBC}}$ (RVM + expanded basis function calibration)	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$
$\mathbf{RVM_{SIMEX}}$	same as $\mathbf{RVM_{naive}}$	same as $\mathbf{RVM_{naive}}$
BRS (bayesian regression splines)	fundamental model parameters $\boldsymbol{\omega}$ (sampled in Gibbs- step), true covariate values ξ_i (sampled in MH- step), smoothing parameter α (sampled in Gibbs-step)	$\lambda := \frac{\sigma_\xi^2}{\sigma_\xi^2+\sigma_\delta^2/m}$ (discrete Gibbs), μ_ξ (sampled in Gibbs-step), σ_ξ^2(sampled in Gibbs-step)
$\mathbf{RVM_{MCMC}}$ (MCMC approach imitation the original RVM)	same as *BRS*, but no smoothing parameter needed, instead selection of basis function via $\boldsymbol{\gamma}$ (sampled in MH-step)	same as **BRS**
$\mathbf{RVM_{LatVar}}$ (combination of data augmentation and structural quasi likelihood)	same as $\mathbf{RVM_{naive}}$ plus artificial responses $\mathbb{E}(z_i\|y_i, x_i)$ (approximated, see Section 4.1.5)	same as $\mathbf{RVM_{naive}}$ plus σ^{**^2}(marginal likelihood optimization)

4.2.3 The results

All correction methods are again compared with respect to the mean squared error (MSE) and pointwise bias. No further properties like the effective number of kernels or selected η are considered, since it is not expected, that the binary regression results deviate dramatically from the Gaussian results discussed in Section 3.2.3. The approach combining data augmentation and calibration, as described in Section 4.1.5, is only of secondary interest, since it is a genuine ad-hoc method that needs further consideration. Thus, it is compared to the other methods only with respect to the mean squared error.

MSE:

The mean squared error is computed over a grid of 101 equidistant values in the given interval $[a, b]$. The specific values for a and b are stated above in the description of the data cases (cf. Section 4.2.1).

$$\text{MSE} = \frac{1}{101} \sum_{k=1}^{101} \Big(f(\xi_k) - \hat{f}(\xi_k) \Big)^2,$$

where $f(\xi_k) = \mathbb{E}(y_k|\xi_k)$ is the true mean function, i.e. the true probability of $y_k = 1$ given ξ_k. And $\hat{f}(\xi_k)$ is the estimated for this probability given ξ_k.
Table 4.2 presents summary results for the MSE from the 200 simulations for each data scenario. The smallest mean MSE value among the naive analysis and the implemented correction methods in each scenario is in boldface.
In contrast to the simulation study investigating the Gaussian case, the correction quality of the presented methods is mixed. While the correction methods are usually still superior to the naive estimation, there is no method that distinctly dominates the scene here.
Case 1 is particularly well fitted by the **BRS** implementation, which is attributable to the underlying quadratic function of the covariate that is theoretically exactly representable by the utilized second order P-splines. The $\mathbf{RVM_{BC}}$, $\mathbf{RVM_{EBC}}$ and $\mathbf{RVM_{SIMEX}}$ work well in case 2 and case 3, where the $\mathbf{RVM_{SIMEX}}$ approach seems to be slightly superior when the

measurement error variance is higher (case 3: $\sigma_\delta^2 = 0.5^2$, case 4: $\sigma_\delta^2 = 0.8^2$). The $\mathbf{RVM_{MCMC}}$ displays particular strength when the measurement error is moderate as in case 3. $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{EBC}}$ seem to lose their corrective power with higher σ_δ^2, however and most astonishingly the ad-hoc $\mathbf{RVM_{LatVar}}$ still seems to correct properly for higher error as far as can be judged from the MSE values. For case 5 there is obviously cure in neither of the presented strategies and only the $\mathbf{RVM_{LatVar}}$ tends to alleviate the adverse effect of measurement error. This difficulty of the RVM in fitting functions of varying frequencies has already been presented by Krause & Tutz (2003) in the Gaussian case. An improvement for fitting case 5 might rely on the possibility to select locally different kernel parameters instead of selecting only one global kernel parameter η. Though this enhancement is straightforward and requires only a slight modification of the existing programs it has not been implemented, yet.

In the binary regression, where the responses carry less information compared to the Gaussian case flexible regression (even without covariate measurement error) is a demanding task, since a complex probability function must be found from dichotomous outcomes. From case 6, which resembles case 4, but employing $N = 1000$ observations per data set, it seems like the number of observations not only improves the error free analysis but particularly boosts the performance of the correction methods. Case 6 displays an even more pronounced gain for the $\mathbf{RVM_{SIMEX}}$ and now a distinctive gain in using $\mathbf{RVM_{MCMC}}$ and indicates at least a slight gain for the structural correction.

It is again $\mathbf{RVM_{SIMEX}}$ and $\mathbf{RVM_{MCMC}}$ which head the other methods in the misspecification case 7, where the true covariate is not normally distributed anymore.

Obviously all correction methods have difficulties when the true underlying probability is hard to model with radial basis functions and quadratic P-splines, respectively. This could possibly again be improved by allowing for the selection of locally different kernel parameters η.

Mean squared error Mean (SE) / Median (all $\times 10^2$)				
Method	Case 1	Case 2	Case 3	Case 4
RVM	0.17 (.01) / 0.12	0.42 (.02) / 0.36	0.43 (.02) / 0.39	0.40 (.02) / 0.36
$\mathbf{RVM_{naive}}$	0.20 (.01) / 0.13	0.59 (.02) / 0.55	1.76 (.03) / 1.77	2.91 (.04) / 2.79
$\mathbf{RVM_{BC}}$	0.20 (.01) / 0.16	**0.53** (.02) / 0.47	1.54 (.04) / 1.57	3.27 (.05) / 3.42
$\mathbf{RVM_{EBC}}$	0.20 (.01) / 0.16	0.54 (.02) / 0.47	1.55 (.05) / 1.55	3.25 (.05) / 3.38
$\mathbf{RVM_{SIMEX}}$	0.28 (.02) / 0.19	0.61 (.02) / 0.56	1.40 (.03) / 1.39	**2.36** (.06) / 2.14
BRS	**0.16** (.01) / 0.12	0.63 (.03) / 0.58	2.67 (.06) / 2.63	3.74 (.03) / 3.80
$\mathbf{RVM_{MCMC}}$	0.45 (.01) / 0.51	0.62 (.02) / 0.59	**1.24** (.05) / 1.09	2.83 (.08) / 2.61
$\mathbf{RVM_{LatVar}}$	0.21 (.01) / 0.16	0.55 (.02) / 0.49	1.51 (.03)/ 1.53	2.41 (.03) / 2.35

Method	Case 5	Case 6	Case 7	Case 8
RVM	0.45 (.01) / 0.42	0.21 (.01) / 0.19	0.40 (.02) / 0.36	0.95 (.02) / 0.92
$\mathbf{RVM_{naive}}$	0.69 (.02) / 0.67	2.66 (.02) / 2.64	1.72 (.03) / 1.67	**1.41** (.02) / 1.36
$\mathbf{RVM_{BC}}$	0.71 (.02) / 0.68	2.61 (.05) / 2.64	1.34 (.05) / 1.23	1.44 (.02) / 1.38
$\mathbf{RVM_{EBC}}$	0.71 (.02) / 0.68	2.52 (.05) / 2.57	1.35 (.05) / 1.23	1.45 (.02) / 1.39
$\mathbf{RVM_{SIMEX}}$	0.82 (.02) / 0.80	1.99 (.03) / 1.93	**1.05** (.05) / 0.87	**1.41** (.02) / 1.31
BRS	1.12 (.02) / 1.14	3.07 (.03) / 3.06	2.58 (.06) / 2.54	1.88 (.01) / 1.88
$\mathbf{RVM_{MCMC}}$	0.92 (.02) / 0.94	**1.92** (.04) / 1.97	1.16 (.04) / 1.06	1.61 (.02) / 1.56
$\mathbf{RVM_{LatVar}}$	**0.64** (.02) / 0.62	2.13 (.02) / 2.20	1.22 (.03) / 1.19	**1.41** (.02) / 1.34

Table 4.2: The mean squared error results for the simulation. In each column, the smallest mean value among the naive analysis and the implemented correction methods is in boldface.

Pointwise bias:

The pointwise bias of the methods under investigation can be seen from the visualization of their mean predictions for $\mathbb{E}(Y|\xi_k) = P(Y = 1|\xi_k)$ over the 200 simulations in Figure 4.3 (for cases 1-4) and Figure 4.4 (for cases 5-8).
In accordance with the MSE results from Table 4.2, $\mathbf{RVM_{SIMEX}}$ and the sampling method $\mathbf{RVM_{MCMC}}$ show the best bias performance across all scenarios, with exception of case 1. The $\mathbf{RVM_{MCMC}}$ has problems in finding the true probability function for case 1 which is most probably attributable to the fact that this method is not restricted to have a minimum number of 1 basis functions in the model. Since the true probability is rather weakly dependent on the covariate, the $\mathbf{RVM_{MCMC}}$ seems to choose the sparse zero model relatively often, which in turn results in undersmoothing. In contrast to the former MSE results (cf. Table 4.2), where the $\mathbf{RVM_{naive}}$, $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{EBC}}$ have more or less the same MSE value, the $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{EBC}}$ method (indistinguishable in the graphic) show now less deviation from the true probability function than the naive approach does.
The graphs of cases 2-4 witness the strength of $\mathbf{RVM_{SIMEX}}$ and $\mathbf{RVM_{MCMC}}$.
Case 5 is quite surprisingly well fitted by the naive method compared to the others. Among the correction methods, $\mathbf{RVM_{SIMEX}}$ does comparatively well in detecting the true probability function for $\xi_k \leq 0$, while all methods (except BRS) seem to estimate the true probability function adequately for $\xi_k > 0$.
All methods benefit (at least slightly) from the additional amount of available data ($N = 1000$) in case 6 compared to case 4.
All developed correction methods apparently reduces bias in case 7, where ξ was not normally distributed. This is most clearly seen for $\mathbf{RVM_{SIMEX}}$ and $\mathbf{RVM_{MCMC}}$. Especially, the structural methods like, $\mathbf{RVM_{BC}}$, $\mathbf{RVM_{EBC}}$ and $\mathbf{RVM_{MCMC}}$, which assume ξ being normally distributed, were expected to suffer from this distributional misspecification.
Neither of the methods is able to recover the true probability in case 8.

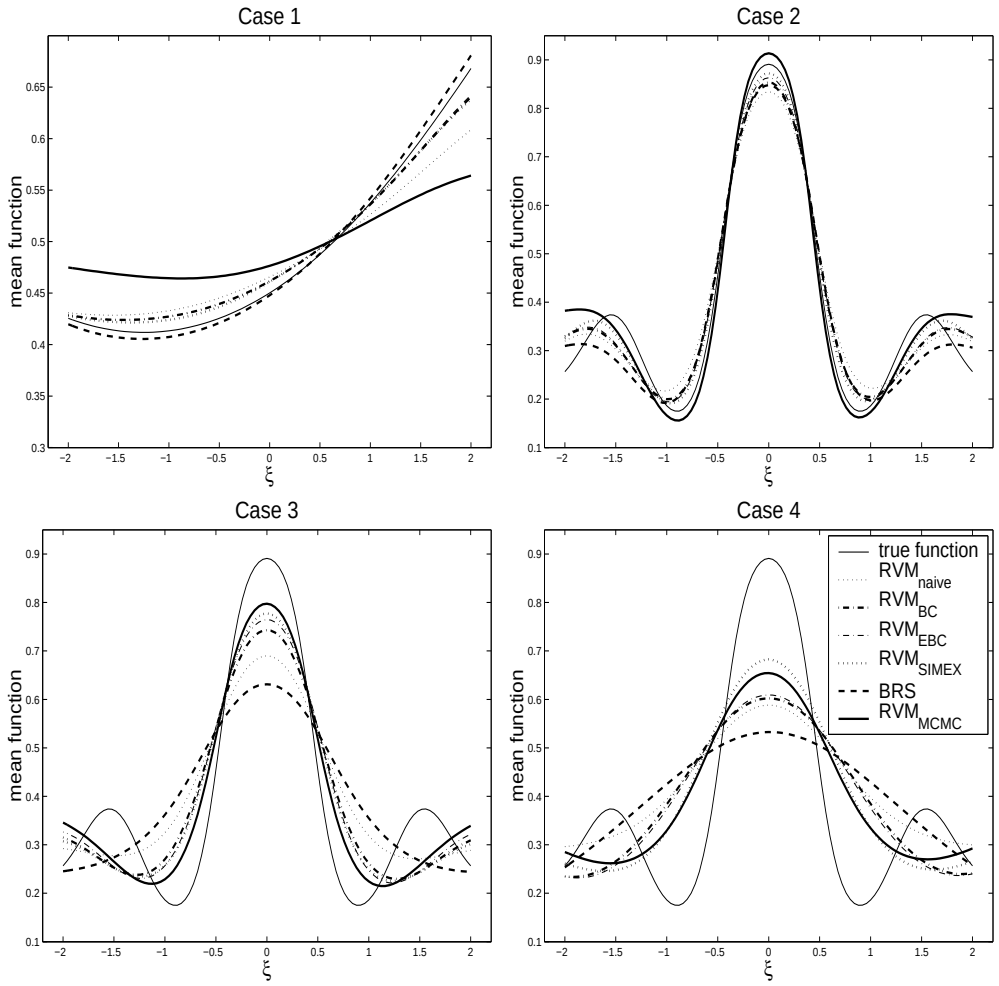

Figure 4.3: The mean functions over 200 simulations for cases 1-4. Data has been generated from the underlying probability function $P(Y = 1|\xi) = (1 + \exp(-m(\xi)))^{-1}$. *Case 1 reflects a weak quadratic relationship between* ξ *and* $m(\xi)$*, with particularly* $\mathbf{RVM_{MCMC}}$ *displaying difficulties. Cases 2-4 represent an oscillating* $m(\xi)$ *yielding true probabilities ranging from* 17.53% *to* 89.09%*. These cases exclusively differ in the amount of measurement error which is* $\sigma_\delta^2 = 0.2^2, 0.5^2, 0.8^2$ *respectively. The strength of* $\mathbf{RVM_{SIMEX}}$ *becomes obvious here for higher measurement error variance. The RVM without measurement error is left out here for sake of visibility.* $\mathbf{RVM_{BC}}$ *and* $\mathbf{RVM_{EBC}}$ *visually coincide in most cases.*

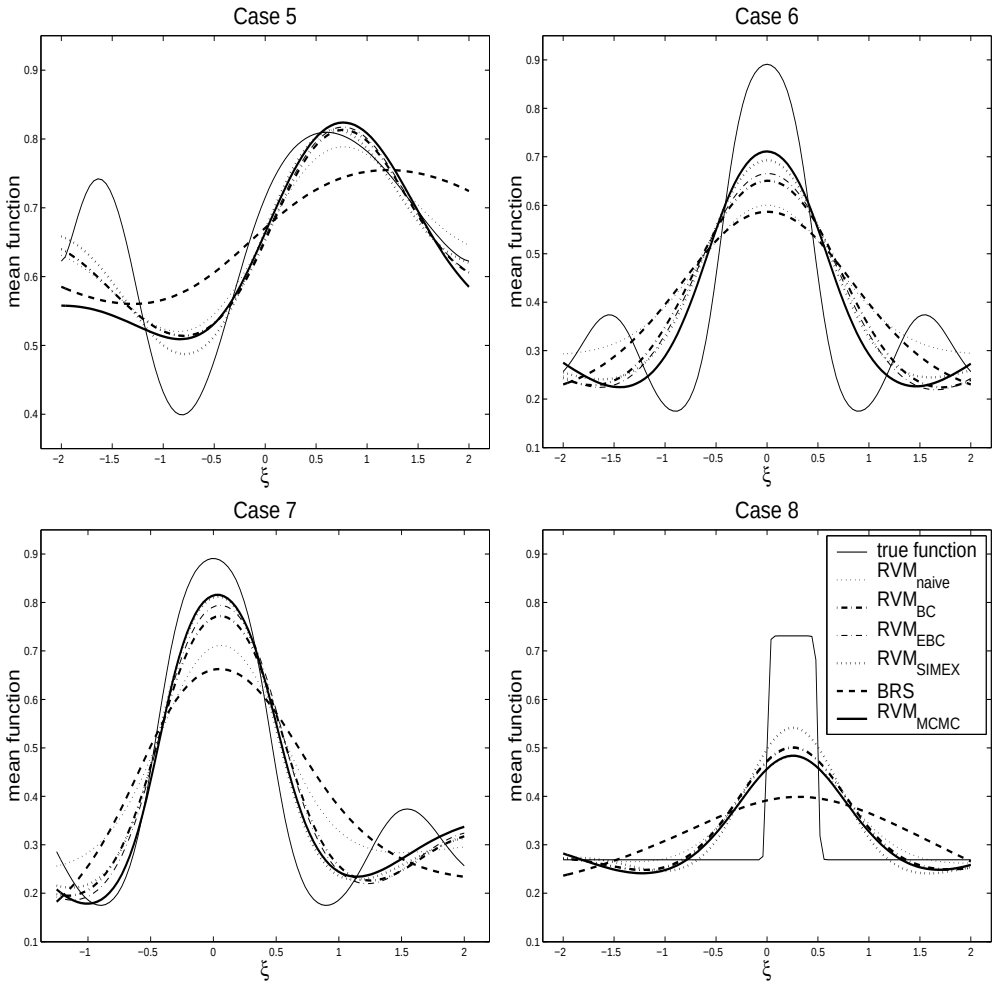

Figure 4.4: The mean functions over 200 simulations for cases 5-8. Case 5 comprises an oscillating $m(\xi)$ with locally different frequencies. Case 6 is identical to case 4, but now supplying more sample data for the methods. Case 7 employs non-normally distributed ξ_i's and case 8 represents a mean function adopting a plateau function which is difficult to fit with RBF kernels and 2nd order truncated power series, respectively. Only a moderate gain of using error correction methods is attested in case 5. In case 6 all correction methods benefit from the larger sample size compared to case 4. Deviation from the normal distribution for ξ in case 7 seems not to harm the power of the (structural) correction methods profoundly. In contrast, neither of the correction strategies can even roughly approximate case 8.

4.3 A Real data example[1]

This section presents some of the previously introduced correction methods of this chapter in a real data example. For this purpose, data from the German panel of the WHO MONICA project (MONItoring of trends and determinants in CArdiovascular disease, cf. , e.g. Döring & Kußmaul (1997), Keil (2000)) is analyzed.
The main target is to quantify the influence of nutrition on cardiovascular disease and mortality, where the focus is on animal and plant protein intake. For a subpopulation of $N = 892$ male respondents, data of a mortality follow up for more than ten years is available. Besides information about nutritional habits and confounders like Cholesterol, it contains also the age, daily alcohol consumption, presence of Hypertonia and smoking status of the participants. Though also information about morbidity is available, this is not considered here.
The nutritional details were obtained by a comprehensive diary where the study participants had to fill in all meals for seven consecutive days. The individual plant and animal protein intake was calculated from the raw data based on nutritional data containing standard recipes. The intake values, computed in such a way, are suspect in double regard: a seven day observation of all meals is a questionable operationalization of the targeted variable 'nutritional habits'. Second, although high attention has been paid to exactly distill the protein intake from the meals, substantial measurement error is unavoidable.
Augustin (2002) investigates the effects on the survival times using a Cox model under regression calibration. However, here the dichotomous information of surviving/not surviving the monitoring period is chosen as response for the flexible binary regression applying error correction. In this work, these data are for the first time analyzed using error correction methods and

[1]This section is a contribution to a common research project with Angela Döring and H.-Erich Wichman (GSF - National Research Center for Environment and Health, Neuherberg).

flexible regression. Thus the obtained results may reveal interesting new insights in how the protein intake affects the organism.

The particular challenge is to adequately consider the fact that **two** covariates are error-prone and thus facing a so-called additive model, where both are modeled in a flexible way. The confounders are modeled linearly. To the author's knowledge there is only one reference work on this topic of additive models under measurement error published so far, which is by Ganguli et al. (2005).
The following section shows how basis function calibration and SIMEX is generalized in order to suit the present case. Though application of expanded basis function calibration is also possible, this is omitted here because it shows nearly indistinguishable performance from basis function calibration in the previous simulations (cf. Section 4.2). Application of a Bayesian MCMC approach would be highly desirable, which, however, requires appropriate consideration of the correlation structure between the unobserved covariates and the confounder variables, which is not yet realized.

4.3.1 Naive approach and basis functions calibration

The naive approach simply uses the seven-days averages of animal and plant protein intake $\mathbf{X} = (X_1, X_2)$ with X_1 = averaged animal protein intake [g/day] and X_2 = averaged plant protein intake [g/day] as surrogates for the true intake. The following confounders $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4, Z_5)$ are included: Z_1 = Cholesterol [mg/dl], Z_2 = Hypertonia [1=yes,0=no], Z_3 = age at event/end of study, Z_4 = smoking status [1=yes,0=no], Z_5 = alcohol consumption [g/day]. These confounders enter linearly into the analysis and their corresponding parameters are not penalized, i.e. the corresponding prior precisions α_{Z_1}, α_{Z_2}, α_{Z_3}, α_{Z_4} and α_{Z_5} are chosen very small a priori in (4.1). The use of Z_3 as a covariate might be critical since it contains information about the time under study until an event/ end of study and, thus, can also be thought of as a dependent variable.

The basis function calibration approach accounts for measurement error by utilizing the calibrated basis functions $\mathbb{E}(\phi(\xi)|\mathbf{X}, \mathbf{Z}^*)$. It uses Z_3^* = the age of the responders at study entry for calibration instead of Z_3 = the age at event/end of study, which explains the notation $\mathbf{Z}^*$ in the calibrated basis. The variable Z_3^* is used for calibration, because one would suspect that the age at study entry Z_3^* has more information about the plant and animal protein intake at that time than the variable Z_3. The basis function approach considers the correlation structure between all covariates.
An important feature of the data is that **two** covariates are error-prone and both should be modeled in a flexible way. Recall the additive weighted basis function structure in the ideal mean model of the RVM from 2.1.1, here for two main effects (for simplicity without any confounders)

$$\mathbb{E}(Y|\boldsymbol{\xi}) = G\left(\sum_{d=1}^{2}\sum_{j=1}^{J_d}\omega_{dj}\phi_{dj}(\xi_d) + \omega_0\right),$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2)$. Basis function calibration replaces the latent basis functions $\phi_{dj}(\xi_d)$ in this ideal mean model by their calibrated versions. This yields the approximate observed model

$$\mathbb{E}(Y|\mathbf{X}, \mathbf{Z}^*) \approx G\left(\sum_{d=1}^{2}\sum_{j=1}^{J_d}\omega_{dj}\,\mathbb{E}\left(\phi_{dj}(\xi_d)|\mathbf{X}, \mathbf{Z}^*\right) + \omega_0\right). \tag{4.57}$$

In order to compute $\mathbb{E}\left(\phi_{dj}(\xi_d)|\mathbf{X}, \mathbf{Z}^*\right)$, the distribution $f_{\xi_d|\mathbf{X},\mathbf{Z}^*}$ is required. Therefore, in a first step, the joint conditional distribution $f_{\boldsymbol{\xi}|\mathbf{X},\mathbf{Z}^*}$ is calculated.
If all variables $\boldsymbol{\xi}, \mathbf{X}, \mathbf{Z}^*$ can reasonably be assumed being normal then one can find the mean and variance of the conditional distribution $f_{\boldsymbol{\xi}|\mathbf{X},\mathbf{Z}^*}$, which is again a normal distribution, by the following theorem:

Theorem:
For two vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ that are jointly normal, i.e.

$$\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{2,1} & \Sigma_{2,2}\end{pmatrix}\right),$$

the conditional distribution of $\mathbf{X}_1|\mathbf{X}_2$ is given by

$$\mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \Sigma_{1|2})$$

with

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1} \\ \Sigma_{1|2} &= \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}. \end{aligned}$$

However, Z_2 = Hypertonia and Z_4 = smoking status enter as binary variables into the model and thus are far from being normally distributed. Consequently, it is necessary to stratify the sample according to these variables yielding four sub-samples defined by the four possible combinations of Hypertonia and smoking status.
Multivariate analysis of variance formulas are required to compute the components of the mean vector and covariance matrix of the joint distribution $f_{\boldsymbol{\xi},\mathbf{X},Z_1,Z_2=z_2,Z_3^*,Z_4=z_4,Z_5}$ for each stratum defined by the absence/presence of Hypertonia $z_2 \in \{0,1\}$ and smoking habit $z_4 \in \{0,1\}$ (cf. Carroll et al. (1995)). The measurement errors are again assumed being independently distributed and their variances can be computed from the seven replicate measurements as before (cf. Section 2.3.1). Most conveniently, the required marginal conditional distribution $f_{\xi_d|\mathbf{X},Z_1,Z_2=z_2,Z_3^*,Z_4=z_4,Z_5}$ in order to perform the calibration in (4.57) is also Gaussian. Its moments are obtained by picking the dth element from the mean vector and diagonal element from the covariance matrix in $f_{\boldsymbol{\xi}|\mathbf{X},Z_1,Z_2=z_2,Z_3^*,Z_4=z_4,Z_5}$. Care must be taken, because there are four different conditional distributions used for calibration and each observation's Hypertonia/smoking status decides which one to use.
Correcting for heteroscedastic additive error, as presented by Augustin (2002), could be straightforwardly realized here by accordingly modifying the joint distribution of $\boldsymbol{\xi}$ given the observed variables which is however not presented here.

The smooth function estimates modeling the influence of animal and plant protein on the risk of dying are displayed in Figure 4.5 for the naive and

corrected analysis under basis function calibration. The results are discussed here in turn and compared to the naive and corrected analysis by Augustin (2002).

The naive analysis

The naive analysis judges low animal protein intake having an adverse effect, while there is no indication of high protein intake doing so, though there are reasonable arguments that both types of extreme low and high intakes could be detrimental. Higher plant protein intake up to a certain amount seems to lower the risk of dying, but consuming beyond an amount of 30g/day appears to invert this effect.
In Table 4.3 the naive parameter estimates and their corresponding p-values based on the posterior covariance matrix are given. Though the concepts of p-value and significance are squeamishly reserved to frequentistic statistics, they are used here under the viewpoint that parameter estimation of the main parameters $\boldsymbol{\omega}$ in the Bayesian RVM is equivalent to parameter estimation in a frequentistic penalized likelihood setting (cf. Section 2.1.2).
Only one basis function is used here to model the influence of animal protein and its corresponding parameter is significant on the 5% level. There are two radial basis functions modeling the influence of plant protein. Since they are located at proximate knots, this can not be seen from Figure 4.5. Only for one of these basis functions, the corresponding parameter estimate is significantly different from zero and only this estimate is displayed in Table 4.3.
Hypertonia and smoking status are found to be relevant risk determinants. The slightly negative - however significant - effect of age is contributable to the type of study: people who went through the complete ten years follow-up are generally older than those, who died during the follow-up period.
These naive results are in concordance with the naive results from the Cox model analysis as obtained by Augustin (2002). However, the beneficial effect of low and medium doses of plant protein intake is only significant for the flexible regression.

The corrected analysis

The influence of animal and plant protein are modeled by only a single basis function, respectively. The adverse influence of low doses of animal protein intake is much more pronounced in the corrected analysis compared to the naive analysis. However, the beneficial effect of low and medium doses of plant protein is no longer significant on the 5 % level. The effects of the confounders are confirmed by the corrected analysis since parameter estimates and p-values remain here unchanged in essence. These corrected results are in concordance with the results from the Cox model analysis correcting for homoscedastic measurement error as developed by Augustin (2002). However, the detrimental effect of low animal protein intake becomes significant only in the flexible regression applying error correction.
It must, however, be noted that the standard errors on which the p-values are based do not account for the uncertainty in estimating the calibrated basis functions $\mathbb{E}(\phi(\xi)|\mathbf{X}, \mathbf{Z}^*)$ and thus can only communicate a rough impression of significance for the parameters. Correct estimators for standard errors is a open problem for flexible regression using calibration methods.

Naive and corrected estimates for basis function calibration				
Parameter	Naive estimate	p-value	Corrected estimate	p-value
ω_{animal}	0.7164	0.0430	1.1291	0.0139
ω_{plant}	-0.5811	0.0317	-0.5611	0.0850
Cholesterol	-0.0006	0.7737	-0.0008	0.7137
Hypertonia	0.7415	0.0005	0.7143	0.0009
age	-0.0423	0.0000	-0.0436	0.0000
smoking status	1.0646	0.0006	1.0708	0.0006
alcohol	-0.0005	0.8780	-0.0002	0.9452

Table 4.3: Naive and corrected parameter estimates under basis function calibration of animal/plant protein intake and of confounders determining the risk of dying from cardiovascular disease. The p-values are calculated from the posterior covariance matrix.

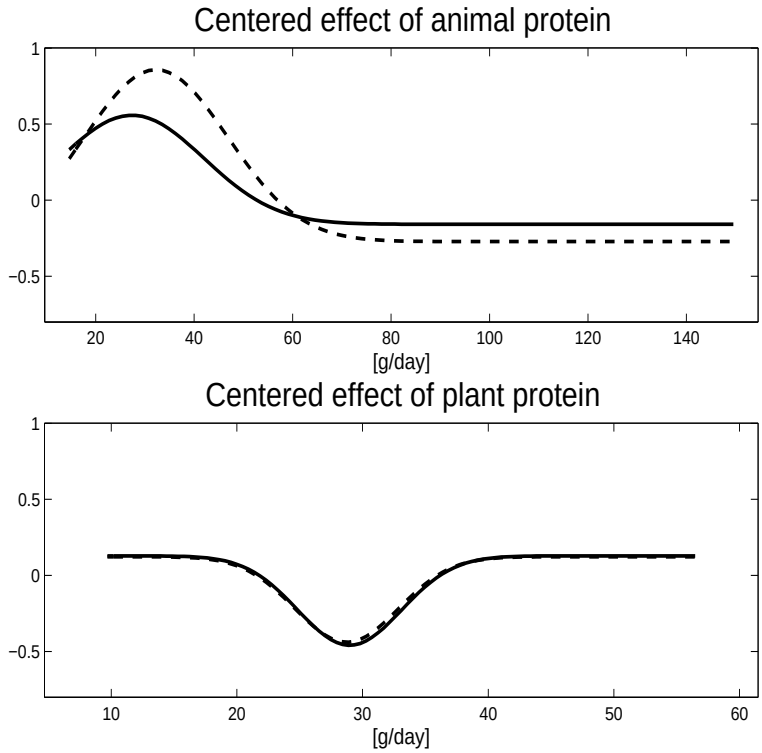

Figure 4.5: Centered smooth influence of animal and plant protein intake on the risk of dying from cardiovascular disease. Solid line represents naive estimate and dashed line represents estimate after basis function calibration.

4.3.2 SIMEX

To the author's knowledge, there exists no literature on SIMEX correction in additive models. A reasonable requirement seems to be that the SIMEX prediction is again a sum of the influence of animal and the influence of plant protein intake. This can be realized by studying the effect of additional measurement error separately on each determinant. Thus the risk associated with animal and the risk associated with plant protein are corrected separately. Therefore, the mean model is split into three parts

$$f = G(f_{animal} + f_{plant} + \mathbf{Z}\boldsymbol{\omega}_Z),$$

which belong to the influence of animal and plant protein intake and the confounders. The influence of animal and plant protein intake will again be modeled with the help of radial basis functions and the confounder enter linearly.
The effect of additional measurement error is then studied separately on these three model parts and corrected SIMEX estimates are computed for f_{animal}, f_{plant} and $\boldsymbol{\omega}_Z$.

Figure 4.6 contains the SIMEX corrected predictors $\widehat{f}_{animal}$ and $\widehat{f}_{plant}$ and the naive prediction for comparison. For animal protein, there seems to be

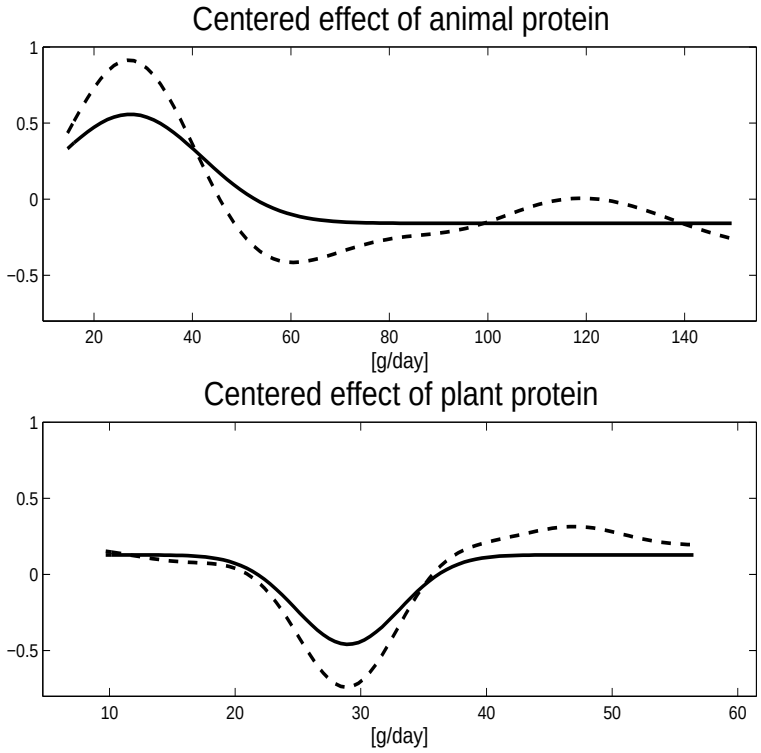

Figure 4.6: Centered smooth influence of animal and plant protein intake on the risk of dying from cardiovascular disease. Solid line represents naive estimate and dashed line represents estimate under SIMEX correction.

an optimal daily intake of $\approx$ 60g. Lower intake is associated with rather high

risk, while consuming beyond the optimal intake seems to increase the risk slightly with every additional gram. Low animal protein intake is obviously more detrimental than high intake.
The beneficial effect of plant protein intake is pronounced after correction. High plant protein intake appears to be slightly more detrimental than low intake. The SIMEX corrected parameter estimates of the confounders are given in Table 4.4. The estimated effects of the confounder remain essentially unaffected by the SIMEX correction.

Naive and corrected estimates for SIMEX		
Parameter	Naive estimate	Corrected estimate
Cholesterol	-0.0007	-0.0002
Hypertonia	0.7461	0.7101
age	-0.0424	-0.0407
smoking status	1.0582	1.0647
alcohol	-0.0003	-0.0004

Table 4.4: Naive and SIMEX corrected parameter estimates of the confounders determining the risk of dying from cardiovascular disease.

4.4 Error correction in flexible models for binary longitudinal data

Often, the target of epidemiological studies is to quantify how the study objects' health is affected by e.g. environmental influences or dietary habits. Therefore, the objects are typically traced over a certain period while the variables of interest are measured repeatedly at certain time points.
The modeling requirements to analyze such longitudinal data appropriately include the allowance for person specific effects and serial correlation of the errors. The previous sections of this chapter only consider cross-sectional data. The case of individuals providing multiple measurements, representing a dynamic development over time, was excluded, though it is highly relevant in practice.
Chib & Jeliazkov (2006) present an MCMC sampling approach to the analysis of semiparametric models with serially correlated errors. However, to the author's knowledge the problem of covariate measurement error has not yet been treated in this complex context. Based on the MCMC error correction approach developed in 4.1.4, the necessary amendments to satisfy the modeling needs of longitudinal data are presented in this section.
Firstly, the person specific random effect and serially correlated responses are included into the Bayesian probit model. The new random effects are simply additional unknown parameters in the mean model and the serial correlation modifies the formerly diagonal covariance matrix of the (response) errors into a block diagonal matrix. Finally, the serially correlated latent covariate observations are considered in much the same way as the responses.
After the specification of the prior distributions over the newly introduced model parameters, the complete sampling scheme is explained. This section concludes with the demonstration of this approach by means of a few data examples.

4.4.1 The model setup

A single binary response is now denoted as y_{it}. The indices i and t ($i = 1, \ldots, n,\ t = 1, \ldots, T_i$) with $\sum_i^n T_i = N$ refer to individual objects (more generally termed 'clusters') and time points at which a measurement is taken, respectively.

The longitudinal Bayesian probit model with Bayesian model selection via the vector $\boldsymbol{\gamma}$ is then

$$\begin{aligned} y_{it} &= \begin{cases} 1 & : \text{ if } z_{it} > 0 \\ 0 & : \text{ otherwise} \end{cases} \\ z_{it} &= b_i + \Phi_{\boldsymbol{\gamma}}(\xi_{it})\boldsymbol{\omega}_{\boldsymbol{\gamma}} + \epsilon_{it}, \\ & \quad b_i \sim \mathcal{N}(0, D), \end{aligned} \tag{4.58}$$

where the exact specification of the errors ϵ_{it} is delayed for a moment.
The random intercepts b_i are introduced to identify cluster specific effects and follow a Gaussian distribution. Of course, also random slopes are conceivable representing e.g. the individual (in contrast to all-over) impact of a covariate. However, to keep the model (notationally) clear the present study is confined to the very popular case of a random intercept to demonstrate the approach. Besides the random effects, there are two more important sources of inter-temporal dependence in the observations $y_{it}, t = 1, \ldots, T_i$ from cluster i. One source is due to lags that capture the so-called 'state dependence', where the probability of a certain outcome may depend on (a series of) past responses. Though this is easily realized in the model (cf. Chib & Jeliazkov (2006)), it is again not captured here in order to keep the model simple. The second source is the presence of serial correlation in the errors $\epsilon_{it}, t = 1, \ldots, T_i$ within a cluster. In order to account for this in the model, a zero mean stationary pth order autoregressive, AR(p), process can be specified, for instance. It is parameterized in terms of the autocorrelation parameters $\rho, \ldots, \rho_p$ and given as

$$\epsilon_{it} = \rho\epsilon_{it-1} + \ldots + \rho_p\epsilon_{it-p} + v_{it}, \quad v_{it} \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{4.59}$$

An AR(1) process will be sufficient in most cases and will be considered for the rest of this section.
It is convenient to combine the T_i observations belonging to the ith cluster into a vector, which yields

$$\mathbf{z}_i \;=\; b_i + \Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \Omega_i). \tag{4.60}$$

Here, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iT_i})^{\mathrm{T}}$ and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})^{\mathrm{T}}$ are column vectors and $\Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)$ is the matrix of basis functions at the cluster values $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iT_i})^{\mathrm{T}}$. The errors within a cluster are normally distributed like in the cross-sectional case, before (cf. (4.33)), but are now longer independent. The covariance matrix Ω_i is the $T_i \times T_i$ Toeplitz matrix as implied by the autoregressive process (4.59). For the $AR(1)$ process applied here, the matrix elements are given by $\Omega_i[j,k] = \rho^{|j-k|}/(1-\rho^2)$, $1 \leq j,k \leq T_i$, where $[j,k]$ denotes the kth element of the jth row of the matrix.

Now, the covariate model describing the generative process of the ξ's is characterized. Due to the longitudinal data structure, it is no longer reasonable to assume that the latent covariate observations $\xi_{it}, t = 1, \ldots, T_i$ are independent as before (cf. (4.36)). Instead, assuming an AR(1) correlation structure for the true covariate observations within a cluster yields the model for the covariate observations

$$\boldsymbol{\xi}_i \;=\; \mu_\xi + \boldsymbol{\epsilon}_{\boldsymbol{\xi}_i}, \quad \boldsymbol{\epsilon}_{\boldsymbol{\xi}_i} \sim \mathcal{N}(0, \Omega_{\boldsymbol{\xi}_i}\sigma_\xi^2). \tag{4.61}$$

The deviations from the mean are again assumed being normally distributed as in the cross-sectional case (cf. 4.36). However, the covariance matrix $\Omega_{\boldsymbol{\xi}_i}$ is a $T_i \times T_i$ Toeplitz matrix according to the AR(1) process. The elements are given by $\Omega_{\boldsymbol{\xi}_i}[j,k] = \rho_\xi^{|j-k|}/(1-\rho_\xi^2)$, $1 \leq j,k \leq T_i$ with ρ_ξ denoting the autocorrelation parameter for the latent covariate observations. Most notably, if there is correlation in the data, i.e. $\Omega_{\boldsymbol{\xi}_i} \neq \mathbf{I}$, then σ_ξ^2 is no longer the variance of ξ_{it} but besides the factor $1/(1-\rho_\xi^2)$ merely one component of $\mathbb{V}(\xi_{it}) = \sigma_\xi^2/(1-\rho_\xi^2)$.
One can also think of including a random intercept in the covariate model (4.61) in order to account for cluster specific effects. This can be handled in

complete analogy to the random intercept in the response model. Another alternative, to relax the assumption of all clusters coming from a normal distribution with same mean and variance structure, can be adopted from Carroll et al. (1999). They enhance a Bayesian method for estimating the parameters of a mixture of k normals in the face of inherent measurement error. Again, only the relatively simple model (4.61) is considered in the present demonstration.

To improve readability, the n clusters are stacked to give the model in terms of all N observations

$$\mathbf{z} \;=\; W\mathbf{b} + \Phi_{\boldsymbol{\gamma}}\boldsymbol{\omega}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \Omega), \tag{4.62}$$

where the latent response vector $\mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$ contains the cluster specific responses (4.60) and has dimension $(N \times 1)$.
The $(N \times n)$ design matrix for the cluster specific effects is of block diagonal structure and includes the $(T_i \times 1)$ intercept vectors denoted as $\mathbf{1}_i$. It is given by

$$W = \begin{pmatrix} \mathbf{1}_1 & & \\ & \ddots & \\ & & \mathbf{1}_n \end{pmatrix},$$

with related $(n \times 1)$ parameter vector $\mathbf{b} = (b_1, \ldots, b_n)^{\mathrm{T}}$.
The covariance matrix of the errors in (4.62) is of block diagonal structure and defined as

$$\Omega = \begin{pmatrix} \Omega_1 & & \\ & \ddots & \\ & & \Omega_n \end{pmatrix},$$

containing the covariance matrices Ω_i from (4.60).
The covariate model, describing the generative process of all N latent covariate observations is given by

$$\boldsymbol{\xi} \;=\; \mu_\xi + \boldsymbol{\epsilon}_{\boldsymbol{\xi}}, \quad \boldsymbol{\epsilon}_{\boldsymbol{\xi}} \sim \mathcal{N}(0, \Sigma_{\boldsymbol{\xi}}), \tag{4.63}$$

where the vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\xi}_n^{\mathrm{T}})^{\mathrm{T}}$ is composed of the sub-vectors from (4.61). The covariance matrix for all latent covariate observations is of block

diagonal form and defined as

$$\Sigma_{\boldsymbol{\xi}} = \sigma_\xi^2 \Omega_{\boldsymbol{\xi}} = \begin{pmatrix} \Sigma_{\boldsymbol{\xi}_1} & & \\ & \ddots & \\ & & \Sigma_{\boldsymbol{\xi}_n} \end{pmatrix},$$

$$\text{with} \quad \Sigma_{\boldsymbol{\xi}_i} := \sigma_\xi^2 \Omega_{\boldsymbol{\xi}_i}, \quad i = 1, \ldots, n. \tag{4.64}$$

Here, the covariance matrices $\Omega_{\boldsymbol{\xi}_i}$ are from (4.61).
To complete the model specification, the measurement error model, relating true and observed covariate observations, is required. Again, a classical additive measurement error model is specified, which remains essentially unchanged to the cross-sectional case (4.39) in Section 4.1.4, but now the replicate measurements are taken for cluster i at a certain point in time t. The parameter λ is in the longitudinal case defined as

$$\lambda = \frac{\sigma_\xi^2}{\sigma_X^2(1-\rho_\xi^2)}, \tag{4.65}$$

where $\frac{\sigma_\xi^2}{1-\rho_\xi^2}$ is the variance of ξ. After the response model, the covariate model and the measurement error model have been specified, the specific prior distributions are considered.

Specification of prior distributions

The prior over the weights $\boldsymbol{\omega}$ is again a Gaussian distribution as in the cross-sectional MCMC approach (cf. (4.34)).
To perform Bayesian model averaging the additional parameter vector $\boldsymbol{\gamma}$ is introduced into the model with a prior distribution as given in (4.35).
The priors over μ_ξ and σ_ξ^2 are, as in the cross-sectional case, chosen as a normal distribution (cf. (4.37)) and an inverse Gamma distribution (cf. (4.38)), respectively.
The parameter specifications for these priors are adopted from Section 4.1.4. However, the robustness of the method with respect to different prior specifications will be checked later. As above, a uniform prior is chosen for the

attenuation factor λ.

New prior specifications are required for the variance of the cluster specific random effect and both autocorrelation parameters.
A Wishart prior is defined over the inverse variance of the random intercept

$$D^{-1} \sim \mathrm{W}(r_0, R_0)$$

and truncated Gaussian priors over the autocorrelation coefficients of the latent responses and covariate observations, respectively

$$\begin{aligned} \rho &\sim \mathcal{N}(\rho_0, P_0) I(-1 < \rho < 1) \\ \rho_\xi &\sim \mathcal{N}(\rho_{\xi 0}, P_{\xi 0}) I(-1 < \rho_\xi < 1). \end{aligned}$$

The normalizing constant making the truncated normal a genuine density is suppressed in the prior distributions over the autocorrelation parameters.
General priors can be adopted quite easily and analysis proceeds then via weighted resampling of the MCMC draws obtained from a model using the priors presented here.
In the data examples presented below, the following priors specifications are used, if not stated otherwise: $r_0 = 2, R_0 = 2.5$ and $\rho_0 = \rho_{\xi 0} = 0, P_0 = P_{\xi 0} = 0.5^2$.

4.4.2 Inference

The introduction of serially correlated responses and latent covariate observations and the introduction of a cluster specific effect, affect almost all full conditional distributions. The details of the new sampling scheme, now accounting for the longitudinal structure in the Bayesian probit model, are explained in this subsection.

In random coefficient models it is convenient to marginalize over the random effect, which yields the following marginal response model

$$\mathbf{z} = \Phi_{\boldsymbol{\gamma}} \boldsymbol{\omega}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* \sim \mathcal{N}(0, \Omega^*). \tag{4.66}$$

The block diagonal covariance matrix of the vector $\boldsymbol{\epsilon}^*$

$$\Omega^* = \begin{pmatrix} \Omega_1^* & & \\ & \ddots & \\ & & \Omega_n^* \end{pmatrix} \tag{4.67}$$

contains the blocks $\Omega_i^* = D + \Omega_i$, where the variance of the random intercept D is added to each element in Ω_i. This model is then used in computing the full conditional distribution of the weights, which is slightly different from the cross-sectional case in Section 4.1.4, where the response error covariance matrix was an identity matrix

$$\begin{aligned} p(\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\mathbf{z}, D, \rho, \boldsymbol{\xi}) \quad &= \quad \mathcal{N}(\mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}, \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}) \\ \text{where} \quad &\mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} = \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \Phi_{\boldsymbol{\gamma}}^{\mathrm{T}} {\Omega^*}^{-1} \mathbf{z} \\ &\Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} = (v^{-1}\mathbf{I} + \Phi_{\boldsymbol{\gamma}}^{\mathrm{T}} {\Omega^*}^{-1} \Phi_{\boldsymbol{\gamma}})^{-1} \end{aligned} \tag{4.68}$$

and v denotes the prior variance of $\boldsymbol{\omega}$ coming from its prior specification (4.34).

The mechanism of basis function selection again works via a Metropolis Hastings step for the parameter vector $\boldsymbol{\gamma}$, which names the selected basis functions. First, a proposal is generated of which basis function to introduce into/exclude from the current model (cf. Section 4.1.4). Then it must be decided whether to accept this proposal or not. Therefore, the acceptance probability, as given by (4.43), needs to be computed, which in turn requires computation of the marginal likelihood.
The marginal likelihood accounts for the autocorrelation in the responses and is based on the moments of the full conditional distribution (4.68)

$$\begin{aligned} p(\mathbf{z}|D, \rho, \boldsymbol{\xi}, \boldsymbol{\gamma}) \quad &= \quad (2\pi)^{-\frac{N}{2}} \frac{\mid \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot} \mid^{\frac{1}{2}}}{\mid v_{\boldsymbol{\gamma}} \mid^{\frac{1}{2}} \mid \Omega^* \mid^{\frac{1}{2}}} \\ &\exp\left(-\frac{1}{2} (\mathbf{z}^{\mathrm{T}} {\Omega^*}^{-1} \mathbf{z} - \mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}^{\mathrm{T}} \Sigma_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}^{-1} \mu_{\boldsymbol{\omega}_{\boldsymbol{\gamma}}|\cdot}) \right), \end{aligned}$$

where Ω^* is the block diagonal covariance matrix as defined in (4.67). Most conveniently, the determinant of the covariance matrix Ω^* and the term

$\mathbf{z}^{\mathrm{T}}\Omega^{*^{-1}}\mathbf{z}$ cancel out in the acceptance probability since both are independent of $\boldsymbol{\gamma}$. Here, $v_{\boldsymbol{\gamma}}$ denotes the prior covariance matrix over the weights for a model as defined by $\boldsymbol{\gamma}$.

In order to sample the latent responses z_{it}, several strategies are conceivable. The updating rule developed here, particularly considers computational efficiency and minimization of posterior correlation between $\mathbf{z}$ and $\boldsymbol{\omega}_{\boldsymbol{\gamma}}$. In the case of cross-sectional data, the single elements in $\mathbf{z}$ are sampled from their respective full conditional after marginalizing over the parameter vector $\boldsymbol{\omega}$ in order to reduce posterior correlation (cf. Holmes & Held (2006)).
The resulting full conditional of a single z_i after marginalizing over $\boldsymbol{\omega}$ is briefly recalled here for the cross-sectional case (4.42) as

$$p(z_i|\mathbf{z}_{-i}, y_i, D, \rho, \xi_i, \boldsymbol{\gamma}) \quad \propto \quad \begin{cases} \mathcal{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i > 0) & \text{if } y_i = 1 \\ \mathcal{N}(\mu_{z_i|\cdot}, \Sigma_{z_i|\cdot})I(z_i < 0) & \text{otherwise} \end{cases} \tag{4.69}$$

However, in the longitudinal case, these observational units z_i that formerly have been scalars are now vectors $\mathbf{z}_i$ containing the serially correlated elements z_{it} of a cluster. Furthermore, a random intercept is now present in the model as well. Thus, the updating scheme for the latent responses must be fundamentally reconsidered.
In a first step, the full conditional of the vector $\mathbf{z}_i$ being marginal of $\mathbf{b}$ and $\boldsymbol{\omega}$ is required. This is a truncated normal distribution and the calculation of the respective moments is in analogy to the earlier derivation of the moments in (4.42) and is presented in the following.
The moments of the conditional distribution (4.69) are the well known leave-one-out mean and variance that are used in the leave-one-out cross validation e.g. for finding optimal smoothing parameters in spline regression. Instead, for the longitudinal case the 'leave-one-block-out' moments (referring to a 'block' of correlated responses within a cluster) are sought now.
The full conditional of a single cluster i is again a multivariate truncated normal

$$p(\mathbf{z}_i|\mathbf{z}_{-i}, \mathbf{y}_i, D, \rho, \boldsymbol{\xi}, \boldsymbol{\gamma}) \quad \propto \quad \begin{cases} \mathcal{N}(\mu_{\mathbf{z}_i|\cdot}, \Sigma_{\mathbf{z}_i|\cdot})I(\mathbf{z}_i > 0) & \text{if } \mathbf{y}_i = 1 \\ \mathcal{N}(\mu_{\mathbf{z}_i|\cdot}, \Sigma_{\mathbf{z}_i|\cdot})I(\mathbf{z}_i < 0) & \text{otherwise} \end{cases} \tag{4.70}$$

where the indicator function $I(\cdot)$ and the comparison $\mathbf{y}_i = 1$ work on the individual elements in $\mathbf{z}_i$ and $\mathbf{y}_i$, respectively.
Applying the small rank adjustment formula (cf. Horn & Johnson (1985)) and some basic matrix algebra, this leave-one-block-out covariance matrix of cluster i can be found to be

$$\Sigma_{\mathbf{z}_i|\cdot} = \Omega_i^*(\Omega_i^* - H_i)^{-1}\Omega_i^*, \tag{4.71}$$

with Ω_i^* being the ith diagonal block of the covariance matrix of the response error as defined in (4.67). The matrix $H_i = \Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)\Sigma_{\boldsymbol{\omega}_\gamma|\cdot}\Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)^{\mathrm{T}}$ contains $\Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)$, which is a cutout of the design matrix comprising the T_i row vectors associated with the elements in $\boldsymbol{\xi}_i$. The matrix $\Sigma_{\boldsymbol{\omega}_\gamma|\cdot}$ is borrowed from the full conditional of the weights above, cf. (4.68). H_i can equivalently be viewed as the $T_i \times T_i$ block matrices along the diagonal of the Bayesian hat matrix $H := \Phi_{\boldsymbol{\gamma}}\Sigma_{\boldsymbol{\omega}_\gamma|\cdot}\Phi_{\boldsymbol{\gamma}}^{\mathrm{T}}$.
The calculation of the leave-one block-out mean vector in (4.70) involves the previous result (4.71) and is given by

$$\mu_{\mathbf{z}_i|\cdot} = \Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)\mu_{\boldsymbol{\omega}_\gamma|\cdot} - H_i(\Omega_i^* - H_i)^{-1}\left(\mathbf{z}_i - \Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)\mu_{\boldsymbol{\omega}_\gamma|\cdot}\right), \tag{4.72}$$

where $\mu_{\boldsymbol{\omega}_\gamma|\cdot}$ is the mean of the full conditional distribution of the weights (4.68). Now, the full conditional distribution (4.70) could be used to sample values of $\mathbf{z}_i$, but direct sampling from the multivariate truncated normal is known to be difficult.
More conveniently, the univariate distribution of a single z_{it}, given all other observations, i.e. $z_{it}|\mathbf{z}_{-i}, \mathbf{z}_{i-t}$, is sought for generating the samples. Applying the results from Robert (1995), the parameters of the univariate full conditional

$$p(z_{it}|\mathbf{z}_{-i}, \mathbf{z}_{i-t}, y_{it}, D, \rho, \boldsymbol{\xi}, \boldsymbol{\gamma}) \propto \begin{cases} \mathcal{N}(\mu_{z_{it}|\cdot}, \Sigma_{z_{it}|\cdot})I(z_{it} > 0) & \text{if } y_{it} = 1 \\ \mathcal{N}(\mu_{z_{it}|\cdot}, \Sigma_{z_{it}|\cdot})I(z_{it} < 0) & \text{otherwise} \end{cases}$$

can be easily computed as

$$\begin{aligned} \mu_{z_{it}|\cdot} &= \mu_{\mathbf{z}_i|\cdot}[t] + \Sigma_{\mathbf{z}_i|\cdot}[t, -t]\left(\Sigma_{\mathbf{z}_i|\cdot}[-t, -t]\right)^{-1}(\mathbf{z}_{i-t} - \mu_{\mathbf{z}_i|\cdot}[-t]), \\ \Sigma_{z_{it}|\cdot} &= \Sigma_{\mathbf{z}_i|\cdot}[t, t] - \Sigma_{\mathbf{z}_i|\cdot}[t, -t]\left(\Sigma_{\mathbf{z}_i|\cdot}[-t, -t]\right)^{-1}\Sigma_{\mathbf{z}_i|\cdot}[-t, t], \end{aligned}$$

where $\mu_{\mathbf{z}_i|\cdot}$ and $\Sigma_{\mathbf{z}_i|\cdot}$ are from (4.72) and (4.71) and $\mathbf{z}_{i-t}$ denotes the remaining observations of cluster i, after having z_{it} removed. Here, $\mu_{\mathbf{z}_i|\cdot}[t]$ is the tth element in $\mu_{\mathbf{z}_i|\cdot}$ and $\mu_{\mathbf{z}_i|\cdot}[-t]$ are the remaining $T_i - 1$ elements after removing the tth element. Furthermore, $\Sigma_{\mathbf{z}_i|\cdot}[-t,-t]$ is the $(T_i-1)\times(T_i-1)$ matrix derived from $\Sigma_{\mathbf{z}_i|\cdot}$ by removing its tth row and tth column and $\Sigma_{\mathbf{z}_i|\cdot}[-t,t]$ is the (T_i-1) vector derived from the tth column of $\Sigma_{\mathbf{z}_i|\cdot}$ by eliminating the tth row term.
Most importantly from a computational point of view, there is no need to invert all the matrices $\Sigma_{\mathbf{z}_i|\cdot}[-t,-t]$. These inverses can be derived from the 'global' inverse $S := \Sigma^{-1}_{\mathbf{z}_i|\cdot}$ since it holds

$$\left(\Sigma_{\mathbf{z}_i|\cdot}[-t,-t]\right)^{-1} = S[-t,-t] - S[t,-t]S[t,-t]^{\mathrm{T}}/S[t,t].$$

Alternatively, the results of Robert (1995) could be directly applied to the marginal multivariate distribution

$$p(\mathbf{z}|\mathbf{y}, D, \rho, \boldsymbol{\xi}, \boldsymbol{\gamma}) \propto \mathcal{N}(0, \Omega^*) I(\mathbf{y}, \mathbf{z}),$$

where $I(\mathbf{y}, \mathbf{z})$ may denote the appropriate indicator function. However, this approach does not exploit the block-diagonal structure in the error matrix Ω^*. The here proposed two step approach via the leave-one-block-out procedure is substantially more efficient since computation of the univariate moments is then based on $T_i \times T_i$ matrices while the alternative strategy uses $(N-1)\times(N-1)$ matrices.

The full conditionals of the random intercept and the inverse covariance of the random effect is Gaussian and Wishart, respectively. Both are easy to sample from their respective full conditionals

$$\begin{aligned} p(\mathbf{b}|\mathbf{z}, D, \rho, \boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\gamma}) \quad &= \quad \mathcal{N}(\mu_{\mathbf{b}|\cdot}, \Sigma_{\mathbf{b}|\cdot}) \\ &\text{where} \quad \mu_{\mathbf{b}|\cdot} = \Sigma_{\mathbf{b}|\cdot} W^{\mathrm{T}} \Omega^{-1} (\mathbf{z} - \Phi_{\boldsymbol{\gamma}} \boldsymbol{\omega}_{\boldsymbol{\gamma}}) \\ &\qquad\quad \Sigma_{\mathbf{b}|\cdot} = \left(D^{-1}\mathbf{I} + W^{\mathrm{T}} \Omega^{-1} W\right)^{-1}, \end{aligned} \tag{4.73}$$

and

$$p(D^{-1}|\mathbf{b}) \quad = \quad \mathrm{W}(r_0 + n, (R_0^{-1} + \mathbf{b}^{\mathrm{T}}\mathbf{b})^{-1}). \tag{4.74}$$

Ideally, the autocorrelation parameter of the latent responses is sampled from its full conditional distribution

$$p(\rho|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\omega}, \mathbf{b}, \boldsymbol{\gamma}) \propto \Psi(\rho) \times \mathcal{N}(\mu_{\rho|\cdot}, \Sigma_{\rho|\cdot}) I(-1 < \rho < 1). \tag{4.75}$$

This is, however, not a standard density and thus a Metropolis Hastings (MH) step is necessary in order to draw samples from (4.75).
It is useful to define the following quantities, which is here done specifically for the AR(1) error correlation structure: $e_{it} = z_{it} - \Phi_{\boldsymbol{\gamma}}(\xi_{it})\boldsymbol{\omega}_{\boldsymbol{\gamma}} - b_i$, which is equivalent to ϵ_{it} from the response model (4.58) if there are no lagged dependent variables in the model as it is the case here; $\mathbf{e}_i = (e_{i2}, \ldots, e_{iT_i})^{\mathrm{T}}$ and $\mathbf{e} = (\mathbf{e}_1^{\mathrm{T}}, \ldots, \mathbf{e}_n^{\mathrm{T}})^{\mathrm{T}}$ are column vectors; E is a column vector of length $(N - n)$ containing the one order lag of the elements in $\mathbf{e}$, i.e. $E = \big((e_{11}, \ldots, e_{1(T_1-1)}), \ldots, (e_{n1}, \ldots, e_{n(T_n-1)})\big)^{\mathrm{T}}$.
The stationary covariance of the AR(1) process is denoted by $\Omega_{AR(1)} = 1/(1 - \rho^2)$. The moments of the Gaussian part in the full conditional (4.75) are then computable as

$$\begin{aligned} \Sigma_{\rho|\cdot} &= (P_0^{-1} + E^{\mathrm{T}}E)^{-1} \\ \mu_{\rho|\cdot} &= \Sigma_{\rho|\cdot}(P_0^{-1}\rho_0 + E^{\mathrm{T}}\mathbf{e}) \end{aligned} \tag{4.76}$$

and furthermore

$$\Psi(\rho) = \mid \Omega_{AR(1)} \mid^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{e_{i1}^2}{\Omega_{AR(1)}}\right).$$

Firstly, a proposal draw ρ' is generated from the truncated normal part of (4.75), i.e. $\mathcal{N}(\mu_\rho, \Sigma_{\rho|\cdot})I(-1 < \rho < 1)$. This is subsequently accepted with probability

$$\alpha = \min\left\{1, \frac{\Psi(\rho')}{\Psi(\rho)}\right\}.$$

The necessary amendments to suit the general case of AR(p) processes and lagged dependent variables are presented by Chib & Jeliazkov (2006).

The full conditionals of the covariate model parameters underly only minor

modifications compared to their cross-sectional specifications in Section 4.1.4. They are given as

$$\begin{aligned}
p(\mu_\xi|\boldsymbol{\xi},\sigma_\xi^2,\rho_\xi) &= \mathcal{N}\left(\mu_{\mu_\xi|\cdot},\sigma^2_{\mu_\xi|\cdot}\right)\\
\mu_{\mu_\xi|\cdot} &= \frac{\left(\sum_{i=1}^n\sum_{t=1}^{T_i}\xi_{it}\right)g^2 + f\frac{\sigma_\xi^2}{1-\rho_\xi^2}}{Ng^2+\frac{\sigma_\xi^2}{1-\rho_\xi^2}},\\
\sigma^2_{\mu_\xi|\cdot} &= \frac{\frac{\sigma_\xi^2}{1-\rho_\xi^2}g^2}{Ng^2+\frac{\sigma_\xi^2}{1-\rho_\xi^2}}\\
p(\sigma_\xi^2|\mathbf{x},\boldsymbol{\xi},\mu_\xi,\rho_\xi,\lambda) &= IG\left(A_{\xi|\cdot},\frac{1}{B_{\xi|\cdot}}\right)\\
A_{\xi|\cdot} &= A_\xi + \frac{1}{2}\sum_{i=1}^n\sum_{t=1}^{T_i} m_{it} + \frac{N}{2},\\
B_{\xi|\cdot} &= B_\xi^{-1} + \frac{\lambda(1-\rho_\xi^2)}{2(1-\lambda)}\sum_{i=1}^n\sum_{t=1}^{T_i}\sum_{j=1}^{m_{it}}(x_{itj}-\xi_{it})^2\\
&\quad + \frac{1}{2}(\boldsymbol{\xi}-\mu_\xi)^{\mathrm{T}}\Omega_{\boldsymbol{\xi}}^{-1}(\boldsymbol{\xi}-\mu_\xi),
\end{aligned}$$

where the matrix Ω_ξ is only one part of the covariance matrix of the latent ξ_i's, cf. (4.64). The attenuation factor λ is defined as in (4.65) and again sampled in a discrete Gibbs step as in the cross-sectional case (cf. Section 4.1.4) based on the full conditional

$$\begin{aligned}
p(\lambda|\boldsymbol{\xi},\rho_\xi,\sigma_\xi^2) &\propto I(\lambda_L<\lambda<\lambda_H)\left(\frac{\lambda}{1-\lambda}\right)^a\exp\left(-\frac{\lambda(1-\rho_\xi^2)\cdot b}{2(1-\lambda)\sigma_\xi^2}\right)\\
a &= \sum_{i,t} m_{it}/2\\
b &= \sum_{i,t} m_{it}(\overline{x}_{it}-\xi_{it})^2 + \widehat{\sigma_\delta^2}\sum_{i,t}(m_{it}-1),
\end{aligned} \tag{4.77}$$

with $\overline{x}_{it} = \frac{1}{m_{it}}\sum_{j=1}^{m_{it}} x_{itj}$ denoting the average over a subjects's replicates at a certain time point.

Finally, the autocorrelation parameter of the latent covariate ρ_ξ has to be sampled under usage of a Metropolis Hastings (MH) step similar to the sampling scheme of ρ described above. Its full conditional distribution is specified by

$$p(\rho_\xi|\boldsymbol{\xi}, \mu_\xi, \sigma_\xi^2) \propto \Psi(\rho_\xi) \times \mathcal{N}(\mu_{\rho_\xi|\cdot}, \Sigma_{\rho_\xi|\cdot}) I(-1 < \rho_\xi < 1). \tag{4.78}$$

In analogy to above, it is again useful to introduce the following quantities (specifically for the AR(1) case): $e_{it} = \xi_{it} - \mu_\xi$, $\mathbf{e}_i = (e_{i2}, \ldots, e_{iT_i})^{\mathrm{T}}$, $\mathbf{e} = (\mathbf{e}_1^{\mathrm{T}}, \ldots, \mathbf{e}_n^{\mathrm{T}})^{\mathrm{T}}$; the column vector E is of length $(N - n)$ and contains the lag of order one of the elements in $\mathbf{e}$, i.e. $E = \big((e_{11}, \ldots, e_{1(T_1-1)}), \ldots, (e_{n1}, \ldots, e_{n(T_n-1)})\big)^{\mathrm{T}}$. The stationary covariance matrix of the AR(1) process is $\Omega_{\xi AR(1)} = 1/(1 - \rho_\xi^2)$. The terms in the full conditional (4.78) can then be computed as

$$\begin{aligned} \Sigma_{\rho_\xi|\cdot} &= \left(P_{\xi 0}^{-1} + \frac{E^{\mathrm{T}} E}{\sigma_\xi^2} \right)^{-1} \\ \mu_{\rho_\xi|\cdot} &= \Sigma_{\rho_\xi|\cdot} \left(P_{\xi 0}^{-1} \rho_{\xi 0} + \frac{E^{\mathrm{T}} \mathbf{e}}{\sigma_\xi^2} \right) \end{aligned}$$

and furthermore

$$\Psi(\rho_\xi) = \mid \Omega_{\xi AR(1)} \sigma_\xi^2 \mid^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{e_{i1}^2}{\Omega_{\xi AR(1)} \sigma_\xi^2} \right). \tag{4.79}$$

A proposal draw ρ'_ξ is generated from the truncated normal density $\mathcal{N}(\mu_{\rho_\xi|\cdot}, \Sigma_{\rho_\xi|\cdot})\ I(-1 < \rho_\xi < 1)$ and subsequently accepted with probability

$$\alpha = \min\left\{ 1, \frac{\Psi(\rho'_\xi)}{\Psi(\rho_\xi)} \right\}. \tag{4.80}$$

The necessary amendments to suit the general case of AR(p) processes are analog to the AR(p) serial correlation of the latent responses, which are given in the article of Chib & Jeliazkov (2006).

Somewhat more complicated is the updating scheme of the latent covariate observations ξ_{it}. The contribution of a single ξ_{it} to the likelihood can

no longer be separated out because of the correlation structure in the data. While in the cross-sectional case the likelihood contribution of a single latent observation and its proposal is calculated to compute the acceptance probability (cf. (4.45)), in the longitudinal context the observations within a cluster contribute jointly to the likelihood. It is thus reasonable to generate 'proposal blocks' of latent observations $\boldsymbol{\xi}_i'$ and compute the ratio of the joint likelihood contributions of these proposed cluster values to the current cluster values $\boldsymbol{\xi}_i$.

Like in the cross-sectional case (cf. 4.46), a multivariate symmetric random walk proposal is specified. This is a Gaussian with the current sample $\boldsymbol{\xi}_i$ as mean vector and with covariance matrix being half the conditional covariance matrix of $\boldsymbol{\xi}_i$ given the observed values $\mathbf{x}_i$. Thus, a proposal $\boldsymbol{\xi}_i'$ is generated from the distribution

$$\begin{aligned} p(\boldsymbol{\xi}_i') \quad &= \quad \mathcal{N}\left(\mu_{\boldsymbol{\xi}_i'}, \frac{1}{2}\Sigma_{\boldsymbol{\xi}_i'}\right), \\ &\text{where} \quad \mu_{\boldsymbol{\xi}_i'} = \boldsymbol{\xi}_i \\ &\qquad\quad \Sigma_{\boldsymbol{\xi}_i'} = \left(\frac{m_i}{\sigma_\delta^2}\mathbf{I} + \Sigma_{\boldsymbol{\xi}_i}^{-1}\right)^{-1}. \end{aligned} \tag{4.81}$$

Here, m_i denotes the number of replicate measurements, where it is assumed that cluster i has the same number of replicate measurements at every time point $t = 1, \ldots, T_i$. If a different number of replicate measurements are taken at the several time points, then $m_i\sigma_\delta^{-2}\mathbf{I}$ in (4.81) is substituted by $diag(m_{i1}, \ldots, m_{iT_i})\sigma_\delta^{-2}$. In the proposal distribution (4.81), the matrix $\Sigma_{\boldsymbol{\xi}_i}^{-1}$ is as previously defined in (4.64) and the measurement error variance σ_δ^2 is calculated from the current samples of σ_ξ^2, ρ_ξ and λ, which gives $\sigma_\delta^2 = \frac{\sigma_\xi^2(1-\lambda)}{(1-\rho_\xi^2)\lambda}$, cf. (4.40).

After the proposal is generated from (4.81) it is accepted with probability

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\xi}_i'|\mathbf{z}, \mathbf{x}_i, \rho, D, \mu_\xi, \sigma_\xi^2, \rho_\xi, \lambda, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})}{p(\boldsymbol{\xi}_i|\mathbf{z}, \mathbf{x}_i, \rho, D, \mu_\xi, \sigma_\xi^2, \rho_\xi, \lambda, \boldsymbol{\omega}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})}\right\}. \tag{4.82}$$

The required conditional density of the true covariate observations within

cluster i is given by

$$p(\boldsymbol{\xi}_i|\mathbf{z}, \mathbf{x}_i, \rho, D, \mu_\xi, \sigma_\xi^2, \rho_\xi, \lambda, \boldsymbol{\omega}_\gamma, \boldsymbol{\gamma}) \propto \exp\left(-\frac{1}{2} \left(\mathbf{z}_i - \Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}}\right)^{\mathrm{T}} \Omega_i^{*^{-1}} \left(\mathbf{z}_i - \Phi_{\boldsymbol{\gamma}}(\boldsymbol{\xi}_i)\boldsymbol{\omega}_{\boldsymbol{\gamma}}\right) - \frac{(1-\rho_\xi^2)\lambda}{2\sigma_\xi^2(1-\lambda)} \sum_{t=1}^{T_i} \sum_{j=1}^{m_{it}} (x_{itj} - \xi_{it})^2 - \frac{1}{2} \left(\boldsymbol{\xi}_i - \mu_\xi\right)^{\mathrm{T}} \Sigma_{\xi_i}^{-1} \left(\boldsymbol{\xi}_i - \mu_\xi\right) \right) \quad (4.83)$$

The triple subscript denotes the cluster $i = 1, \ldots, n$, the time $t = 1, \ldots, T_i$ and the replicate measurements of an cluster at one point in time $j = 1, \ldots, m_{it}$. The vector $\mathbf{x}_i$ denotes here, and only here in (4.83), all available replicates m_{it} for all time points t for this person i – only these are required in the full conditional distribution of $\boldsymbol{\xi}_i$.
The conditional density (4.83) is easily evaluated at both positions $\boldsymbol{\xi}_i'$ and $\boldsymbol{\xi}_i$ to give the desired acceptance probability (4.82). Now, for each cluster i only a single uniform random number is generated to decide whether to accept or reject the complete proposed cluster. If the candidate values $\boldsymbol{\xi}_i'$ are rejected, all current values in $\boldsymbol{\xi}_i$ are repeated as the next values of the MCMC sample. A potential improvement may lie in viewing the kernel parameter η of the radial basis functions (cf. (2.3)) also as unknown parameter and insert a further MH-step to draw samples for η. In the data examples of the next section an estimate for η comes from the naive RVM neither considering measurement error nor the aspects of longitudinal data. Doing so, is in analogy to the MCMC approach in the cross-sectional case.

Firstly, starting values for the unknown parameters are given and then after a burn-in period of 8000 draws for each parameter (vector) another 12000 runs are executed, where samples are collected for parameter estimation. Given a previously unseen ξ^*, predicting the probability $P(y^* = 1|\xi^*, \mathbf{y})$ is accomplished by firstly averaging over the samples for the linear predictor $z^* := \Phi(\xi^*)\boldsymbol{\omega}$ from each MCMC run (after the burn-in period is finished)

$$\widehat{\mathbb{E}}(z^*|\mathbf{y}) \;=\; \frac{1}{12000} \sum_{j=1}^{12000} (\Phi(\xi^*)\boldsymbol{\omega}^{[j]}|\mathbf{y}),$$

which is the histogram estimator for the posterior mean of the linear predictor (cf. (2.30)). Then, this estimate is transformed to the desired estimate for the probability by using the cumulative distribution function of a standard Gaussian density with mean zero and variance one to give

$$\widehat{P}(y^* = 1|\xi^*, \mathbf{y}) = \Phi\left(\widehat{\mathbb{E}}(z^*|\mathbf{y}), 0, 1\right),$$

where $\Phi(a, b, c)$ denotes the Gaussian cumulative distribution function at value a with mean b and variance c.
The Bayesian hope lies in the samples $\mathbf{z}^{[j]}$, $\rho^{[j]}$, $\mathbf{b}^{[j]}$, $D^{[j]}$, $\boldsymbol{\omega}^{[j]}$, $\boldsymbol{\xi}^{[j]}$, $\rho_\xi^{[j]}$, $\mu_\xi^{[j]}$, $(\sigma_\xi^2)^{[j]}$, $\lambda^{[j]}$, $\boldsymbol{\gamma}^{[j]}$, $j = 1, \ldots, 12000$ behaving as if coming from their joint posterior distribution and thus $\widehat{\mathbb{E}}(z^*|\mathbf{y})$ being a decent approximation to the mean of the marginal posterior distribution for the linear predictor z^*. The estimate $\widehat{P}(y^* = 1|\xi^*, \mathbf{y})$ obtained here is not equivalent to the posterior mean estimate of $P(y^* = 1|\xi^*, \mathbf{y})$, but rather a transformation of the mean estimate $\widehat{\mathbb{E}}(z^*|\mathbf{y})$, cf. end of Section 4.1.4 and Section 4.2.2.

4.4.3 A few data examples

The flexible modeling of longitudinal binary data under covariate measurement error is rather complex and requires the specification of a load of parameters that have to be estimated from the MCMC draws. The idea of an extensive simulation study is discarded here in favor of a few relevant data examples with an increased number of MCMC runs accounting for the complexity of the model. That is, a burn-in period of 8000 runs was taken and parameter estimation relied on the following 12000 runs.
The binary responses are generated from a probit model with linear predictor $m(\xi)$ and characterized as follows

Case 1: An oscillating function of the covariate with

$$m(\xi) = \begin{cases} -0.9 + 3\sin(5\xi)/(5\xi) & \xi \neq 0 \\ 2.1 & \text{else} \end{cases},$$

$n = 100$, $T_i = 10, i = 1, \ldots, n$, $\rho = -0.6$,$D = 0.3$, $\sigma_\delta^2 = 0.5^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5$, $\rho_\xi = 0.5$

Case 1 is further investigated under varying specifications for the prior distributions in order to check, whether the results are robust against modifications (cf. Table 4.5).

Case 2: Same as case 1 but now with $n = 200$ and $T_i = 5, i = 1, \ldots, n$

Case 3: Same as case 1 but now with $n = 100$, $T_i = 10, i = 1, \ldots, n$, $\rho = 0.6$, $D = 0.5$, $\sigma_\delta^2 = 0.5^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5$, $\rho_\xi = -0.5$

Before the results for these cases are given, it is emphasized that in all three cases ρ and ρ_{xi} carry contrary signs – a fact that makes the data difficult to analyze since it might blur the true impact of the covariate on the response. This should be kept in mind when interpreting the results.

The Results

Case 1:

The first data case represents the prototype of a small-sized study, where only $n = 100$ subjects are followed over ten time points. The measurement error is small while autocorrelation of the responses and the true latent covariate observations is pronounced as well as the impact of the random effect.

Figure 4.7 displays the model fit and shows how well the underlying true probability function is recovered by the sampling strategy described above, here termed '$\mathbf{RVM_{MCMC}}$'.

The 90% prediction band is constructed from the 0.05 and 0.95 quantile of the empirical posterior density of the parameter $\boldsymbol{\omega}$. Although it does not contain the true probability at every covariate point, the rough course of this function is very well discovered in the face of autocorrelation, person specific effects and covariate measurement error. A refined estimation routine for the basis function parameter η, determining the width of a basis, may lead to a further improvement.

The inspection of the sampling paths for some key parameters in Figure 4.8

indicates no severe problems like paths getting stuck in certain regions of the parameter space or revealing a trend as if working their way to a more representative part of the posterior distribution. However, the coarse sampling path of a randomly chosen latent observation (here ξ_2) and the sampling path of the number of basis functions in the model witness a lower acceptance rate than e.g. for both autocorrelation parameters that are just as well drawn in MH-steps.
Figure 4.9 depicts the histogram of the 12000 samples used for inference. The empirical marginal posterior distributions are symmetric (except for the number of basis functions) and thus the posterior mean estimate is expected to be a reliable summary statistic of the distribution. The histogram estimator (2.30) based on these samples gave the following posterior mean estimates (true values in brackets): $\mu_\xi = -0.0223(0.0)$, $\widehat{\sigma_\xi^2} = 1.4069(1.5)$, $\widehat{\rho_\xi} = 0.4673(0.5)$, $\widehat{\sigma_\delta^2} = 0.2411(0.25)$, $\widehat{\rho} = -0.3312(-0.6)$ and $\widehat{D} = 0.1576(0.3)$. The mean squared error for this example is 0.0103 and is monitored under varying prior specifications in Table 4.5.

Case 2:
The second data case mirrors the problem of having rather few measurements for each individual cluster. Here, $n = 200$ cluster/individuals are followed over a time period of only five units. The remaining setup is unchanged to case 1.
The model fit in Figure 4.10 indicates again a good quality in detecting at least roughly the tenor of the underlying true probability function. The sampling paths of almost all parameters in Figure 4.11 are located in vicinity to the true values and do not display a pattern as if working their way to another part of the posterior distribution.
Figure 4.12 depicts the histogram of the 12000 samples used for inference. The histogram estimator (2.30) based on these samples gave the following posterior mean estimates (true values again in brackets): $\widehat{\mu_\xi} = -0.0913(0.0)$, $\widehat{\sigma_\xi^2} = 1.5249(1.5)$, $\widehat{\rho_\xi} = 0.4221(0.5)$, $\widehat{\sigma_\delta^2} = 0.2363(0.25)$, $\widehat{\rho} = -0.3953(-0.6)$ and $\widehat{D} = 0.1281(0.3)$.

MSE under alternative prior specifications				
	Case 1	Case 1 b)	Case 1 c)	Case 1 d)
MSE	0.0103	0.0126	0.0145	0.0115
$\omega_j \sim \mathcal{N}(0, v)$	$v = 100$	$v = 1000$	$v = 1000$	$v = 1000$
$\mu_\xi \sim \mathcal{N}(f, g)$	$f = 0$, $g = 100$	$f = 0$, $g = 1000$	$f = 0$, $g = 1000$	$f = 0$, $g = 1000$
$\sigma_\xi^2 \sim IG(A_\xi, B_\xi)$	$A_\xi = 1$, $B_\xi = 1$	$A_\xi = 1$, $B_\xi = 100$	$A_\xi = 1$, $B_\xi = 100$	$A_\xi = 100$, $B_\xi = 1$
$\rho_\xi \sim \mathcal{N}(\rho_{\xi 0}, P_{\xi 0}) I_{[-1,+1]}$	$\rho_{\xi 0} = 0$, $P_{\xi 0} = 0.25$	$\rho_{\xi 0} = 0$, $P_{\xi 0} = 0.25$	$\rho_{\xi 0} = 0.5$, $P_{\xi 0} = 1$	$\rho_{\xi 0} = -0.5$, $P_{\xi 0} = 1$
$\rho \sim \mathcal{N}(\rho_0, P_0) I_{[-1,+1]}$	$\rho_0 = 0$, $P_0 = 0.25$	$\rho_0 = 0$, $P_0 = 0.25$	$\rho_0 = 0.5$, $P_0 = 1$	$\rho_0 = -0.5$, $P_0 = 1$
$D^{-1} \sim W(r_o, R_0)$	$r_0 = 2$, $R_0 = 2.5$	$r_0 = 2$, $R_0 = 2.5$	$r_0 = 4$, $R_0 = 1.25$	$r_0 = 6$, $R_0 = 0.33$

Table 4.5: The data from case 1 are also analyzed under alternative prior specifications. The mean squared error criterion indicates, except for case 1 c), only a slight impact of the modified prior specifications on the final estimation. Here, the normalizing constant is suppressed in the priors over the autocorrelation parameters and $I_{[-1,+1]}$ stands for the indicator function.

Case 3:

Similar to case 1, ten dynamic measurements T_i are available for each of the $n = 100$ individuals. However, here the direction of autocorrelation has been switched for response and covariate observations and the person specific effect is slightly increased compared to case 1.

The model fit in Figure 4.13 shows that the $\mathbf{RVM_{MCMC}}$ is able to reconstruct roughly the course of the true underlying probability. Compared to the first case, this reconstruction is less successful which is probably rather attributable to the concrete data samples than to the slight modifications in the underlying data generation setup. The sampling paths in Figure 4.8 do not indicate any severe problems.

The histograms of the 12000 samples used for inference in Figure 4.15 are symmetric (except for the number of basis functions) and thus the posterior

mean estimate is expected to be a reliable summary statistic of the distribution. The histogram estimator based on these samples gave the following posterior mean estimates (true values in brackets): $\widehat{\mu_\xi} = -0.0525(0.0)$, $\widehat{\sigma_\xi^2} = 1.5438(1.5)$, $\widehat{\rho_\xi} = -0.4773(-0.5)$, $\widehat{\sigma_\delta^2} = 0.2585(0.25)$, $\widehat{\rho} = 0.4593(0.6)$ and $\widehat{D} = 0.2222(0.5)$.

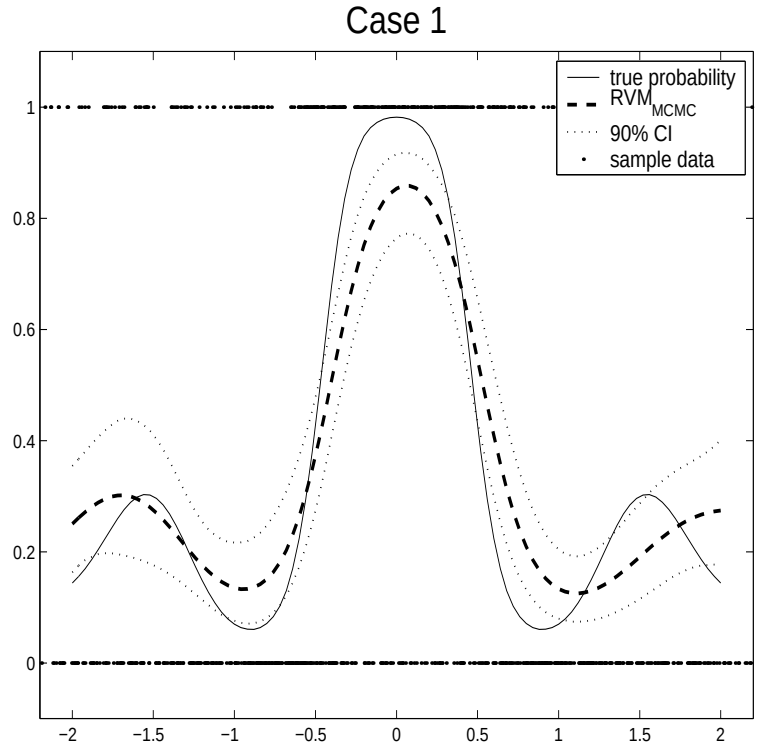

Figure 4.7: Final model fit for case 1. This prediction is calculated from the posterior mean estimate for the linear predictor $z^* := \Phi(\xi^*)\boldsymbol{\omega}$ *at 101 grid points* $-2 \leq \xi^* \leq 2$*. The pointwise 90% credible intervals are calculated from the 0.05 and 0.95 quantiles of the empirical posterior distribution of the linear predictor.*

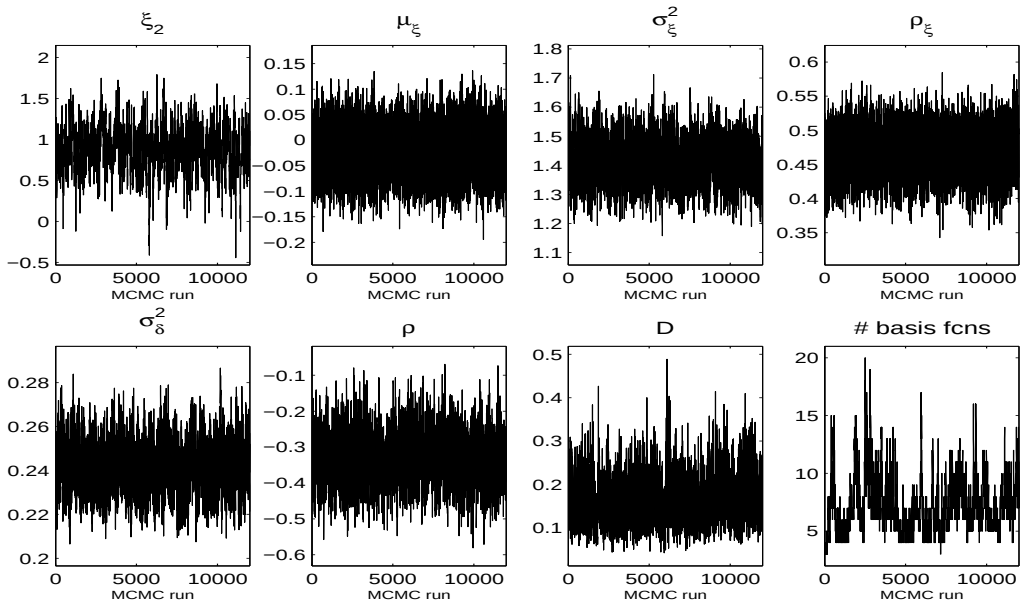

Figure 4.8: Sampling paths of key parameters for case 1.

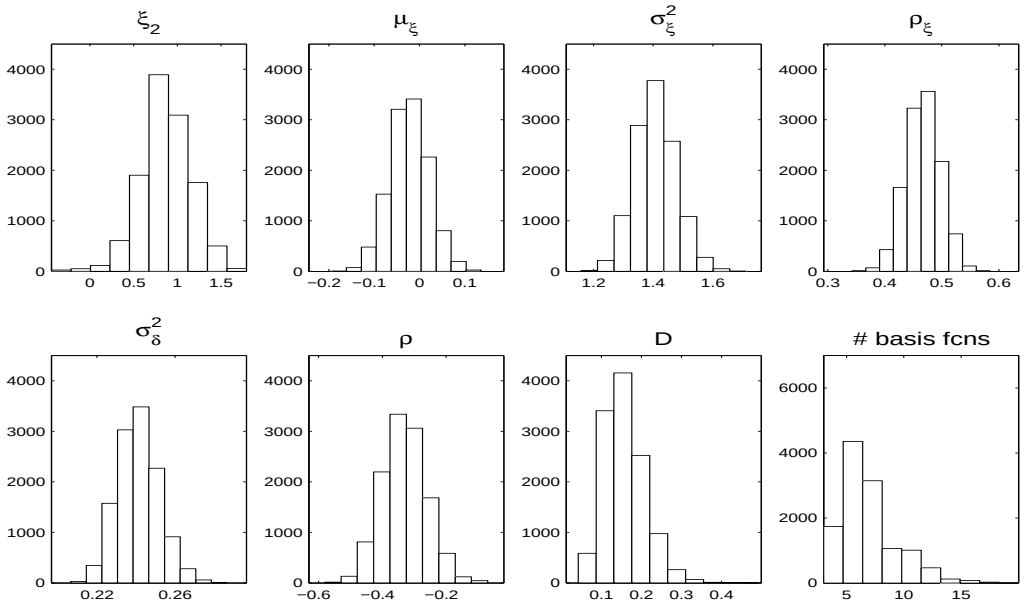

Figure 4.9: Histograms of MCMC samples for key parameters for case 1.

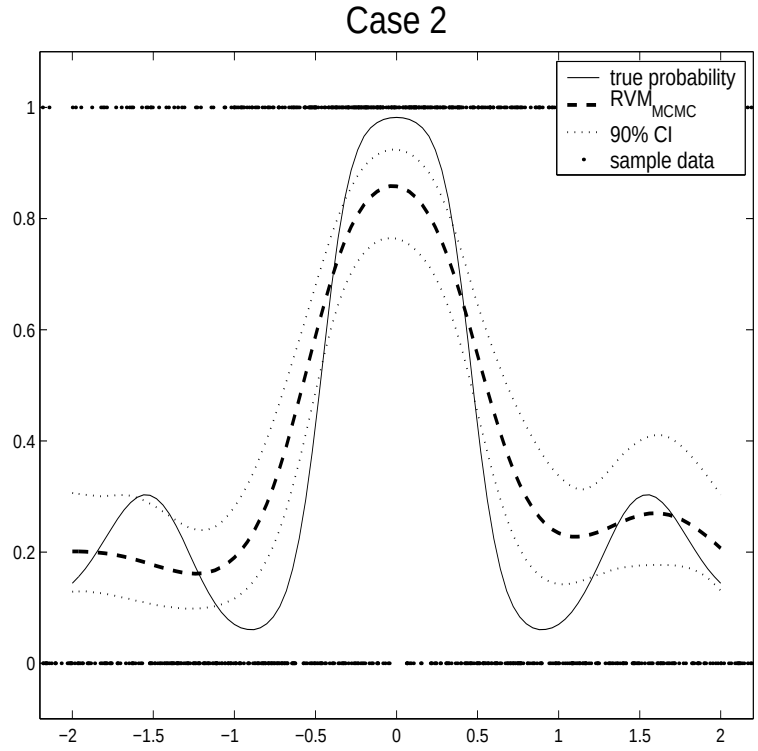

Figure 4.10: Final model fit for case 2. This prediction is calculated from the posterior mean estimate for the linear predictor $z^* := \Phi(\xi^*)\boldsymbol{\omega}$ *at 101 grid points* $-2 \leq \xi^* \leq 2$*. The pointwise 90% credible intervals are calculated from the 0.05 and 0.95 quantiles of the empirical posterior distribution of the linear predictor.*

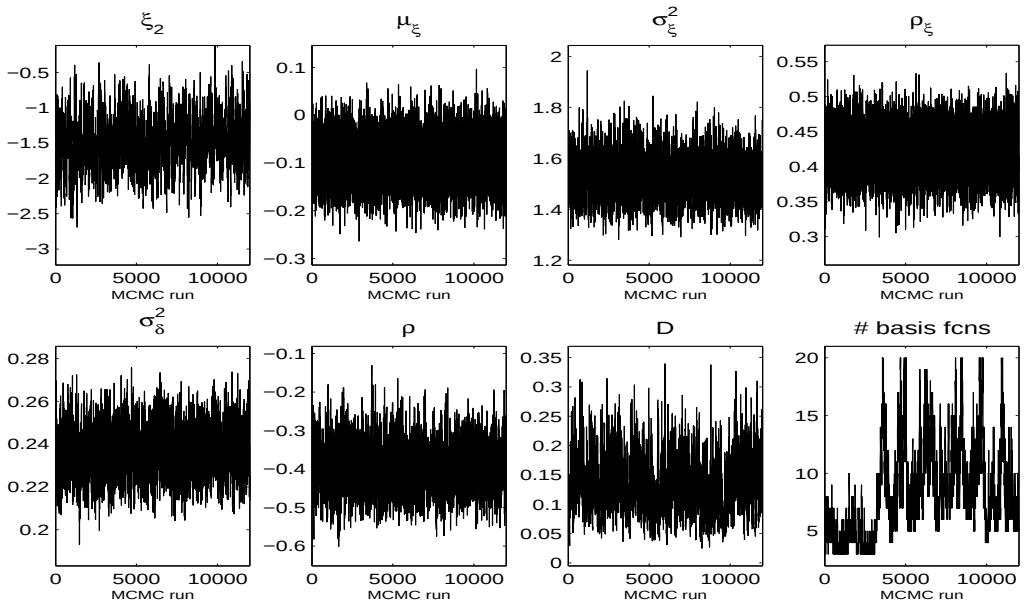

Figure 4.11: Sampling paths of key parameters for case 2.

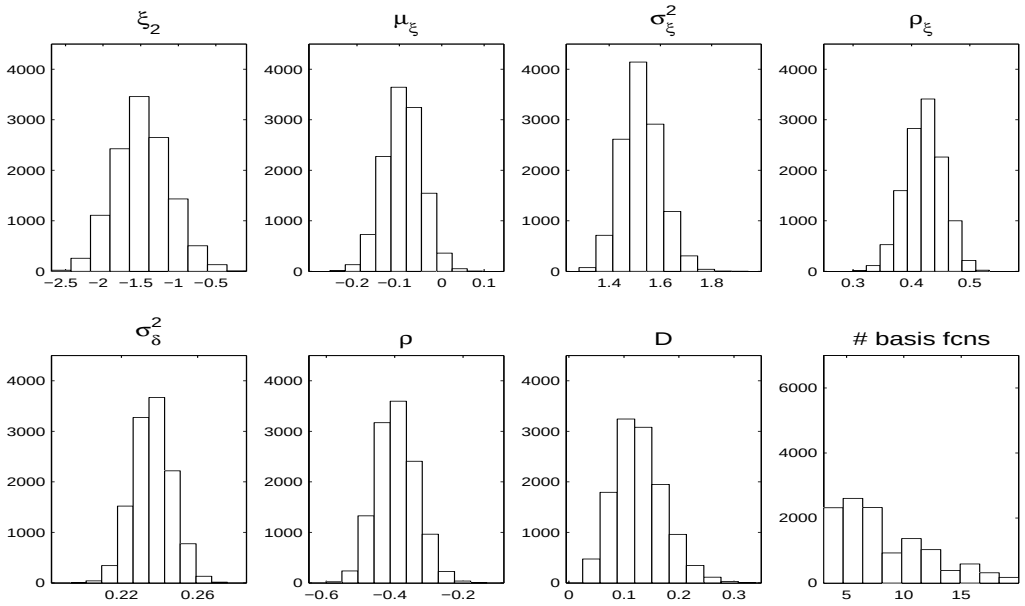

Figure 4.12: Histograms of MCMC samples for key parameters for case 2.

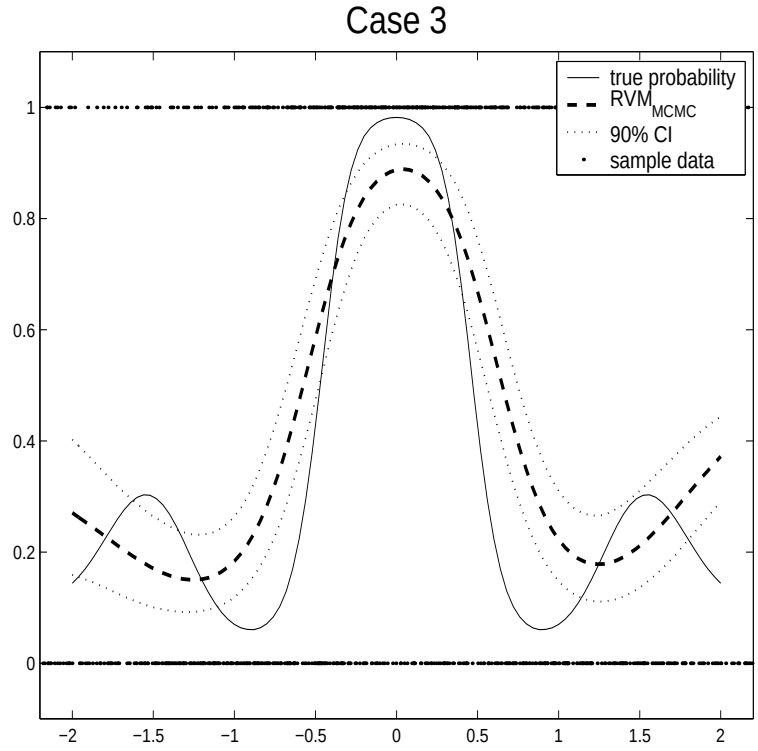

Figure 4.13: Final model fit for case 3. This prediction is calculated from the posterior mean estimate for the linear predictor $z^* := \Phi(\xi^*)\boldsymbol{\omega}$ *at 101 grid points* $-2 \leq \xi^* \leq 2$*. The pointwise 90% credible intervals are calculated from the 0.05 and 0.95 quantiles of the empirical posterior distribution of the linear predictor.*

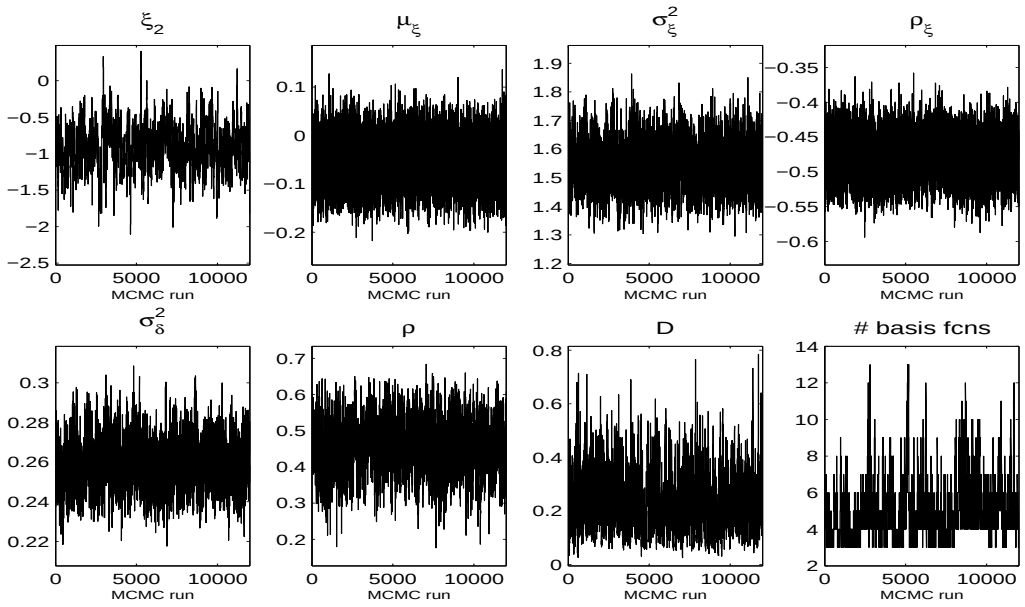

Figure 4.14: Sampling paths of key parameters for case 3.

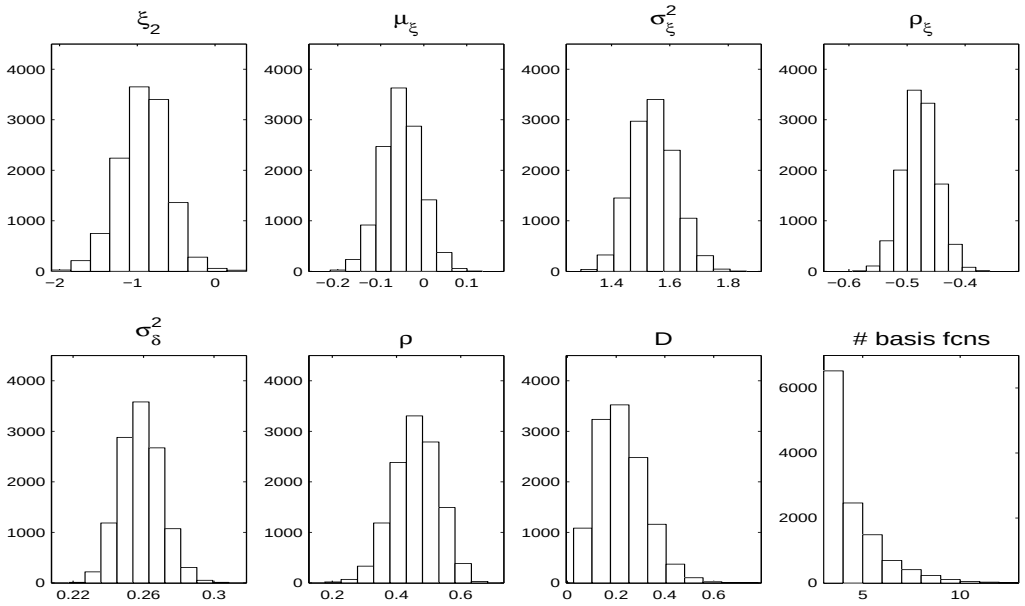

Figure 4.15: Histograms of MCMC samples for key parameters for case 3.

Chapter 5

Covariate measurement error in flexible Poisson regression

This chapter considers the flexible Poisson regression model under covariate measurement error. Chapter 4 already developed strategies of how to correct for error in the flexible binary regression. It is revealing to see, how a part of these methods can be generalized, usually under only minor modifications, to suit Poisson responses and how they behave in this case. This chapter is mainly left with giving the relevant modifications and computational details and concludes with a simulation study.
To keep this chapter self-contained the Poisson RVM regression model is briefly recalled, before Section 5.1 surveys the employed correction methods together with the necessary amendments.

Resorting to chapter 2, the Poisson RVM is of the form

$$Y = G(\Phi(\xi)\boldsymbol{\omega}) + \epsilon, \tag{5.1}$$

where $Y \in \mathbb{N}_0$ and $\mathbb{E}(\epsilon) = 0$. The response function is typically chosen as $G(z) = \exp(z)$, and (5.1) is then called a log link Poisson model.
All methods described in this chapter employ the Gaussian prior distribution

over the fundamental mean model parameters

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{j=0}^{J} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{\alpha_j}{2}\omega_j^2\right). \tag{5.2}$$

Further, Gamma hyperpriors are specified over those hyperparameters collected in $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_J)^{\mathrm{T}}$ and also over σ^2 if dispersion is included in the model (cf. Section 2.1.1).
Thus, the 'ideal mean model' and the 'ideal variance model' are here given as

$$\begin{aligned} \mathbb{E}(Y|\xi) &= G\left(\Phi(\xi)\boldsymbol{\omega}\right) && (5.3) \\ \mathbb{V}(Y|\xi) &= \sigma^2 G\left(\Phi(\xi)\boldsymbol{\omega}\right), && (5.4) \end{aligned}$$

with dispersion parameter σ^2. The variance $\mathbb{V}(Y|\xi)$ is again heteroscedastic as in the binary case (cf. (4.3)). Throughout this chapter, the dispersion parameter $\sigma^2 \overset{!}{=} 1$, unless stated otherwise. Though the mean model parameters $\boldsymbol{\omega}$ are random in a Bayesian context they are again suppressed on the left hand side of (5.3) and (5.4) for notational clarity and will be throughout this chapter for the sake of clarity.

5.1 The arsenal of correction methods

The correction for covariate measurement error in Poisson regression is like in binary regression a more demanding task than in the Gaussian case. It is complicated by the non-linear response function $G(z)$ in the ideal models (5.3) and (5.4).
Basis calibration, as described earlier in Section 4.1.1, can be straightforwardly applied, but again merely approximates the observed mean model in terms of the true model parameters $\boldsymbol{\omega}$. The expanded basis function calibration, as introduced in Section 4.1.2, refines the approximation to the observed mean and variance model. However, it requires careful consideration of the specific form of the ideal moments (5.3) and (5.4) in the Poisson case. The

SIMEX (cf. Section 4.1.3) is easily adapted to the Poisson case. The three methods are compared to the naive RVM regression in a concluding simulation study. There, it is most interesting to see how the informational gain in the Poisson responses, in contrast to the binary responses, is responsible for the clear superiority of the correction methods compared to the naive analysis.
An MCMC approach to suit the Poisson regression model under covariate error is conceivable, e.g. via the Binomial approximation of the Poisson case (cf. Denison et al. (2002)). This is not realized, yet. This could, however, prove to be particularly rewarding in the case of longitudinal data, where the methodology from the longitudinal binary case (cf. Section 4.4) can be adopted. Such an approach is not yet discussed in the existing literature.

5.1.1 Basis function calibration

The core of basis function calibration is the replacement of the design matrix Φ, formulated in terms of the latent observations $\xi_i, i = 1, \ldots, N$, by its calibrated version Φ_c in the ideal mean model (5.3). This calibrated design matrix Φ_c is made up of the row vectors $\mu_{\Phi(\xi)|X} = \mathbb{E}(\Phi(\xi)|X)$ as described at great length in the Gaussian case (cf. Section 3.1.1).
Like in the binary case, this replacement does no longer yield the exact representation of $\mathbb{E}(Y|X)$ in terms of the fundamental model parameters $\boldsymbol{\omega}$, because

$$\mathbb{E}(Y|X) \quad \neq \quad G\left(\mu_{\Phi(\xi)|X}\boldsymbol{\omega}\right). \tag{5.5}$$

In Poisson regression, substituting Φ_c into Φ yields merely an working model for the observed mean.
Parameter estimation proceeds via Fisher scoring as described for the binary case in Section 4.1.1, but now with the exponential response function for $G(z)$. The usage of the modified basis functions should intuitively lead to larger estimates for the mean model parameters $\boldsymbol{\omega}$ (cf. Figure 3.1) and so ideally alleviate the oversmoothing that is typically inherent in the naive

analysis.
Hyperparameter estimation is done via the maximization of the marginal likelihood as depicted for the binary case in Section 4.1.1.

5.1.2 Expanded basis function calibration

The basic idea of expanded basis function calibration is to refine the approximation of the observed mean and variance model under retainment of the fundamental parameters of the ideal mean (5.3) and variance model (5.4). This concept utilizes the idea of basis function calibration and uses Taylor series expansion for the improved approximation to the observed moments. The underlying idea has bees discussed in detail in Section 4.1.2. From the theoretical aspect, the adoption for the Poisson case is straightforward, however, the technical details, required for a successful implementation, need somewhat more considerate explanation.

In the Poisson regression case (with dispersion parameter σ^2) the ideal mean and variance model are based on the true but latent covariate ξ

$$\begin{aligned} \mathbb{E}(Y|\xi) &= f(\Phi(\xi), \boldsymbol{\omega}) && (5.6) \\ \mathbb{V}(Y|\xi) &= \sigma^2 g^2(\Phi(\xi), \boldsymbol{\omega}, \theta) = \sigma^2 f(\Phi(\xi), \boldsymbol{\omega}), && (5.7) \end{aligned}$$

where $f(\Phi(\xi), \boldsymbol{\omega}) : \mathbb{R}^{(J+1)} \to \mathbb{R}$ is viewed as working on the domain of the individual basis functions i.e. mapping the row vector $\Phi(\xi)$ containing the values of all basis functions at position ξ to a scalar.

Like the structural quasi likelihood approach (cf. Section 3.1.2), expanded basis function calibration also requires the conditional mean and variance of the basis function vectors $\Phi(\xi)$ given X

$$\begin{aligned} \mathbb{E}(\Phi(\xi)|X) &= \mu_{\Phi(\xi)|X} \\ \mathbb{V}(\Phi(\xi)|X) &= \Sigma_{\Phi(\xi)|X}, \end{aligned}$$

to approximate the observed mean $\mathbb{E}(Y|X)$ and observed variance $\mathbb{V}(Y|X)$. Both quantities can be calculated with the aid of the formulas (3.8) and

(3.18) in Section 3.1.2
Again some notational shortcuts are introduced in order to make the formulas involved in this approach more lucid.

Insertion: Notational details

The ideal mean model (5.6) and the ideal variance model (5.7) only differ in the dispersion parameter σ^2. Hence, only half of the shortcuts, compared to the binary case in Section 4.1.2, are needed here:

$$\begin{aligned} f &:= f(\Phi(\xi), \boldsymbol{\omega}) = G(\Phi(\xi)\boldsymbol{\omega}) \\ f_\mu &:= f(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) = G(\mu_{\Phi(\xi)|X}\boldsymbol{\omega}) \\ f' &:= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})}{\partial \Phi(\xi)} \\ f'_\mu &:= f'(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) \\ f'' &:= \frac{\partial G(\Phi(\xi)\boldsymbol{\omega})}{\partial \Phi(\xi)^{\mathrm{T}} \Phi(\xi)} \\ f''_\mu &:= f''(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}). \end{aligned}$$

Here, all first derivatives are row vectors, while the second derivatives are matrices. Poisson regression adopts the exponential response function $G(z) = \exp(z)$ and the specific form of the derivatives for this response function will be given later.

Basis function calibration (cf. Section 3.1.1) substitutes the latent design matrix Φ by its calibrated version Φ_c, which yields the following approximations for the observed mean and variance model

$$\begin{aligned} \mathbb{E}(Y|X) &\approx f(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}) \\ \mathbb{V}(Y|X) &\approx \sigma^2 f(\mu_{\Phi(\xi)|X}, \boldsymbol{\omega}, \theta). \end{aligned}$$

Expanded basis function calibration now refines these approximations employed in basis function calibration by adopting Taylor series expansion around $\mu_{\Phi(\xi)|X}$. Expanding the Taylor series around $\mu_{\Phi(\xi)|X}$ is justified by

the 'basis function model'

$$\begin{aligned} \Phi(\xi) &= \mu_{\Phi(\xi)|X} + V \\ &\text{where} \quad \mathbb{E}(V|X) = 0, \mathbb{V}(V|X) = \Sigma_{\Phi(\xi)|X}. \end{aligned}$$

The improved mean model approximation is then derived under application of the law of iterated expectations, as described for the binary case in (4.15), as

$$\mathbb{E}(Y|X) \approx f_\mu + \frac{1}{2}\text{tr}\left(\Sigma_{\Phi(\xi)|X} f''_\mu\right), \tag{5.8}$$

where 'tr' denotes the trace function applied to the matrix product $(\Sigma_{\Phi(\xi)|X} f''_\mu)$. A refined approximation of the observed variance uses the variance decomposition formula, and is in analogy to the binary case (4.16) given by

$$\mathbb{V}(Y|X) \approx f'^{\text{T}}_\mu \Sigma_{\Phi(\xi)|X} f'_\mu + \sigma^2 f_\mu + \sigma^2 \frac{1}{2}\text{tr}\left(\Sigma_{\Phi(\xi)|X} f''_\mu\right). \tag{5.9}$$

The necessary first and second derivatives of f for the exponential response function $G(z) = \exp(z)$ are given as:

$$\begin{aligned} f' &= f\boldsymbol{\omega}^{\text{T}} \\ f'' &= -\boldsymbol{\omega} f' \end{aligned}$$

Since f'' depends on the parameters $\boldsymbol{\omega}$, recalculation of this matrix has to be performed in every step of the optimization algorithm, making the procedure computer intensive.

The approximations of the observed moments (5.8) and (5.9) are again not range preserving, since they might become negative. A range preserving alternative is discussed for the binary case in Section 4.1.2 and can be easily adopted for the Poisson case. Instead an alternative strategy is used, where the Fisher scoring for $\boldsymbol{\omega}$ is exited if the approximation for $\mathbb{E}(Y|X)$ becomes negative and further $\boldsymbol{\omega}$ updating is not performed, before a hyperparameter update is accomplished.

The approximated observed moments (5.8) and (5.9) replace their ideal counterparts in the penalized quasi score function

$$s^X(Y, X, \boldsymbol{\omega}) = \sum_{i=1}^{N} \frac{\delta\, \mathbb{E}^*(y_i|x_i)}{\delta \boldsymbol{\omega}} \left(\frac{y_i - \mathbb{E}^*(y_i|x_i)}{\mathbb{V}^*(y_i|x_i)} \right) - \boldsymbol{\omega} A. \tag{5.10}$$

Parameter estimation is performed via Fisher scoring (cf. Section 4.1.2) and involves, besides the score function (5.10), the expected Fisher matrix

$$F(\boldsymbol{\omega}) \;=\; \left(\Phi_c^{*\mathrm{T}} B_c^* \Phi_c^* + A \right). \tag{5.11}$$

The symbol $*$ indicates that the approximative moments (5.9) and (5.8) are involved here. The Fisher matrix (5.11) comprises the following two quantities

$$\Phi_c^* = \begin{pmatrix} \frac{\partial\, \mathbb{E}^*(y_1|x_1)}{\partial \boldsymbol{\omega}} \\ \frac{\partial\, \mathbb{E}^*(y_2|x_2)}{\partial \boldsymbol{\omega}} \\ \cdots \\ \frac{\partial\, \mathbb{E}^*(y_N|x_N)}{\partial \boldsymbol{\omega}} \end{pmatrix}, \quad B_c^* = \begin{pmatrix} \mathbb{V}^*(y_1|x_1)^{-1} & & \\ & \ddots & \\ & & \mathbb{V}^*(y_N|x_N)^{-1}. \end{pmatrix}$$

The matrix Φ_c^* is constructed from the vectors of derivatives of the observed mean model with respect to the weights (cf. (4.23)). The diagonal matrix B_c^* contains the observed variances for each individual. The subscript 'c' indicates that the calibrated design matrix Φ_c is involved here.

The posterior covariance matrix is approximated by the inverse expected Fisher matrix and thus the posterior moments of the parameter vector $\boldsymbol{\omega}$ are given as

$$\Sigma \;=\; F(\boldsymbol{\omega})^{-1} = \left(\Phi_c^{*\mathrm{T}} B_c^* \Phi_c^* + A \right)^{-1}, \tag{5.12}$$

$$\boldsymbol{\mu} \;=\; \Sigma \Phi_c^{*\mathrm{T}} B_c^* \mathbf{y}^*, \quad \text{where} \quad \mathbf{y}^* = \Phi_c^* \boldsymbol{\omega} + (\mathbf{y} - \mathbb{E}^*(\mathbf{y}|\mathbf{x})). \tag{5.13}$$

The working observations $\mathbf{y}^*$ include the column vector $\mathbb{E}^*(\mathbf{y}|\mathbf{x})$, which is based on the approximation in (5.8) with $\mathbf{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ and $\mathbf{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$.
The so computed standard errors, based on (5.12), are not properly corrected

for the inherent covariate measurement error, which is a open problem, yet.

According to Tipping (2001), the marginal likelihood of the working observations is then approximated by a Gaussian essentially following a Laplace approximation. Like in the binary case, the required working observations are here not derived from the 'iteratively weighted least squares' representation of the Fisher scoring (5.13), but (for computational reasons) defined as

$$\mathbf{y}^{**} \;:=\; \Phi_c \boldsymbol{\omega} + D^{-1}(\mathbf{y} - \mathbb{E}^*(\mathbf{y}|\mathbf{x})) \tag{5.14}$$

with the diagonal matrix D having elements

$$D_{ii} = \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})}{\partial(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})} \right).$$

For the binary case, this particular specification of the working observations is discussed at great length in the paragraph 'Estimating the hyperparameters' of Section 4.1.2. The expression $\mathbb{E}^*(\mathbf{y}|\mathbf{x})$ in (5.14) denotes again the approximation for the observed mean model (5.8) for all observations in the data set.
This yields the marginal likelihood of the working observations that will be used in the hyperparameter optimization

$$p(\mathbf{y}^{**}|\boldsymbol{\alpha}) \;=\; \mathcal{N}\left(0, B^{**} + \Phi_c A \Phi_c^{\mathrm{T}}\right).$$

The elements of the diagonal matrix B^{**} are given by

$$B^{**}_{ii} = \left(\frac{\partial G(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})}{\partial(\mu_{\Phi(\xi)|x_i}\boldsymbol{\omega})} \right)^2 / \, \mathbb{V}^*(y_i|x_i)$$

involving the first derivative of the response function with respect to the linear predictor and $\mathbb{V}^*(y_i|x_i)$, the approximated observed variance from (5.9).

5.1.3 SIMEX

The SIMEX for Poisson regression can be straightforwardly adopted from the binary case in Section 4.1.3. Again, the error effect is studied on the linear

predictor $\widehat{f^*}(\xi_k) := \Phi(\xi_k)\widehat{\boldsymbol{\omega}}$ at points of interest ξ_k, instead of investigating its effect directly on $\widehat{f}(\xi_k) := \exp(\Phi(\xi_k)\widehat{\boldsymbol{\omega}})$. This modification guarantees the final SIMEX prediction function $\widehat{f}_{SIMEX}(\xi_k) = G\left(\widehat{f^*}_{SIMEX}(\xi_k)\right)$ to be non-negative as postulated for Poisson regression. The recipe for SIMEX is unchanged from the binary case, but now the mean function estimate is given by $\widehat{f}_{SIMEX}(\xi_k) = \exp\left(\widehat{f^*}_{SIMEX}(\xi_k)\right)$. Here, $\widehat{f^*}(\xi_k)_{SIMEX}$ denotes the SIMEX corrected linear predictor.

In the following simulation study, a quadratic extrapolation, based on the naive analysis (c=0) and the mean estimates over $B = 50$ repetitions for multiples $c \in \{0, 0.5, 1, 1.5, 2\}$ of the original measurement error variance, is used to attain the SIMEX estimates.

5.2 Simulation study

Basis function calibration, expanded basis function calibration and SIMEX are compared in a simulation study. Firstly, the various data scenarios considered in the simulation study are described. Then, some general settings are recalled. Finally, the results of the simulation study are presented and discussed.

5.2.1 The data

For each data scenario 200 data sets are simulated.
There are always two replicates ($m_i = 2$) available containing classical additive measurement error with $\mu_\delta = 0$. Thus, each surrogate observation $x_i, i = 1, \ldots, N$ represents the average over these two replicates. The measurement error variance can then be estimated from these replicates and so the money for the fortune teller can be saved.
In seven of the eight scenarios the $\xi_i, i = 1, \ldots, N$ are generated as independent normal random variables with mean μ_ξ and variance σ_ξ^2. Case 7

adopts a standardized $\chi^2(4)$ distribution for the generation of the ξ_i's. Each data set contains usually $N = 500$ samples, with exception of case 6, where $N = 1000$ samples are available. The level of measurement error variance is different for the data scenarios. As a consequence of having two replicates, the measurement error variance of the surrogates $x_i = \frac{x_{i1}+x_{i2}}{2}$ is only half the error variance that is stated below in the respective cases.
For the purpose of mean squared error calculations, the predictions of the correction methods were computed on a grid of 101 points in the interval $[a, b]$. The mean squared error was computed over this grid, which is expected to cover most of the distribution for ξ.
The series of simulation includes eight data cases, where the Poisson responses are generated from the underlying mean function $f(\xi) = \exp(-m(\xi))$ with functional argument $m(\xi)$. No under-/overdispersion is specified here. The general setup of this simulation study is almost identical to the one in the binary case:

Case 1: A quadratic function of the covariate with $m(\xi) = -0.2 + 0.25\xi + 0.1\xi^2$, with $N = 500$, $a = -2.0$, $b = 2.0$, $\sigma_\delta^2 = 0.8^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5$

Case 2: An oscillating function of the covariate with

$$m(\xi) = \begin{cases} 0.9 + 1.8\sin(5\xi)/(5\xi) & \xi \neq 0 \\ 2.7 & \xi = 0 \end{cases},$$

$N = 500$, $a = -2.0$, $b = 2.0$, $\sigma_\delta^2 = 0.2^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5$

Case 3: Same as Case 2 except $\sigma_\delta^2 = 0.5^2$

Case 4: Same as Case 2 except $\sigma_\delta^2 = 0.8^2$

Case 5: Another oscillating function of the covariate with

$$m(\xi) = \begin{cases} 1 + 2\frac{\sqrt{(0.25\xi+0.5)(1-(0.25\xi+0.5))}\sin(2\pi(1+2^{(9-4j)/5}))}{\frac{\xi}{4}+0.5+2^{(9-4j)/5}} & -2 \leq \xi \leq 2 \\ 1 & \text{otherwise} \end{cases}$$

for $j = 3$, with $N = 500$, $a = -2$, $b = 2$, $\sigma_\delta^2 = 0.5^2$, $\mu_\xi = 0$ and $\sigma_\xi^2 = 1.5^2$

Case 6: Same as case 4 except N=1000

The violation of the assumption of ξ being normally distributed is studied in

Case 7: The same as case 3 above except that ξ is a standardized $\chi^2(4)$ random variable. The MSE will be evaluated on $[-1.25, 2.00]$.

A plateau function is difficult to model with the RVM methods using RBF kernels or the MCMC approach using 2nd order truncated power series. This model misspecification is investigated in

Case 8: The same as case 3 above except that

$$\begin{aligned} m(\xi) &= 3 + 0.5(-2 + H(100\xi) + H(100(\xi - 0.5))), \\ &\text{where } H(\xi) = (1 + \exp(-\xi))^{-1}. \end{aligned}$$

Figure 5.1 and 5.2 display example data sets for each scenario as well as the true mean function. Only a single measurement is displayed here, although there are two replicate measurements available and usually the average is taken as surrogate to perform the analysis. Compared to the binary data examples in the respective Section 4.2.1, the Poisson responses rather allow for inferring the underlying mean function 'by eye' - at least when the measurement error is small.

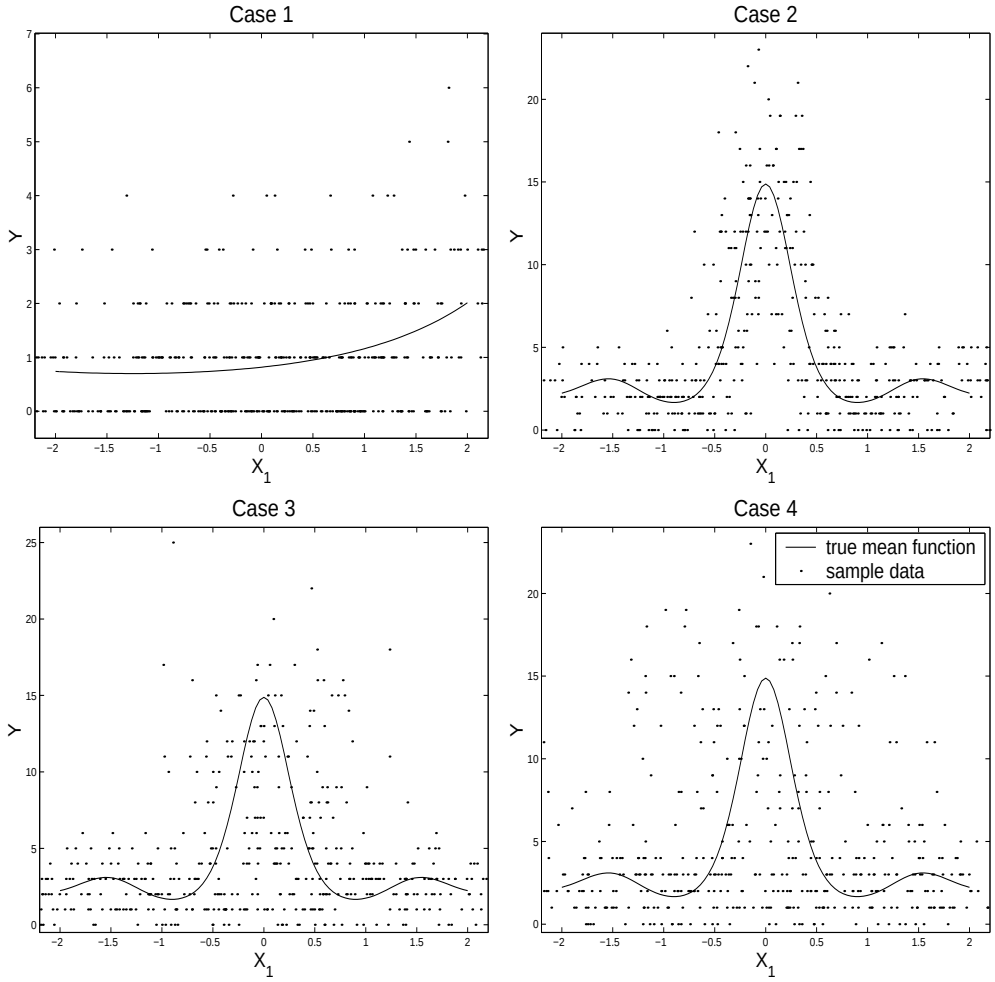

Figure 5.1: Example data sets for cases 1-4 and the respective true underlying mean function. The response is plotted versus one error-prone measurement, i.e. no averaging over the two available replicates is done here. Only the range $[a, b]$ on which the methods will be evaluated is shown.

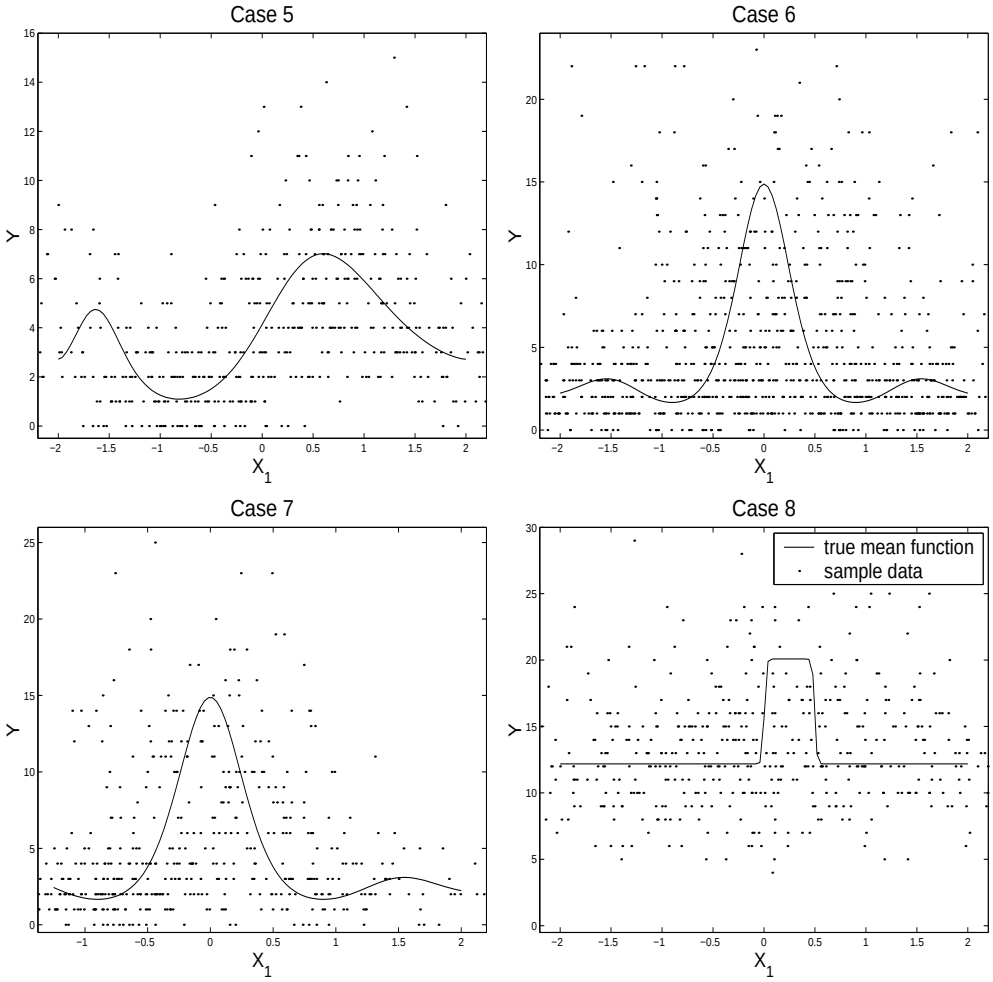

Figure 5.2: Example data sets for cases 5-8 and the respective true underlying mean function. The response is plotted versus one error-prone measurement, i.e. no averaging over the two available replicates is done here. Only the range $[a, b]$ on which the methods will be evaluated is shown.

5.2.2 Specification details of the methods

Basis calibration, expanded basis function calibration and SIMEX are implemented as described above with an arsenal of 100 radial basis functions centered on knots located at the quantiles of the observed data and one intercept. For SIMEX, the knots are again located at the quantiles of the artificially generated observations in each simulation step.
The kernel parameter η of the radial basis functions is selected from a list of admissible values like in the Gaussian implementation (cf. Section 3.2.2). Basis calibration and expanded basis function calibration copy the optimal kernel parameter selected by the naive approach.
All methods use the analytic updating scheme of the hyperparameter vector $\boldsymbol{\alpha}$ as described for the non-Gaussian case of the RVM in Section 2.1.2 and in Section 4.1.2 for the modified version in the expanded basis function calibration approach.

For an compact overview of the methods, the interested reader is referred to Table 4.1 of the previous chapter, which contains an overview of the compared methods in some essential respects.

5.2.3 The results

The quality of the correction methods is again investigated by mean squared error and pointwise bias.

MSE:

The mean squared error is computed over a grid of 101 equidistant values in the given interval $[a, b]$ (specific values see in Section 5.2.1 above)

$$\text{MSE} = \frac{1}{101} \sum_{k=1}^{101} \Big(f(\xi_k) - \hat{f}(\xi_k) \Big)^2 ,$$

where $f(\xi_k) = \mathbb{E}(y_k|\xi_k)$ is the true mean function and $\hat{f}(\xi_k)$ an estimate.
Table 3.2 presents summary results for the MSE from the 200 simulations for each data scenario. The smallest average MSE value among the naive analysis and the implemented correction methods in each scenario is in boldface. Like in the simulation study investigating the binary models (cf. Section 4.2.3) the correction quality of the presented methods is heterogeneous, though all are decidedly superior to the naive analysis (except for cases 1 and 8). Unlike the binary case, where higher measurement error variance suggests $\mathbf{RVM_{SIMEX}}$ correction, the results from cases 2-4 give no unambiguous recommendation when to use which method.
Case 5 is an oscillating function with two different frequencies and is generally hard to fit with the RVM models (see also simulation study in Krause & Tutz (2003)). However, even here the corrective power of the presented methods becomes obvious.
As can be seen from cases 4 and 6, increasing the sample size affects the correction methods only marginally in the Poisson regression compared to the pronounced effects in binary regression (cf. MSE results for binary regression in Table 4.2). That may be due to the fact that Poisson responses contain much more information than binary responses and there is no improvement in using 500 more (error-prone) observations in detecting the underlying complex mean function. This hypothesis is supported by the results of the $\mathbf{RVM_{naive}}$ for cases 4 and 6, which indicate no further improvement when using $N = 1000$ observations and consequently there appears to be no huge potential in making 500 additional observations available for the correction methods. Comparing the observed effects in binary and Poisson regression, when increasing the sample size, there seems to be the following relationship: the correction methods benefit from a larger sample size only if the naive analysis benefits, too. This is the case for binary regression, but not for Poisson regression.
All methods, most notably, including the structural ones, $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{EBC}}$, show distinctive correction power in case 7, although ξ is generated here as a standardized $\chi^2(4)$ random variable. Since $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{EBC}}$ are based on the additional assumption of ξ being normally dis-

tributed, these methods were expected to fail in that case.
When the functional form of the true mean function is not presentable as a sum of weighted basis functions as in case 8, then error correction seems to have little impact on the analysis.

Mean squared error Mean (SE) / Median (all $\times 10^2$)				
Method	Case 1	Case 2	Case 3	Case 4
RVM	1.10 (.05) / .92	9.56 (.35) / 8.73	8.91 (.39) / 7.91	9.81 (.40) / 8.45
$\mathbf{RVM_{naive}}$	**1.52** (.09) / 1.06	40.21 (1.23) / 37.20	345.22 (4.32) / 344.26	663.59 (4.89) / 655.66
$\mathbf{RVM_{BC}}$	2.16 (.14) / 1.68	26.31 (.92) / 23.97	194.09 (3.61) / 190.41	**385.96** (5.69) / 372.34
$\mathbf{RVM_{EBC}}$	1.62 (.09) / 1.28	**20.85** (.87) / 19.30	204.28 (4.61) / 203.70	394.77 (6.02) / 374.47
$\mathbf{RVM_{SIMEX}}$	2.36 (.16) / 1.59	25.11 (1.12) / 21.86	**164.62** (4.58) / 159.79	413.08 (7.80) / 386.69

Method	Case 5	Case 6	Case 7	Case 8
RVM	14.39 (.41) / 14.30	4.38 (.17) / 4.32	10.27 (.37) / 8.98	143.65 (2.65) / 134.60
$\mathbf{RVM_{naive}}$	56.70 (1.00) / 55.52	665.12 (3.93) / 665.12	417.85 (5.70) / 409.21	362.04 (2.74) / 364.62
$\mathbf{RVM_{BC}}$	44.68 (.99) / 42.97	**366.73** (3.97) / 358.14	206.94 (7.80) / 191.14	343.16 (2.86) / 342.42
$\mathbf{RVM_{EBC}}$	43.31 (.94) / 41.85	379.92 (4.64) / 365.55	194.97 (6.69) / 185.59	345.05 (2.87) / 346.87
$\mathbf{RVM_{SIMEX}}$	**42.50** (1.29) / 38.53	400.95 (6.13) / 384.33	**166.76** (8.28) / 138.91	**328.57** (2.95) / 327.54

Table 5.1: The mean squared error results for the simulation. In each column, the smallest mean value among the naive analysis and the implemented correction methods is in boldface.

Pointwise bias:

The pointwise bias of the methods under investigation can be seen from the visualization of the mean predictions for $\mathbb{E}(Y|\xi_k)$ over the 200 simulations in Figure 5.3 (for cases 1-4) and Figure 5.4 (for cases 5-8).
The cases 2-4 display impressively the growing difference between naive estimation and correction methods, when the error variance increases. While in case 1 all correction approaches seem to be more or less indistinguishable, the strength of $\mathbf{RVM_{BC}}$ is revealed for higher measurement error variance.
Case 5 can be fitted comparatively well by the naive method, while all correction methods slightly underestimate the true curve for $\xi_k \leq 0$, and overestimate it for $\xi_k > 0$.
Increasing the sample size does not seem to help to improve the corrective power, as already indicated by the MSE results (cf. Table 5.1).
Even under wrongly assuming that ξ is normally distributed, the structural methods $\mathbf{RVM_{BC}}$ and $\mathbf{RVM_{EBC}}$ maintain their prominence in case 7.
The plateau function in case 8 is badly fit by all approaches and only marginal effects of error correction are observable.

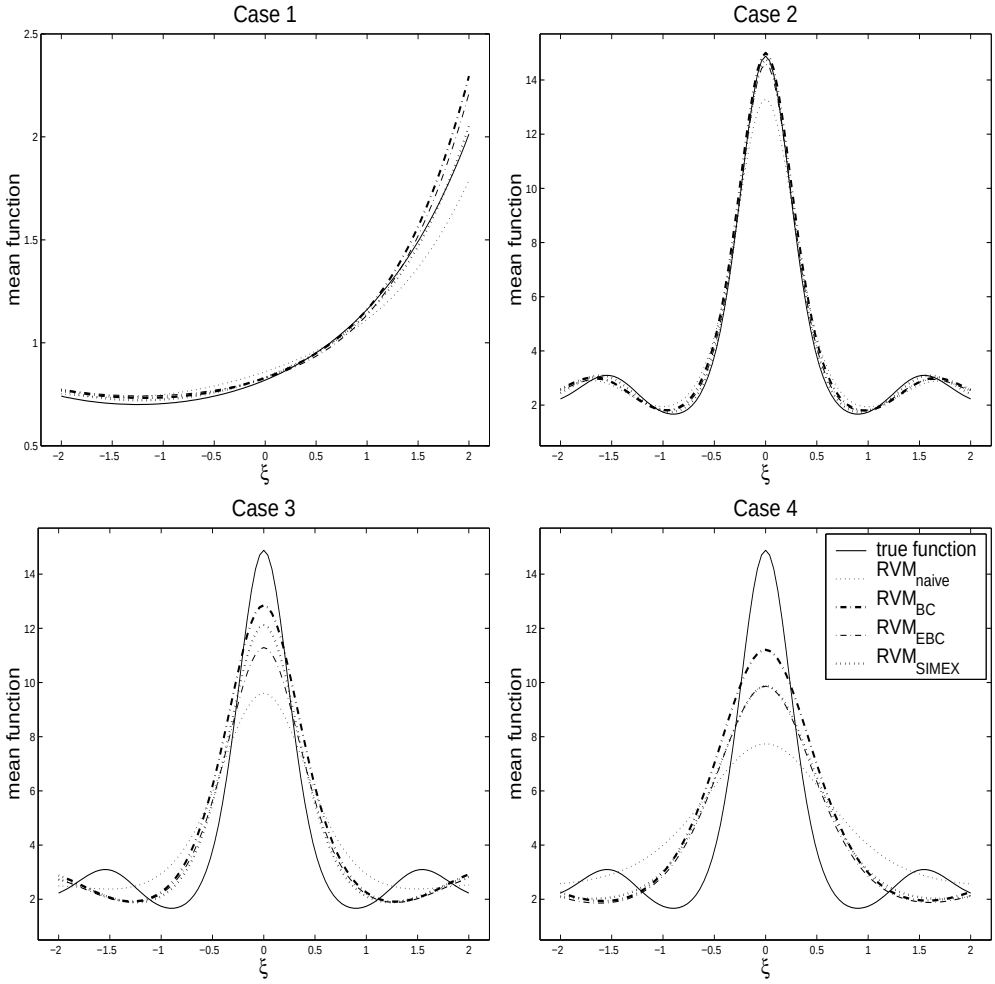

Figure 5.3: The mean functions over 200 simulations for cases 1-4. Data has been generated from the underlying mean function $\mathbb{E}(Y|\xi) = \exp(m(\xi))$. *Case 1 reflects a weak quadratic relationship between* ξ *and* $m(\xi)$*, cases 2-4 employ an oscillating* $m(\xi)$ *yielding a true mean function ranging from* 1.67 *to* 14.88*. These cases exclusively differ in the amount of measurement error, which is* $\sigma_\delta^2 = 0.2^2, 0.5^2, 0.8^2$*, respectively. Here, the* $\mathbf{RVM_{BC}}$ *shows to be somewhat superior for higher measurement error variance. However, there lies a clear improvement in all correction strategies. The RVM without measurement error is left out here for the sake of visibility.*

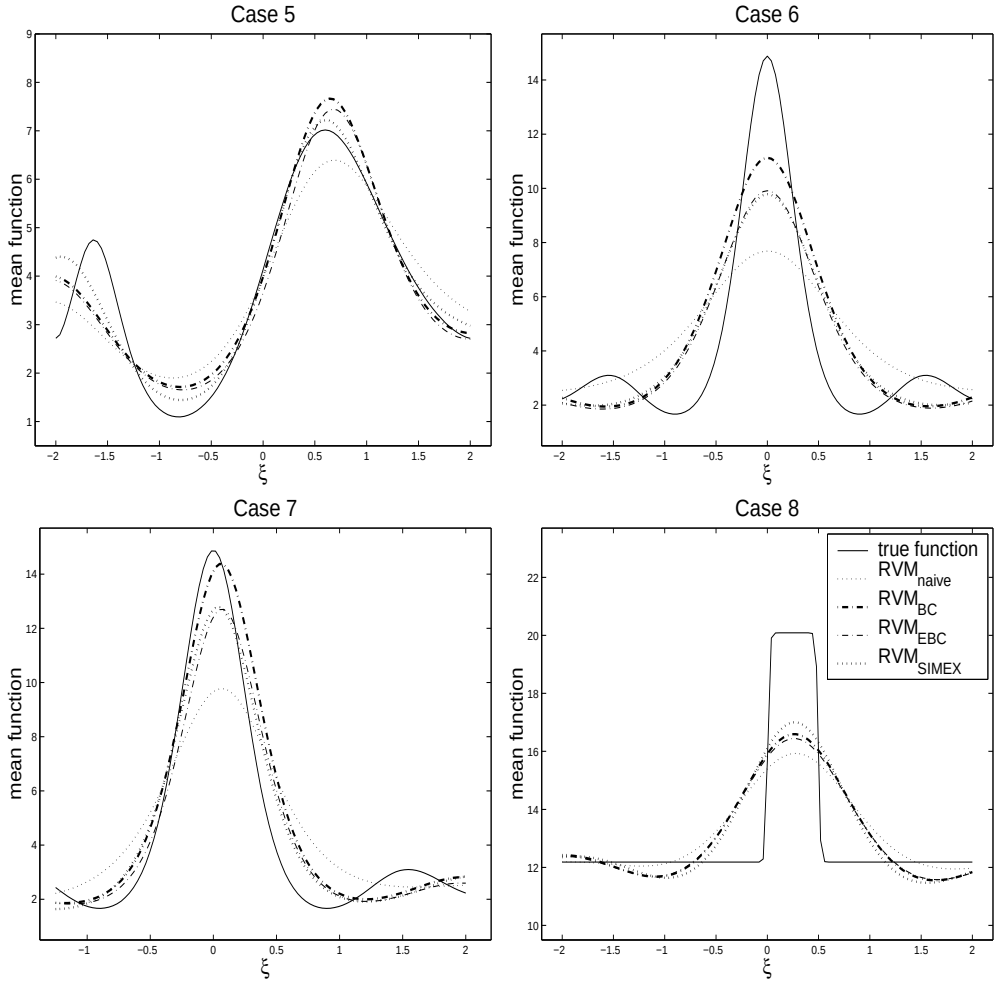

Figure 5.4: The mean functions over 200 simulations for cases 5-8. Data has been generated from the underlying mean function $\mathbb{E}(Y|\xi) = \exp(m(\xi))$. *Case 5 comprises an oscillating* $m(\xi)$ *with locally different frequencies. Case 6 is identical to case 4 but now supplying more sample data for the analysis. Case 7 employs non-normally distributed* ξ_i*'s and case 8 represents a mean function adopting a plateau function which is difficult to fit with RBF kernels. With exception of case 8, a clear improvement of the prediction is indicated for every correction method.* $\mathbf{RVM_{SIMEX}}$ *and* $\mathbf{RVM_{BC}}$ *seem to be slightly preferable in these cases.*

Chapter 6

Concluding discussion

There is a strong interest in generating knowledge on the basis of empirical observations. The basic requirements are a set of hypotheses, empirical data and, for instance, a statistical regression model for inference. Problems occur, when the hypotheses include complex covariates that can not be operationalized completely or can not be measured correctly. Applying the chosen statistical model without accounting for the intrinsic measurement error leads to deceptive conclusions about the true relation between covariates and response.

The cure lies in methods that account for covariate measurement error and this work develops such strategies for flexible regression using the relevance vector machine (RVM) in Gaussian, binary and Poisson regression. The RVM specifies a parametric, yet flexible, model and thus belongs to a subclass of nonparametric methods.

Only correction for covariate measurement error is investigated throughout the present work, since the impact of mismeasured responses on the analysis is usually far less pronounced. Emphasis is on the development of structural methods that use distributional information about the not directly observable covariate ξ. These include, on the one hand, the Markov Chain Monte Carlo (MCMC) sampling techniques and on the other hand so-called calibration methods. The latter ones are based on the equivalence between the

posterior mode estimator, adopted by the RVM, and Fisher scoring in a penalized likelihood setting for parameter estimation. Finally, the SIMulation EXtrapolation (SIMEX) approach presented here, is an adaption of the nonparametric SIMEX developed in Carroll et al. (1999) for the RVM.

The next paragraph summarizes the main points made in this thesis. Afterwards, a number of important, yet unsolved, problems and promising aspects of further research are presented.

Summary

In chapter 3, the correction for covariate measurement error in flexible models for Gaussian responses is discussed. This work develops two new correction methods for the RVM termed 'basis function calibration' and 'structural quasi likelihood'. These are particularly attractive, since they do not rely on computer intensive simulation like SIMEX or costly MCMC sampling like the state-of-the-art Bayesian P-splines for measurement error problems by Berry et al. (2002). Furthermore, basis function calibration can also be used, at least approximatively, for the non-Gaussian response cases.
The discussed methods are compared in a simulation study, where the method of Berry et al. (2002) is taken as a reference. All correction methods show a strong improvement compared to the naive analysis ignoring measurement error. The simulation study presents evidence that basis function calibration and structural quasi likelihood are the most powerful tools here to combat the adverse effects of measurement error.

Chapter 4 presents the correction of flexible models for binary responses. This work develops an expanded version of basis function calibration, termed 'expanded basis function calibration' and an MCMC version of the RVM using Bayesian measurement error correction.
The results of the attached simulation study are somewhat more heterogenous than in the Gaussian case. Judging by the pointwise bias, all correction methods appear to improve the naive results in the investigated data sce-

narios. Considering the MSE criterium, only the ad-hoc development that combines the idea of Bayesian 'data augmentation' and calibration shows a consistently strong improvement in all data cases. Little if any gain is indicated for the other calibration methods. However, the MCMC version of the RVM and the SIMEX adoption for the RVM do particularly well when the measurement error is moderate and the sample size is high. All correction methods clearly improve the analysis when the sample size is increased. Particularly in flexible binary regression, the sample size seems to be a strong determinant of successful error correction. Against that background, it would be interesting to re-run a part of the simulation study with increased sample size.
A part of the correction methods developed here are also applied in re-analyzing data from the MONICA study (MONItoring of trends and determinants in CArdiovascular disease). Here, the nutritional variables, animal and plant protein intake, are both suspected to influence mortality, however, their observations contain substantial measurement error. The results of the applied correction methods deviate clearly from the naive analysis and thus underline the impact of taking the error into account.
This work also contributes to the highly relevant problem of error correction in flexible regression for binary longitudinal data. This requires an extremely complex model, which accounts for person specific effects, autocorrelated covariate and response observations and, of course, for covariate measurement error. The parameter inference proceeds here via a subtle MCMC sampling algorithm.

Chapter 5 shows, how basis function calibration, expanded basis function calibration and SIMEX can be generalized to the case of Poisson responses. All correction methods show here a convincing performance on the simulated toy data.

Future work

This paragraph specifies some challenging, but probably rewarding prospects of future work based on the current achievements.

Correct estimators of standard errors for the presented (non-MCMC) correction methods are desirable. They must take into account the inherent measurement error in the covariate. Further research for the SIMEX method could be based on the theoretical developments for SIMEX using regression splines by Carroll et al. (1999). For the basis function calibration, the results for standard regression calibration (cf. Carroll et al. (1995), Subsection 3.12.2) and structural regression splines (cf. Carroll et al. (1999)) may be of interest. For structural quasi likelihood and expanded basis function calibration the asymptotic standard errors can be derived from the so-called sandwich formula (cf. Carroll & Stefanski (1990)), but must account for the penalization in the RVM setup.

It has been shown that the method for estimating the posterior mode of the mean model parameters in the RVM is exactly equivalent to Fisher scoring in a penalized likelihood setting, where the latter represents a more frequentistic view on this parameter estimation.
The powerful calibration methods, developed in this work, only intervene into this part of the optimization scheme. However, this scheme of optimization under penalization is also applied in a range of very promising alternative approaches to flexible regression. These include mixed model smoothing (cf. Wand (2003)) and structured additive regression (cf. Kneib & Fahrmeir (2005)), which are up to date applied to multinomial, ordinal or survival response data. Using radial basis functions instead truncated power series or B-spline basis functions for these approaches, makes the calibration methodology from the present work directly available for these attractive and fast emerging methods. Otherwise, the calibration of regression splines has already been presented by Carroll et al. (1999) and the calibration of B-spline basis functions as commonly used in the P-splines approach seems to be only

a stone's throw away.

The calibration methods developed in this thesis are readily applicable to the Berkson type covariate measurement error. Setting up a simulation study considering this error type or a mixture of classical and Berkson error will give further insight into the correction properties of these methods.

A special focus, when designing a new simulation study, should lie on the investigation of the 'byproduct approach', which combines data augmentation and calibration for binary regression. In contrast to the other methods, this is not firmly rooted in theory, but rather relies on several ad-hoc approximations. Nevertheless, it works surprisingly well in the simulation study presented here.

The core idea of all calibration methods is to find an (approximative) representation of the observed moments in terms of the true parameters. A completely different view of error correction is to approximate the ideal moments required for parameter inference. This strategy, described by Schennach (2004), can be easily extended to suit models like the RVM. Therefore, further investigation is attractive and may be fruitful.

Now, turning to the MCMC methods, a correction method for Gaussian longitudinal data can be directly attained from the binary longitudinal case in chapter 4. Solely, time prevented a realization during this work.

Also, for flexible Poisson regression under covariate measurement error an obvious enrichment lies in accounting for longitudinal data. This is an extremely important case for practical applications e.g. from the area of epidemiology. There, the dependent variable is often an aggregated number of events, e.g. death or disease, in a population. In a first step, this could be realized by using the Binomial approximation to the Poisson distribution and adopting the MCMC methodology developed for binary longitudinal data in chapter 4.

Finally, a lot of data sets could be re-analyzed, then appropriately accounting for covariate measurement error, if the developed MCMC approach could be generalized to more than one error-prone covariate. This seems to be manageable for the cross-sectional data case, however, much more difficult when considering longitudinal data.

Carl Friedrich Gauss's least squares method has most prominently made its way into present-day statistics. However, as demonstrated by the MCMC sampling strategies, there are also other approaches for making predictions. This thesis has worked in both fields and it seems probable that many more questions of future research can be answered when considering both areas together.

Bibliography

Albert, J. & Chib, S. (1993). Bayesian analysis of binary and polytomous response data, *Journal of the American Statistical Association* **88**(422): 669–679.

Andrieu, C., de Freitas, N., Doucet, A. & Jordan, M. I. (2003). An introduction to mcmc for machine learning, *Machine Learning* **50**(1-2): 5–43.

Augustin, T. (2002). Survival analysis under measurement error. Postdoctoral thesis at the University of Munich.

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**(457): 160–169.

Besag, J. (1986). On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society* **48**(B): 259–302.

Carroll, R. J. & Gallo, P. P. (1984). Comparisons between maximum likelihood and method of moments in a linear errors-in-variables regression model, *in* T. J. Santner & A. C. Tamhane (eds), *Design of Experiments: Ranking and Selection*, Marcel Dekker.

Carroll, R. J., Küchenhoff, H., Lombard, F. & Stefanski, L. A. (1996). Asymptotics for the simex estimator in structural measurement error models, *Journal of the American Statistical Association* **91**(433): 242–250.

Carroll, R. J., Maca, J. D. & Ruppert, D. (1999). Nonparametric regression in the presence of measurement error, *Biometrika* **86**: 541–554.

Carroll, R. J. & Ruppert, D. (1988). *Transformation and Weighting in Regression*, New York: Chapman & Hall.

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, London: Chapman & Hall/CRC.

Carroll, R. J. & Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors, *Journal of the American Statistical Association* **85**: 652–663.

Carroll, R., Midthune, D., Freedman, L. & Kipnis, V. (2006). Seemingly unrelated measurement error models, with application to nutritional epidemiology. To appear in *Biometrics*.

Carroll, R., Ruppert, D., Crainiceanu, C., Tosteson, T. & Karagas, M. (2004). Nonlinear and nonparametric regression and instrumental variables, *Journal of the American Statistical Association* **99**: 736–750.

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler, *The American Statistician* **46**(3): 167–174.

Chakraborty, S., Gosh, M. & Mallick, B. (2005). Bayesian non linear regression for large p small n problems, *Under revision for Journal of the American Statistical Association* .

Chib, S. (1995). Marginal likelihood from the gibbs output, *Journal of the American Statistical Association* **90**(432): 1313–1321.

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output, *Journal of the American Statistical Association* **96**(453): 270–281.

Chib, S. & Jeliazkov, I. (2006). Inference in semiparametric dynamic models for binary longitudinal data. To appear in *Journal of the American Statistical Association*.

Cook, J. & Stefanski, L. A. (1994). Simulation-extrapolation estimation for parametric measurement error models, *Journal of the American Statistical Association* **89**: 1314–1328.

Davidian, M. & Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density, *Biometrika* **80**: 475–488.

de Boor, C. (1978). *A practical guide to splines*, Springer.

Dellaportas, P. & Roberts, G. O. (2003). An introduction to mcmc, *in* J. Møller (ed.), *Spatial Statistics and Computational Methods, Lecture Notes in Statistics*, Springer, pp. 1–41.

Denison, D., Holmes, C., Mallick, B. & Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*, Wiley.

Döring, A. & Kußmaul, B. (1997). Ernährungsdeterminanten des herzinfarktrisikos, *Technical Report GSF-Fe-7629*, GSF - National research center for environment and health, Neuherberg.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing using b-splines and penalized likelihood (with comments and rejoinder), *Statistical Science* **11**(2): 89–121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: A bayesian perspective, *Statistica Sinica* **14**: 731–761.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Modells*, second edn, Springer.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348–1360.

Fan, J. & Truong, Y. K. (1993). Nonparametric regression with errors in variables, *Annals of Statistics* **21**: 1900–1925.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, New York.

Ganguli, B., Staudenmayer, J. & Wand, M. P. (2005). Additive models with predictors subject to measurement error, *Australian and New Zealand Journal of Statistics* **47**: 193–202.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**: 398–409.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721–741.

Gleser, L. J. (1990). Improvement of the naive estimation in nonlinear errors-in-variables regression, *in* P. J. Brown & W. A. Fuller (eds), *Statistical Analysis of Measurement Error Models and Application*, number 112, Contemporary Mathematics, pp. 99–114.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, MIT Press, Cambridge MA.

Gössl, C. & Küchenhoff, H. (2001). Bayesian analysis of logistic regression with an unknown change point and covariate measurement error, *Statistics in Medicine* **20**: 3109–3121.

Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrica* **82**: 711–732.

Green, P. J. (2001). A primer on mcmc, *in* O. E. Barndorff-Nielsen, D. R. Cox & C. Klüppelberg (eds), *Monographs on Statistics and Applied Probability 87: Complex Stochastic Systems*, Chapman and Hall, pp. 1–62.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**: 383–385.

Hastings, W. (1970). Monte carlo sampling methods using markov chains, and their applications, **57**: 97–109.

Heid, I. M., Gerken, M., Wellmann, J., Küchenhoff, H., Kreienbrock, L. & Wichmann, H. E. (2002). On the potential of measurement error to induce differential bias between groups: an example from radon epidemiology, *Stat Med* **21**: 3261–3278.

Holmes, C. C. & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Analysis* **1**(1): 145–168.

Horn, R. A. & Johnson, C. R. (1985). *Matrix Analysis*, Cambridge University Press.

Jackman, S. (2000). Estimation and inference via bayesian simulation: an introduction to markov chain monte carlo, *American Journal of Political Science* **44**(2): 269–398.

Kass, R. E., Carlin, B. P., Gelman, A. & Neal, R. M. (1998). Markov chain monte carlo in practice: A roundtable discussion, *The American Statistician* **52**: 93–100.

Kass, R. E. & Rafferty, A. E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**: 773–795.

Keil, U. (2000). Monica: Abschluss der welt-herz-studie. kommentar: Bedeutung der ökologischen studie monica, *Deutsche Medizinische Wochenschrift* **125**: A8–A10.

Kneib, T. & Fahrmeir, L. (2005). Structured additive regression for categorical space-time data: A mixed model approach. To appear in *Biometrics*.

Krause, R. & Tutz, G. (2003). *Additive Modeling with Penalized Regression Splines and Genetic Algorithms*, Institut für Statistik. Discussion Paper 312.

Küchenhoff, H. (1995). Schätzmethoden in mehrphasigen regressionsmodellen. Post-doctoral thesis at the University of Munich.

Küchenhoff, H. & Carroll, R. J. (1997). Segmented regression with errors in predictors: semi-parametric and parametric methods, *Statistics in Medicine* **16**: 169–188.

Küchenhoff, H., Mwalili, S. & Lesaffre, E. (2005). A general method for dealing with misclassification in regression: The misclassification simex. To appear in *Biometrics*.

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

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines, *Journal of the Royal Statistical Society, Series B* **61**: 381–400.

Lindley, D. V. (1957). A statistical paradox, *Biometrika* **44**: 187–192.

MacKay, D. J. C. (1994). Models of Neural Networks III, *in* E. Domany, J. L. van Hemmen & K. Schulten (eds), *Bayesian methods for backpropagation networks*, Springer, chapter 6, pp. 211–254.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*, first edn, Cambridge University Press.

McCullagh, P. (1983). Quasi-likelihood functions, *Annals of Statistics* **11**: 59–67.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. H. & Teller, E. (1953). Equation of state calculations for fast computing machines, *Journal of Chemical Physics* **21**(6): 1087–1092.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal, *Biometrika* **58**: 545–554.

Ranyimbo, A. O. & Held, L. (2006). Estimation of the false negative fraction of a diagnostic kit through bayesian regression model averaging, *Statistics in Medicine* **25**: 653–667.

Richardson, S. (1996). Measurement error, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall.

Richardson, S. & Gilks, W. R. (1993a). A bayesian approach to measurement error problems in epidemiology using conditional independence models, *American Journal of Epidemiology* **138**: 430–442.

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

Robert, C. P. (1995). Simulation of truncated normal variables, *Statistics and Computing* **5**: 121–125.

Roeder, K. & Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals, *Journal of the American Statistical Association* **92**: 894–902.

Rummel, D. (2004). Improving the relevance vector machine under covariate measurement error, *in* A. Biggeri, E. Dreassi, C. Lagazio & M. Marchi (eds), *Proceedings of the 19th International Workshop on Statistical Modelling*, Firenze University Press, pp. 254–258.

Rummel, D. (2005). The relevance vector machine under covariate measurement error, *in* C. Weihs & W. Gaul (eds), *Classification - The Ubiquitous Challenge*, Springer, pp. 296–303.

Schafer, D. & Purdy, K. (1996). Likelihood analysis for errors-in-variables regression with replicate measurements, *Biometrika* **83**: 813–824.

Schennach, S. (2004). Estimation of nonlinear models with measurement error, *Econometrica* **72**: 33–75.

Schneeweiß, H. (1990). *Ökonometrie*, Physica, Heidelberg.

Smith, M. & Kohn, R. (1996). Nonparametric regression using bayesian variable selection, *Journal of Econometrics* **75**(2): 317–343.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models, *Communications in Statistics, Part A - Theory and Methods* **18**: 4335–4358.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *The Annals of Statistics* **22**: 1701–1762.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**: 82–86.

Tipping, M. E. (2000). The Relevance Vector Machine, *in* S. A. Solla, T. K. Leen & K. R. Müller (eds), *Advances in Neural Information Processing Systems 12*, MIT Press, pp. 652–658.

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine, *Journal of Machine Learning Research* **1**: 211–244.

Tipping, M. E. & Faul, A. C. (2002). Analysis of Sparse Bayesian Learning, *in* T. G. Dietterich, S. Becker & Z. Ghahramani (eds), *Advances in Neural Information Processing Systems 14*, MIT Press, pp. 383–389.

Tipping, M. E. & Faul, A. C. (2003). Fast marginal Likelihood maximisation for sparse Bayesian Models, *in* C. M. Bishop & B. J. Frey (eds), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley.

Vidakovic, B. (2005). Hierarchical bayes and empirical bayes. Handout for the course BAYESIAN STATISTICS FOR ENGINEERS at Georgia Tech University.
*http://www2.isye.gatech.edu/~brani/isyebayes/bank/handout8.pdf

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression, *Journal of the Royal Statistical Society* **40**(B): 364–372.

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

Wand, M. P. (2003). Smoothing and mixed models, *Computational Statistics* **18**: 223–249.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss newton method, *Biometrika* **61**: 439–447.

Willett, W. (1998). *Nutritional Epidemiology*, Oxford University Press.

Wolf, R. (2004). *Vergleich von funktionalen und strukturellen Messfehlerverfahren*, Berlin: Logos Verlag. Doctoral thesis at the University of Munich.

Lebenslauf

David Rummel - geboren am 9.Juli 1977 in Lohr am Main

Schulausbildung	
1983-1987	Grundschule in Lohr-Sendelbach
1987-1996	Franz-Ludwig-von-Erthal-Gymnasium in Lohr am Main, mathematisch-naturwissenschaftlicher Zweig
Zivildienst	
1996-1997	in der Inneren Abteilung und der Endoskopie am Kreiskrankenhaus Lohr am Main
Studium	
1997-2003	Studium der Statistik an der LMU München mit den Anwendungsgebieten VWL und Soziologie und dem Fach der speziellen Ausrichtung Wissenschaftstheorie
09/2000 - 02/2001	Studium an der University of North London im Rahmen des ERASMUS-Programmes
Beruf	
seit 03/2000	Mitarbeit im Statistischen Beratungslabor (StaBLab) am Institut für Statistik der LMU München
seit 05/2003	wissenschaftlicher Mitarbeiter im Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen – Teilprojekt C2" und am Institut für Statistik der LMU München
seit 03/2004	freiberufliche statistische Beratung, schwerpunktmäßig im Bereich osteopathische Medizin

München, den 8.März 2006