Linear Regression Analysis under Sets of Conjugate Priors

Gero Walter
Department of Statistics,
Ludwig-Maximilians-University
Munich, Germany
gero.walter@campus.lmu.de

Thomas Augustin
Department of Statistics,
Ludwig-Maximilians-University
Munich, Germany
thomas@stat.uni-muenchen.de

Annette Peters
GSF – National Research Center
for Environment and Health,
Institute for Epidemiology
Neuherberg, Germany
peters@gsf.de

Abstract

Regression is *the* central concept in applied statistics for analyzing multivariate, heterogenous data: The influence of a group of variables on one other variable is quantified by the regression parameter β. In this paper, we extend standard Bayesian inference on β in linear regression models by considering imprecise conjugated priors. Inspired by a variation and an extension of a method for inference in i.i.d. exponential families presented at ISIPTA'05 by Quaeghebeur and de Cooman, we develop a general framework for handling linear regression models including analysis of variance models, and discuss obstacles in direct implementation of the method. Then properties of the interval-valued point estimates for a two-regressor model are derived and illustrated with simulated data. As a practical example we take a small data set from the AIRGENE study and consider the influence of age and body mass index on the concentration of an inflammation marker.

Keywords. AIRGENE study, analysis of variance, exponential family, (imprecise) conjugate priors, imprecise probability models, interval probability, prior-data conflict, regression, robust Bayesian inference

1 Introduction and Sketch of the Argument

From engineering science over econometrics to sociology, from psychology over biometrics to medicine, one of the omnipresent questions is how certain variables (called covariates/confounders, regressors, stimulus or independent variables, here denoted by x) influence a certain outcome (called response or dependent variable z). The answer is obtained from regression models, and so regression modelling is *the* central concept in applied statistics.

The most common and simple case, dating back already to Gauß, is linear regression (see Section 3.1 for more details on the model), where, possibly after some transformations, for every unit i, taken from a sample of size k, the dependent variable z_i is assumed to be of metric scale and to be linearly related to p independent variables $x_{i1}, x_{i2} \ldots, x_{ip}$:

$$z_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \,, \qquad (1)$$

where ε_i is a stochastic error term subsuming the residual variation beyond the linear relationship between the variables.

The so-called regression coefficients β_j, $j = 1 \ldots, p$ measure the extent to which the dependent variable is expected to change if the value of the j-th regressor is enlarged by one unit and all other regressors remain unchanged. Often $x_{i1} = 1$ for all i and then β_1 is called intercept, describing some baseline level. The special case where all regressors are categorical (and coded via (several) 0/1 variables) is known as ANOVA (analysis of variance). The coefficients β_j can, for example, be estimated by the classical least squares method, or, relying on the Bayesian paradigm, by assigning a prior on the β_j's and updating it in the light of the data. This update step is especially elegant and simple to perform in situations we will call LUCK-models (described in Section 2.1). In our situation there are several possibilities to construct such a LUCK-model. We will rely on that multivariate normal as the prior for β_j, $j = 1 \ldots, p$, which has become the standard for regression analysis (see, e.g., [13]).

Although imprecise probabilities and related concepts [21, 26, 18] have proven to be quite powerful and convincing in many areas of application, regression analysis has only lived in the shadows there:[1] population heterogeneity, i.e. individual variation related to different covariate values, has mainly been incorporated by means of classification models [30, 1]; generalized inference (in particular along the lines of Walley's generalized Bayes's rule [21]) and decision theory (see the

[1] Very rare exceptions can be found in the robust Bayesian context, including: [6, 12]

survey in [20]) have almost exclusively confined themselves to the case of homogenous populations (i.i.d. case) or two-sample models (like [22, Section 5] and [8]).

Sound regression models would make imprecise probabilities quite attractive for applied scientists. As a step towards this ambitious aim we show that the approach of Quaeghebeur and de Cooman [16], who developed a concept of imprecise conjugated priors that nicely generalizes the widely applied imprecise Dirichlet model (IDM) [22, 2, 3], can be extended in an appropriate way.[2] Indeed, at least two different ways are successful. We briefly comment on the first one, which directly adopts to regression analysis Quaeghebeur and de Cooman's [16] original way to proceed, and investigate in more detail the second one, which is in straight line with the standard model for regression analysis (cf., e.g., [13].) For that purpose we interpret [16]'s approach, beyond its direct application in their work, as a general method for powerfully introducing imprecision into a huge class of Bayesian models, which we will call, for sake of brevity, LUCK-models in this paper, and demonstrate that the standard model for Bayesian regression analysis indeed fits into this framework.

In more detail, the paper is organized as follows: In Section 2 we collect some basic ingredients from Bayesian analysis, identify the special case of Bayesian analysis (LUCK-models) that underlies our basic argument, and then turn to the method for introducing imprecision into conjugate priors [16]. Section 3.1 firstly recalls classical[3] Bayesian regression analysis and puts it into the framework of LUCK-models. After having utilized [16] as a powerful method to extend LUCK-models to imprecise probabilities, we arrive at a general framework for regression analysis under sets of conjugate priors. We then focus consideration on a special case with two regressors, where some complex constraints underlying the general situation can be made easily tractable. The results are illustrated in Section 4 by simulated data and in Section 5 by a small data set from the AIRGENE study [15]. We conclude with some remarks on modifications and extensions, including the possibility to incorporate modelling of prior-data conflict, which was explicitly named by Walley as one of the main arguments for imprecise probabilities [21, p. 6], but which cannot be captured by the original method along the lines of [16].

[2] A different approach to generalize regression analysis has been proposed quite recently [28, Chapter 13] in the framework of the symmetric theory based on logical probability ([28], see also [27]).

[3] We use the term 'classical' for all concepts relying on precise probabilities / linear previsions.

2 Bayes Inference and LUCK-models

2.1 Classical Bayesian Inference and LUCK-models

As a preparation, some basic notions from the *classical Bayesian approach* will be recalled: Central is the assumption that the knowledge on a (possibly multivariate) parameter ϑ can be perfectly expressed by a single precise probability distribution on ϑ. So, inference from a (possibly multidimensional) sample w, the distribution of which is described by a density or probability function $f(w \mid \vartheta)$ (called likelihood in this context), consists in updating the so-called *prior* $p(\vartheta)$ to the so-called *posterior* $p(\vartheta \mid w)$ via Bayes's rule

$$p(\vartheta \mid w) \propto f(w \mid \vartheta) \cdot p(\vartheta)\,. \tag{2}$$

For a Bayesian, the prior describes the knowledge before having seen the sample, and the posterior subsumes the complete knowledge on ϑ after having seen the sample, and therefore it underlies all inferences drawn from the data.

For the intended application presented later on, it is quite convenient to distinguish certain standard situations (called *models with 'Linearly Updated Conjugate prior Knowledge'* (LUCK) here) of Bayesian updating with classical probabilities, where prior and posterior fit nicely together in the sense that

i) they belong to the same class of parametric distributions, a case where they are called *conjugate*, and, in addition,
ii) the updating of one parameter ($y^{(0)}$ below) of the prior is linear.[4]

More precisely, we introduce the following definition:

Definition 1 *Consider classical Bayesian inference on a parameter ϑ based on a sample w as described in (2), and let the prior $p(\vartheta)$ be characterized by the (vectorial) parameter $\vartheta^{(0)}$. The pair $\big(p(\vartheta), p(\vartheta \mid w)\big)$ is said to constitute a* LUCK-model of size q in the natural parameter ψ with prior parameters $n^{(0)} \in \mathbb{R}^+$ and $y^{(0)}$ and sample statistic $\tau(w)$, *iff there exist $q \in \mathbb{N}$ as well as transformations of ϑ into ψ and $\mathbf{b}(\psi)$ and of $\vartheta^{(0)}$ into $n^{(0)}$ and $y^{(0)}$, such that $p(\vartheta)$ and $p(\vartheta \mid w)$ can be rewritten in the following way:*[5]

$$p(\vartheta) \propto \exp\big\{n^{(0)}\big[\langle \psi, y^{(0)}\rangle - \mathbf{b}(\psi)\big]\big\} \tag{3}$$

and

$$p(\vartheta \mid w) \propto \exp\big\{n^{(1)}\big[\langle \psi, y^{(1)}\rangle - \mathbf{b}(\psi)\big]\big\} \tag{4}$$

with

$$n^{(1)} = n^{(0)} + q \quad \textit{and} \quad y^{(1)} = \frac{n^{(0)} y^{(0)} + \tau(w)}{n^{(0)} + q}\,. \tag{5}$$

[4] The second parameter $n^{(0)}$ possesses a vivid interpretation as "prior strength", which will become clearer in Section 2.2.

[5] $\langle a, b\rangle$ denotes the scalar product of a and b.

2.2 Imprecise Priors for Inference in LUCK-models

Several powerful approaches have been proposed to overcome the "dogma of ideal precision" (Walley) underlying classical Bayesian inference (cf., in particular, [14, 7, 5, 16]; see also Section 6). We rely in the following on the work of Quaeghebeur and de Cooman [16], who consider, by utilizing a general result (see, e.g., [4, Proposition 5.4]), certain LUCK-models for Bayesian inference based on independently and identically distributed (i.i.d.) observations from *regular, linear canonical exponential families* [4, p. 202 and p. 272f]. The central idea of [16] is that the seemingly strange parameterization in terms of $y^{(0)}$ and $n^{(0)}$ in (3) and (4) is perfectly suitable to be generalized to credal sets of priors. The crucial point is that these parameters are updated *linearly*, thus allowing for an easily tractable imprecise calculus: When sets of priors are defined via sets of parameters, and these sets of parameters are defined by lower and upper bounds, the lower and upper bounds of the sets of posterior parameters can be obtained directly from (5). So, just as in the popular IDM, which is contained as the special case of a multinomial sampling model with conjugated Dirichlet priors, minimization and maximization problems on the set of posteriors can be reduced to minimization and maximization problems on the set of priors in the case when the parameter $y^{(1)}$ (or a linear function of it) is the quantity of interest.

It shall be noted explicitly that this line of argumentation simply uses the linearity of the updating in the parameters, not the concrete derivation of the conjugate prior. Consequently, Quaeghebeur and de Cooman's technique to construct *imprecise* conjugate priors can be applied to any LUCK-model.

In more detail, the following technique will be applied: Given the situation in Definition 1, let $y^{(0)}$ vary in some set $\mathcal{Y}^{(0)} \subset \mathcal{Y}$, where the parameter space $\mathcal{Y}$ is taken as the convex hull (without the boundary) of the range of $\tau(w_i)$, and take as the imprecise prior the credal set consisting of all convex mixtures of all $p(\vartheta)$ from (3) created by varying $y^{(0)}$ in $\mathcal{Y}^{(0)}$. After having evaluated the sample w, the posterior credal set arising from applying Bayes's rule element by element has to be determined. For its calculation it is sufficient to consider the extreme points, and so it is obtained as the set of all convex mixtures of posteriors $p(\vartheta \,|\, w)$ arising from (4) by varying $y^{(1)}$ in $\mathcal{Y}^{(1)}$, where

$$\mathcal{Y}^{(1)} = \left\{ \frac{n^{(0)} y^{(0)} + \tau(w)}{n^{(0)} + n} \;\middle|\; y^{(0)} \in \mathcal{Y}^{(0)} \right\} \subset \mathcal{Y} \,. \quad (6)$$

$\mathcal{Y}^{(1)}$ can actually be seen as a shifted and rescaled version of $\mathcal{Y}^{(0)}$:

$$\mathcal{Y}^{(1)} = \frac{n^{(0)}}{n^{(0)}+n} \cdot \mathcal{Y}^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{1}{n} \sum_{i=1}^{n} \tau(w_i) \,, \quad (7)$$

which immediately suggests a vivid interpretation of $n^{(0)}$ as "prior strength" or as "pseudocounts", as it plays the same role for the prior as n for the sample. So, $n^{(0)}$ can be interpreted as the size of an imaginary sample that corresponds to the trust on the prior information in the same way as the sample size of a real sample corresponds to the trust in conclusions based on that sample.

For posterior inference, lower and upper posterior expectations are derived as the infimum and the supremum over all classical expectations with $y^{(1)}$ varying in $\mathcal{Y}^{(1)}$. The resulting relations between $n^{(0)}$, prior and posterior bounds are in essence the same as in the IDM:[6] In particular, for $n^{(0)} = n$, the width of the posterior expectation interval is half the width of the prior interval.

The choice of $\mathcal{Y}^{(0)}$ should reflect the prior information on the parameters. When there is very little or no information at all available, $\mathcal{Y}^{(0)}$ should be chosen as large as possible, that is, as the set of all possible parameter values, $\mathcal{Y}^{(0)} = \mathcal{Y}$. However, in most cases this would lead to the posterior set $\mathcal{Y}^{(1)}$ being vacuous as well, whatever the number of observations used for updating; for any $\overline{y}_j^{(0)} = \infty$, it holds that $\overline{y}_j^{(1)} = \infty$ as well. To avoid this, $\mathcal{Y}^{(0)}$ must be bounded by (element-wise) finite lower and upper boundaries.[7] This need to bound $\mathcal{Y}^{(0)}$ is not perceived as a severe restriction in practical application; typically, as in our example in Section 5, the very rough magnitude of reasonable parameter values is known, and there exist some trivial bounds.

3 Towards Imprecise Normal Regression Models

3.1 The Linear Regression Model, and its Classical Bayesian Treatment

In handling the linear regression model it is helpful to arrange the components in column vectors, denoted without index, i.e., $z = (z_1, \ldots, z_k)^\mathsf{T}$, and to collect all regressors column by column in the so-called design matrix $\mathbf{X}$. Equation (1) then reads as

$$z = \mathbf{X}\beta + \varepsilon \,, \quad \mathbf{X} \in \mathbb{R}^{k \times p},\ \beta \in \mathbb{R}^p,\ z \in \mathbb{R}^k,\ \varepsilon \in \mathbb{R}^k \,;$$

ε is assumed to have expected value $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$, i.e. $\mathbb{V}(\varepsilon_i) = \sigma^2$, the variance of ε_i does

[6] $n^{(0)}$ corresponds to the parameter s in the IDM.

[7] For the IDM, this is not necessary, as the parameter space $\mathcal{Y}$ itself is already bounded, being the unit simplex.

not vary among the units (homoscedasticity) and all units are uncorrelated.

There are several methods to construct estimators $\hat{\beta}$ for the regression parameter β. With the least squares method, $\hat{\beta}$ is chosen to minimize the squared difference between the observed response values z and the values estimated by $\mathbf{X}\hat{\beta}$, yielding the well-known *least squares (LS) estimator*

$$\hat{\beta}_{LS} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} z\,. \tag{8}$$

Other estimation techniques additionally assume that the error term ε is normally distributed. Then, as the design matrix $\mathbf{X}$ is considered to be non-stochastic,[8] also z is normally distributed,

$$z \sim N_k(\mathbf{X}\beta, \sigma^2\mathbf{I})\,. \tag{9}$$

This point of view is very helpful for several types of generalizations.[9] Reinterpreting (9) as a likelihood on β and σ^2 and applying the maximum likelihood (ML) principle again leads to the estimator from (8). In the Bayesian context appropriate priors related to the parameters and the likelihood (9) have to be found. Several choices for the prior seem attractive,[10] even different LUCK-models can be produced: In the light of the intended generalization below one natural possibility would be to follow the path of Quaeghebeur and de Cooman closely, by constructing a conjugate prior along the general construction method for LUCK-models (see, e.g., [4, Proposition 5.4]), also mentioned at the beginning of Section 2.2. For the case of known (or in advance estimated) σ^2 considered throughout the paper, one obtains by this procedure

$$p(\beta) \propto \exp\left\{n^{(0)}\left[\langle \beta, y^{(0)}\rangle - \mathbf{b}(\beta)\right]\right\},$$

where $\mathbf{b}(\beta) = \frac{1}{2\sigma^2}\sum_{i=1}^{k}\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)^2$. This prior can be shown to be a normal distribution on β, with its parameters being some transformations of $n^{(0)}$ and $y^{(0)}$ depending on $\mathbf{X}$.[11] It was maybe this strange dependency of the prior on the covariates $\mathbf{X}$ that resulted in this prior rarely being used for estimating regression parameters in the Bayesian framework.

[8] If $\mathbf{X}$ is stochastic, then it is common, and legitimate, practice (cf., e.g., [9]) to perform the analysis conditional on $\mathbf{X}$, hence (9) is replaced by $z \mid \mathbf{X} \sim N_k(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

[9] It makes also clear how the heterogeneity in the data is modelled: Each response z_i is assumed to be normally distributed, but the corresponding mean value depends on the individual characteristics (regressors) $x_{i1}, \ldots, x_{ip}$ and the effect size (expressed by β).

[10] So-called 'objective Bayesian estimation' of β, using the 'non-informative' prior $p(\beta) \propto$ *const.* leads to the same results as LS and ML when the expected or the maximum value of the posterior is used as the estimate. Therefore, when the interval-valued estimations of β proposed in this work are compared with the LS estimate, they are implicity compared to the ML and the objective Bayesian estimates as well.

[11] See [25] for more details on this model.

Instead, the commonly used approach specifies

$$\beta \sim \mathrm{N}_p\left(\beta^{(0)}, \sigma^2\boldsymbol{\Sigma}^{(0)}\right) \tag{10}$$

as the conjugate prior.[12] Advocated, e.g., by [13], it has become the standard, why we call the model based on this prior *normal regression model* throughout the paper. Applying Bayes's rule (2) to (10) yields

$$\beta \,|\, z \sim \mathrm{N}_p\left(\beta^{(1)}, \sigma^2\boldsymbol{\Sigma}^{(1)}\right), \tag{11}$$

where the updated parameters $\beta^{(1)}$ and $\boldsymbol{\Sigma}^{(1)}$ are obtained as

$$\beta^{(1)} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda}^{(0)}\right)^{-1}\left(\mathbf{X}^\mathsf{T} z + \boldsymbol{\Lambda}^{(0)}\beta^{(0)}\right) \tag{12}$$

$$\boldsymbol{\Sigma}^{(1)} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda}^{(0)}\right)^{-1}, \tag{13}$$

$\boldsymbol{\Lambda}^{(0)} = {\boldsymbol{\Sigma}^{(0)}}^{-1}$ being the so-called precision matrix.[13]

3.2 The Normal Regression Model as a LUCK-model

Now the argument that the extension proposed in [16] is neither limited to the i.i.d. case of homogenous populations nor to the special construction of LUCK-models considered there becomes fruitful: The standard Bayesian treatment of regression models based on the prior (10) can be shown to fit into the framework of LUCK-models, a fact that luckily immediately enables an appropriate generalization to imprecise models.

Theorem 2 *Consider the normal regression model described by the prior $p(\beta)$ from (10) with prior parameters $\beta^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$, and the posterior $p(\beta \,|\, z)$ from (11) with (12) and (13).*

Fixing a value $n^{(0)}$, $\big(p(\beta), p(\beta \,|\, z)\big)$ constitutes a LUCK*-model of size* 1 *with prior parameters*

$$y^{(0)} = \frac{1}{n^{(0)}}\begin{pmatrix} \boldsymbol{\Lambda}^{(0)} \\ \boldsymbol{\Lambda}^{(0)}\beta^{(0)} \end{pmatrix} =: \begin{pmatrix} y_a^{(0)} \\ y_b^{(0)} \end{pmatrix} \tag{14}$$

and $n^{(0)}$ and sample statistic

$$\tau(z) = \tau(\mathbf{X}, z) = \begin{pmatrix} \mathbf{X}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T} z \end{pmatrix} =: \begin{pmatrix} \tau_a(\mathbf{X}, z) \\ \tau_b(\mathbf{X}, z) \end{pmatrix}. \tag{15}$$

[12] Throughout the paper we denote prior parameters by the superscript $^{(0)}$ and, if appropriate, corresponding parameters of the posterior by $^{(1)}$. Here, the mean vector $\beta^{(0)} \in \mathbb{R}^p$ and the (positive definite) covariance matrix $\boldsymbol{\Sigma}^{(0)} \in \mathbb{R}^{p\times p}$ are the prior parameters defining the concrete distribution on β.

[13] If, in addition, σ^2 is considered unknown, too, the commonly used prior distribution conjugate to the likelihood in (9) is the so-called normal-inverse gamma distribution (e.g., [13, §9.4]). Unfortunately, this model will turn out to be not generalizable in the same way as it is done here for the normal regression model (cf., [23, Appendix], and also briefly in [25]). A first way out would be to estimate σ^2 in advance, and then to apply the normal regression model with the estimated value of σ^2, a strategy that we followed in our examples in Sections 4 and 5.

Proof: The proof is given in [24].

Knowing now the form of $y^{(0)}$, we can finally start with the imprecise probability calculus: By varying $y^{(0)}$ from (14) in a set $\mathcal{Y}^{(0)} \subset \mathcal{Y}$, the set of priors is generated. Since $\mathcal{T}$, the range of the sample statistic, is the product of the set of positive semidefinite $(p \times p)$ matrices and arbitrary vectors of dimension p, $\mathcal{Y}$ is taken as the convex hull of $\mathcal{T}$ without the boundary, thus $y_a^{(0)}$ having to be a positive definite $(p \times p)$ matrix. On the one hand, $\mathcal{Y}^{(0)}$ is chosen in order to reflect prior knowledge on β; on the other hand, this set must, as mentioned above at the end of Section 2.2, be bounded in order to avoid the possibility of vacuous posterior inference. In the case of a multidimensional parameter space $\mathcal{Y}$, [16] suggest to relate the element-wise bounds to each other. Their suggestion for the multivariate normal distribution is adopted here, leading to the following constraints of positive definiteness (p.d.):

$$\frac{1}{n^{(0)}}\mathbf{\Lambda}^{(0)} \quad \text{p.d.}, \qquad \text{and} \tag{16}$$

$$\frac{1}{n^{(0)}}\left(\mathbf{\Lambda}^{(0)} - \frac{1}{n^{(0)}}\mathbf{\Lambda}^{(0)}\beta^{(0)}\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\right) \quad \text{p.d.} \tag{17}$$

If the normal regression model is to be applied as an imprecise probability model, we have to proceed as follows:

1. Prior knowledge on β must be expressed as a set of values of $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$.
2. This set must be "translated" into a set of values of $y^{(0)}$ in a way such that the resulting set $\mathcal{Y}^{(0)}$ satisfies the constraints (16) and (17).
3. Then each $y^{(0)}$ in $\mathcal{Y}^{(0)}$ is linearly updated by (5) to $y^{(1)}$.
4. The obtained set $\mathcal{Y}^{(1)}$ must be "retranslated" into an interpretable set of values of $\beta^{(1)}$ and $\mathbf{\Lambda}^{(1)}$.

The sets can be defined by lower and upper bounds for each element, e.g., for $\beta^{(0)}$ by

$$\beta_j^{(0)} \in \left[\underline{\beta}_j^{(0)}, \overline{\beta}_j^{(0)}\right] \qquad j = 1, \ldots, p\,.$$

The bounds for the components $\beta_j^{(0)}$ of $\beta^{(0)}$ can be chosen independently of each other, as any vector of reals forms an admissible regression parameter. For $\mathbf{\Lambda}^{(0)}$ the situation is more complex, because all the element-wise bounds $\underline{\lambda}_{ij}^{(0)}$ and $\overline{\lambda}_{ij}^{(0)}$ have to be chosen such that for any combination of values between the bounds the resulting $\mathbf{\Lambda}^{(0)}$ is positive definite. Choosing bounds for the precision matrix $\mathbf{\Lambda}^{(0)}$ instead of bounds for $\mathbf{\Sigma}^{(0)}$ facilitates the "translation" issues in application to a great extent.[14]

In the "translation" step the bounds on $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$ must be turned into bounds on $y^{(0)}$ that have to satisfy conditions (16) and (17). For $y_a^{(0)}$, this is simple, as multiplying by $\frac{1}{n^{(0)}}$ does not change positive definiteness. But deriving bounds on $y_b^{(0)}$ is more difficult, as it holds that

$$\underline{y}_{bi}^{(0)} = \min_{\beta^{(0)}, \mathbf{\Lambda}^{(0)}} \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{ij}^{(0)} \beta_j^{(0)}$$

$$\overline{y}_{bi}^{(0)} = \max_{\beta^{(0)}, \mathbf{\Lambda}^{(0)}} \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{ij}^{(0)} \beta_j^{(0)}\,.$$

The minima and maxima are to be taken over a joint set of $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$ that satisfies the constraint (17). Note that for obtaining the bounds for a *single* $y_{bi}^{(0)}$ the bounds of all elements of $\beta^{(0)}$ and of the i-th row on $\mathbf{\Lambda}^{(0)}$ have to be taken into account on the one hand, but on the other hand maximization and minimization must be executed only on combinations of all values between these bounds that are admissible according to (17). The obstacle is that (16) and (17) are nonlinear constraints (polynomial of degree p when checking wether all eigenvalues are positive), so that $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$ can hardly be calculated analytically. The satisfaction of the highly complex constraint (17) can be taken into account when "translating" to $\mathcal{Y}^{(0)}$ or already when defining the sets for $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$.

3.3 The Case of Two Regressors

In order to be able to give vividly interpretable analytical expressions, we now focus attention on the case of two regressors. Here, (16) turns out to demand only that, for any given $\lambda_{11}^{(0)}$ and $\lambda_{22}^{(0)}$, $\lambda_{12}^{(0)}$ must be chosen such that it leads to a correct non-deterministic correlation ρ. Still, with the five parameters $\beta_1^{(0)}$, $\beta_2^{(0)}$, $\lambda_{11}^{(0)}$, $\lambda_{12}^{(0)}$, $\lambda_{22}^{(0)}$ in this model, (17) turns out be quite complex, leading to an inequality in six parameters (the above five plus $n^{(0)}$) that does not seem to produce an easily interpretable condition on their choice.

Therefore a further simplification was made by assuming $\lambda_{11}^{(0)} = \lambda_{22}^{(0)} =: a$ and $\lambda_{12}^{(0)} = 0$. Then, (16) is trivially satisfied and (17) requires only

$$a\left({\beta_1^{(0)}}^2 + {\beta_2^{(0)}}^2\right) < n^{(0)}\,. \tag{18}$$

If the bounds for $\beta_1^{(0)}$, $\beta_2^{(0)}$ and a are chosen such that all possible combinations of values satisfy this

[14] Defining the bounds for $\mathbf{\Lambda}^{(0)}$ is in fact not as complicated as it might seem as the elements are interpretable in a quite straightforward way (and maybe even closer to intuition than the elements of $\mathbf{\Sigma}^{(0)}$): According to [17] (who is referring to [29, p. 142ff]), it holds that $\lambda_{ii} = [\mathbb{V}(\beta_i \,|\, \beta_{\setminus i})]^{-1}$, where $\beta_{\setminus i}$ is vector β without element i, and $\lambda_{ij} = -(\lambda_{ii}\lambda_{jj})^{\frac{1}{2}} \cdot \rho(\beta_i, \beta_j \,|\, \beta_{\setminus\{i,j\}})$, with the second factor being the correlation of β_i and β_j conditioned on the linear effect of $\beta_{\setminus\{i,j\}}$.

constraint, minimization and maximization can be performed for every parameter independently. Now, most but not all parameters of the posterior can be specified analytically, and the results to be sketched here[15] will turn out to be highly plausible:

We consider the following prior:

$$\beta \sim \mathrm{N}_2\left(\beta^{(0)},\, \frac{\sigma^2}{a}\mathbf{I}\right),$$

with $a \in A := [\underline{a}, \overline{a}]$, $\underline{a} > 0$ and

$$\beta^{(0)} = \begin{pmatrix}\beta_1^{(0)}\\ \beta_2^{(0)}\end{pmatrix} \in B := \begin{pmatrix} B_1 = [\underline{b}_1, \overline{b}_1] \\ B_2 = [\underline{b}_2, \overline{b}_2]\end{pmatrix}.$$

In the description we jump directly to the "retranslated" results. Denoting the elements of the updated covariance matrix $\mathbf{\Sigma}^{(1)}$ by $\sigma_{ij}^{(1)}$, $i,j = 1,2$, we obtain for any $a \in A$ by abbreviating

$$D = \left(\sum_{l=1}^k x_{l1}^2 + a\right)\left(\sum_{l=1}^k x_{l2}^2 + a\right) - \left(\sum_{l=1}^k x_{l1}x_{l2}\right)^2 :$$

$$\begin{aligned}\sigma_{11}^{(1)} &= D^{-1}\cdot\left(\textstyle\sum_{l=1}^k x_{l2}^2 + a\right)\\ \sigma_{22}^{(1)} &= D^{-1}\cdot\left(\textstyle\sum_{l=1}^k x_{l1}^2 + a\right)\\ \sigma_{12}^{(1)} &= D^{-1}\cdot\left(-\textstyle\sum_{l=1}^k x_{l1}x_{l2}\right).\end{aligned}$$

Their basic properties are summarized in

Remark 3

i) As $\frac{\partial}{\partial a}\sigma_{11}^{(1)}$ *and* $\frac{\partial}{\partial a}\sigma_{22}^{(1)}$ *are always negative, the higher the prior precision* a*, the lower the posterior variance of* β_1 *and* β_2*. The trend of the covariance* $\sigma_{12}^{(1)}$ *depends on the sign of* $\sum_{l=1}^k x_{l1}x_{l2}$*.*

ii) $\lim_{a\to 0}\sigma^2\mathbf{\Sigma}^{(1)} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \mathbb{V}(\hat{\beta}_{LS})$*. Therefore, for* $\underline{a} > 0$ *and monotonicity, it holds that the posterior variance of the regression parameters in the imprecise normal regression model is always smaller than the one of the LS estimator.*

iii) $\lim_{a\to\infty}\sigma^2\mathbf{\Sigma}^{(1)} = \mathbf{0}$*: An infinitely high prior precision causes naturally an infinitely small posterior variance.*

Most of the results on $\beta^{(1)}$ are reported in terms of $\beta_1^{(1)}$ only; by noting the symmetry underlying β_1 and β_2, analogous results for $\beta_2^{(1)}$ are immediately achieved mutatis mutandis. We obtain

$$\beta_1^{(1)} = \frac{1}{D}\left\{\left(\sum_{l=1}^k x_{l2}^2 + a\right)\left[a\cdot b_1 + \sum_{l=1}^k x_{l1}z_l\right] - \left(\sum_{l=1}^k x_{l1}x_{l2}\right)\left[a\cdot b_2 + \sum_{l=1}^k x_{l2}z_l\right]\right\}.$$

[15] See [23, Section 4.3] for a detailed derivation.

As these expressions are linear in b_1 and b_2 and optimizations in B can be taken independently of a, it holds that

$$\beta_1^{(1)} \to \max \text{ for } b_1 \to \overline{b}_1 \text{ and } \begin{cases} b_2 \to \overline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} < 0 \\ b_2 \to \underline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} > 0\end{cases}$$

$$\beta_1^{(1)} \to \min \text{ for } b_1 \to \underline{b}_1 \text{ and } \begin{cases} b_2 \to \underline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} < 0 \\ b_2 \to \overline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} > 0\end{cases}.$$

Unfortunately, calculating $\frac{\partial}{\partial a}\beta_1^{(1)}$ yields neither monotonicity nor an easily interpretable condition so that the bounds for $\beta^{(1)}$ can not be given analytically. But still asymptotic results can be obtained, which are summarized in

Remark 4

i) For any $b_j \in B_j$, $j = 1,2 : \lim_{a\to\infty}\beta^{(1)} = (b_1, b_2)^\mathsf{T}$*, and so, for very high values of* a *implying a very high prior precision, each* $b \in B$ *is updated to a value very near to itself; very high trust in the given prior values in* B *means sticking on the prior values and results in learning from the sample only to a very small extent.*

ii) On the other hand, $\lim_{a\to 0}\beta^{(1)} = \hat{\beta}_{LS}$*: Very low trust in the prior values in* B *results in relying almost only on the information given by the sample, and so, any given* $b \in B$ *will be updated to a value close to the least squares estimate* $\hat{\beta}_{LS}$*.*

On a first view, it is disturbing that none of the above formulae for deriving posterior parameters depends on $n^{(0)}$, the second prior parameter. The reason for this is that in proving Theorem 2, the parameter $n^{(0)}$ had to be introduced 'artificially' to match Relations (3) to (5) for the LUCK-model. When 'retranslating' $y^{(1)}$ into $\beta^{(1)}$ and $\mathbf{\Sigma}^{(1)}$, the parameter $n^{(1)}$ is eliminated immediately, and, as a consequence, the dependency on $n^{(0)}$ seems to vanish. In fact, the posterior bounds do actually depend on $n^{(0)}$ via Equation (18). Through this restriction on the prior bounds, the range of posterior bounds is constrained. When using the imprecise normal regression model, the bounds for B are quite easy to derive; a possible strategy is then to set a value for $n^{(0)}$ according to the interpretation as pseudocounts or sample size equivalent, and then to determine $\overline{a}$ from (18).

4 Results Based on Simulated Data

To illustrate the performance of the two-parameter model developed in Section 3.3, three data sets were simulated, each with 20 observations, but with a different arrangement of parameters. For data set 1,

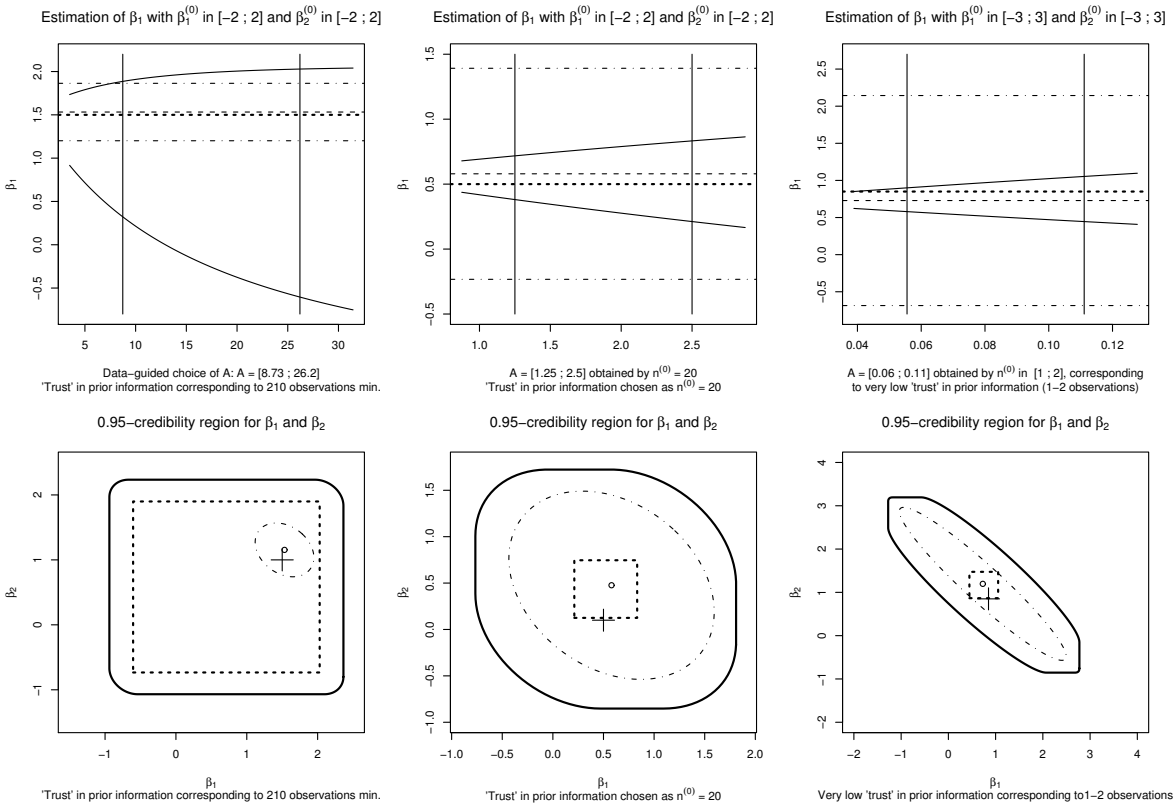

Figure 1: Exemplary results for two-regressor models based on three simulated data sets with 20 observations each.

realizations x_1 and x_2 of two independent standard normal variables were generated as regressors; the error ε was simulated with variance $\sigma^2 = 0.5$. Then the response z was calculated by choosing $\beta_1 = 1.5$, $\beta_2 = 1$. Data set 2 was generated analogously but with $\beta_1 = 0.5$, $\beta_2 = 0.1$ and $\sigma^2 = 3$. In data set 3, multi-collinearity was modeled by simulating x_1 and x_2 by a two-dimensional normal with correlation $\rho = 0.9$, taking $\beta_1 = \beta_2 = 0.85$ and $\sigma^2 = 1$ for calculating z. The regressors were standardized and z was centered with the observed moments in order to make the estimation of an additional intercept unnecessary. Exemplary results are shown in Figure 1, where the graphs from the left to the right show results for data set 1, 2 and 3, respectively.

The upper graphs show the estimation of β_1 for each data set. In each of these graphs, the thick short-dashed line represents the actual value of β_1, the thinner dashed line the LS estimate, and dot-dashed lines indicate the bounds of the 0.95% confidence interval for the LS estimate. The 'horizontal' solid lines represent the estimated lower and upper bound for $\beta_1^{(1)}$ as a function of a, and the vertical lines mark the chosen values of $\underline{a}$ and $\overline{a}$. The lower graphs compare the classical ellipsoid confidence region (dash-dotted line) for the LS estimate of β_1 and β_2 (indicated by the small circle) with the interval-valued estimate (thick short-dashed line) and a 0.95-credibility region for it (thick solid line). The actual value of (β_1, β_2) is indicated by the big cross.

For the "large β, small σ^2" data set 1, relatively high values of a were chosen 'data-guided' by taking the estimated variance of the LS estimator to calculate a central value of A. Because standardized regression parameters are to be estimated, their absolute value is interpretable, and the choice of $B_1 = B_2 = [-2\,;\,2]$ seems reasonable, as higher values are very rare in application. Note that the course of the 'horizontal' solid lines illustrates clearly the statement in Remark 4: The prior assignment results in a quite broad posterior interval for $\beta_1^{(1)}$ (lowest and highest intersection of vertical with 'horizontal' solid lines), as the induced value of $n^{(0)} = 210$ is quite high with respect to the sample size of 20. Consequently, the interval-valued estimate displayed in the lower graph covers a wide area compared to the frequentist confidence region. So does the 0.95-credibility region, which was approximated by the union of 0.95-credibility regions for all combinations of β_1 and β_2 in the interval-valued estimate. Because the maximum posterior variance is lower than the variance for the LS estimate (as mentioned in Remark 3), the distance between the bounds

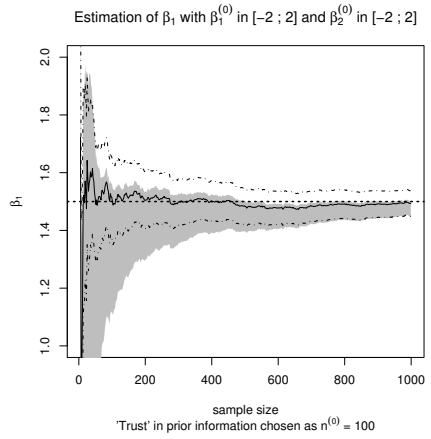

Figure 2: Illustration of asymptotic behavior of interval-valued regression parameter estimates.

of the interval-valued estimate and its credibility region is smaller than the distance between the LS estimate and its confidence region.

In the "small β, large σ^2" situation of data set 2, $\overline{a}$ was chosen such that $n^{(0)} = 20$, implying that the influence of prior and data information are evenly balanced. $\underline{a}$ was chosen ad hoc as $0.5 \cdot \overline{a}$ to illustrate the effect of values of $\underline{a} > 0$ on the variance of $\beta^{(1)}$. (Here, the choice of $\underline{a}$ has no influence on the interval-valued estimates for $\beta^{(1)}$.) Now, as smaller values of $\overline{a}$ result in shorter intervals for $\beta^{(1)}$ (which can be seen clearly from each of the top graphs), the resulting interval-valued estimate for β_1 is less wide, being now shorter than the confidence interval that is quite wide due to the high value of σ^2. This can be seen also in the lower graph, where the confidence region and the credibility region differ to a much lesser extent than in the lower left graph.

For the analysis of data set 3 with a "moderate β_1 and β_2, moderate σ^2" arrangement, A was chosen by using values for $n^{(0)}$ that are commonly suggested for s in the IDM to represent prior ignorance. So $\underline{a}$ was derived from $n^{(0)} = 1$, and $\overline{a}$ from $n^{(0)} = 2$. Together with, as a precaution, even wider prior intervals $B_1 = B_2 = [-3\,;\,3]$, this still yields a very short posterior interval for β_1, as can be seen in the top right graph. Note the exceedingly wide confidence interval for the LS estimate, as this shows the troublesome property of the LS estimate in the case of multi-collinearity: the high resulting variance of estimates can, in many cases, even cause the 'observed' estimates having the wrong sign. In the lower right graph, both the confidence as well as the credibility region show the effect of multi-collinearity through their diagonal shape: estimates for β_1 and β_2 are negatively correlated because both x_1 and x_2 contain similar information. Still, the interval-valued estimate covers a quite small area around the LS estimate, illustrating again the results achieved for the limiting case $a \to 0$.

In Figure 2, asymptotic behavior of the interval-valued estimation for β_1 is illustrated using the situation of data set 1 and choosing $n^{(0)} = 100$. With increasing sample size $k = \dim(z)$, the range of the interval, marked as a gray colored vertical line for each value of k, is becoming shorter and shorter, tightening around the LS estimate, represented by the thin solid line, which approaches the actual value of β_1, marked by the thick short-dashed horizontal line. The dot-dashed lines indicate again the bounds of the 0.95% confidence interval for the LS estimate.

5 The AIRGENE Study

In addition, the model was applied to a data set that is a part of the data collected for the AIRGENE study [15], an EU financed panel study which was conducted to assess the association between air pollutants and inflammation markers in the high-risk group of myocardial infarction survivors. As epidemiological studies show that inflammation markers are associated with the BMI (Body-Mass-Index) and the age of subjects [19], their influence on inflammation marker levels must be taken into account when estimating the effect of air pollutants. To this end, estimations for the parameters of these interfering factors (confounders) are derived in a separate regression model and then are used – in the main analysis not to be presented here – to adjust the main model that estimates the influence of air pollutant variables.

Here, the 200 cases collected by KORA [11] in Augsburg, which was one of the six study centers, are analyzed. The reduced data set consists of the variables `bmi` and `age` as regressors and `log(fib)` as the response variable, being the log of the concentration of the inflammation marker fibrinogen, averaged over the several blood samples collected for each subject during the study period.

Just as for the simulated data sets, the response was centered and the regressors standardized to make an estimation of an intercept unnecessary. Prior bounds for $\beta_{\texttt{bmi}}$ and $\beta_{\texttt{age}}$ were derived each in a straightforward way by considering the lowest and highest possible values (e.g. for `age`, these were, according to the inclusion criterion of the study, 35 and 80 respectively) that were transformed on the standardized scale and then linked to the range of the centered response.When choosing A in the same way as for data set 3 to model weak prior knowledge, the retransformed interval-valued estimates can be combined to the following regression equation:

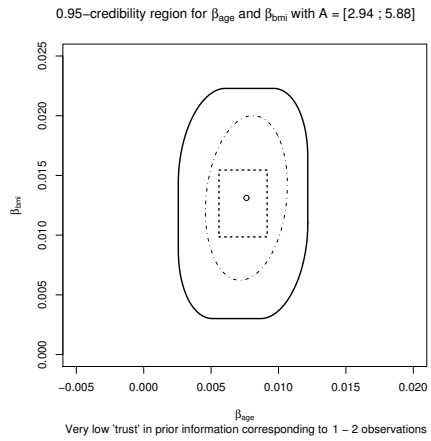

Figure 3: Exemplary results for the AIRGENE data.

$$\begin{aligned}\log(\texttt{fib})_i = \quad & \texttt{age}_i \cdot [0.00558,\, 0.00915] \\ & + \texttt{bmi}_i \cdot [0.00985,\, 0.01545] \\ & + \qquad [0.180,\, 0.562] \qquad\qquad + \varepsilon_i\end{aligned}$$

The fact that the 0.95-credibility region displayed in Figure 3 does not cover the origin is a strong hint that, also when considering complex uncertainty in the prior, `age` and `bmi` have a noteworthy effect on the fibrinogen level. So, the established evidence on this association can be confirmed and set on a more stable base with respect to the model assumptions.

6 Concluding Remarks

We have suggested a first approach to linear regression with imprecise conjugate priors. Of course, the approach needs further investigation, including a comparison to modifications and alternative ways to proceed. This applies in particular to the approach briefly described at the beginning of Section 3.1 where an (imprecise) LUCK-model is constructed directly along the lines of [16].

Further research should also clarify whether other powerful models generalizing classical Bayesian inference in the i.i.d. case (like [14, 7, 5]) can also be extended to linear regression models by similar arguments. The results should also be compared with the approach currently being developed by [28, Chapter 13], whose so-called symmetric theory based on logical probabilities ([28], see also [27]) allows the derivation of probability distributions on parameters without prior modelling.

A possible drawback of the approach introduced by [16], which consequently is shared by the models developed here, is that, in some sense, it does not entirely utilize the expressive power of imprecise probabilities: As $n^{(0)}$ is fixed (like s in the IDM), the behavior of the model – outside the situation of prior ignorance – is not optimal in the case of prior-data conflict in the sense of [21, p. 6]. To see this, note that, if in the situation of Section 2.2 $y^{(0)}$ varies between $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$, then the difference between the updated bounds $\overline{y}^{(1)}$ and $\underline{y}^{(1)}$ is given by $\frac{n^{(0)}\left(\overline{y}^{(0)}-\underline{y}^{(0)}\right)}{n^{(0)}+n}$. So the imprecision decreases by the same amount for any sample of size n, irrespectively whether or not there is substantial discrepancy between prior assignments and the sample. A natural attempt to find a way out would be to vary $n^{(0)}$ in addition. This idea still has to be explored, but the model developed in [21, Ch. 5.4], where such effects are described for an IDM with two categories, may give some hint.

From the applied point of view it is quite important to extend the modelling to generalized linear models, which in particular allow regression analysis for non-metric responses. Here the adaption of auxiliary variable models, considered by [10] in a simulation-based classical Bayesian setting, seems to be very promising.

Acknowledgements

We thank the GSF, KORA and the AIRGENE study group for the help with the data set used for illustration. We are very grateful to the three referees for very helpful and stimulating remarks.

References

[1] J. Abellán and S. Moral. Upper Entropy of Credal Sets. Applications to credal classification. *Intern. J. Approx. Reasoning*, 39: 235–255, 2005.

[2] J.-M. Bernard. An Introduction to the Imprecise Dirichlet Model for Multinomial Data. *Intern. J. Approx. Reasoning*, 39: 123–150, 2005.

[3] J.-M. Bernard. Special Issue on the Imprecise Dirichlet Model. *Intern. J. Approx. Reasoning*, to appear.

[4] J. Bernardo and A. Smith. *Bayesian Theory.* Wiley & Sons, 1993.

[5] A. Boratyńska. Stability of Bayesian Inference in Exponential Families. *Stat. Probabil. Lett.*, 36: 173–178, 1997.

[6] A. Chaturvedi. Robust Bayesian analysis of the linear regression model. *J. Stat. Plan. Inf.*, 50: 175–186, 1996.

[7] F. P. A. Coolen. Imprecise Conjugate Prior Densities for the One-parameter Exponential Family of Distributions. *Stat. Probabil. Lett.*, 16: 337–342, 1993.

[8] F. P. A. Coolen and P. Coolen-Schrijner. Nonparametric Predictive Comparison of Proportions. *J. Stat. Plan. Inf.*, 137: 23–33, 2007.

[9] L. Fahrmeir and G. Tutz. *Multivariate Statistical Models Based on Generalized Linear Models.* Springer, 2001.

[10] C. C. Holmes and L. Held. Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*, 1: 145–168, 2006.

[11] H. Löwel, C. Meisinger, M. Heier, and A. Hormann. The Population-based Acute Myocardial Infarction (AMI) Registry of the MONICA/KORA Study Region of Augsburg. *Gesundheitswesen*, 67, Supplement 1: S31–S37, 2005.

[12] M. Lavine. An approach to evaluating sensitivity in Bayesian regression analyses. *J. Stat. Plan. Inf.*, 40: 233–244, 1994.

[13] A. O'Hagan. *Bayesian Inference*. Kendall´s Advanced Theory of Statistics Vol. 2B, Arnold, 1994.

[14] L. P. Pericchi and P. Walley. Robust Bayesian Credible Intervals and Prior Ignorance. *Int. Stat. Rev.*, 58: 1–23, 1991.

[15] A. Peters, A. Schneider, S. Greven, T. Bellander, F. Forastiere, A. Ibald-Mulli, T. Illig, B. Jacquemin, K. Katsouyanni, W. Koenig, T. Lanki, J. Pekkanen, G. Pershagen, S. Picciotto, R. Rückerl, A. Schaffrath-Rosario, C. Stefanadis, and J. Sunyer. Air Pollution and Inflammatory Response in Myocardial Infarction Survivors: Gene-environment-interactions in a High-risk Group: Study Design of the AIRGENE Study. *Inhalation Toxicology*, in press.

[16] E. Quaeghebeur and G. de Cooman. Imprecise Probability Models for Inference in Exponential Families. In: F. G. Cozman, R. Nau, and T. Seidenfeld (eds.) *Proc. of the 4th Intern. Symp. on Imprecise Probabilities and their Applications (ISIPTA'05)*, 2005.

[17] F. Reithinger. Zusammenhangsstrukturen, *Technical Report, Department of Statistics, LMU Munich*, 2006.
`http://www.statistik.lmu.de/~flo/ss06/strukturen.pdf`

[18] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[19] B. Thorand, J. Baumert, A. Döring, C. Herder, H. Kolb, W. Rathmann, G. Giani, and W. Koenig; Kora Group. National Research Center for Environment and Health, Institute of Epidemiology, Neuherberg, Germany. Sex Differences in the Relation of Body Composition to Markers of Inflammation, *Atherosclerosis*, 184: 216–224, 2006.

[20] M. C. M. Troffaes. Decision Making Under Uncertainty using Imprecise Probabilities. *Intern. J. Approx. Reasoning*, in press.

[21] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[22] P. Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *J. Roy. Stat. Soc. B*, 58: 3–57, 1996.

[23] G. Walter. Robuste Bayes-Regression mit Mengen von Prioris — Ein Beitrag zur Statistik unter komplexer Unsicherheit. *Diploma thesis, Department of Statistics, LMU Munich,* 2006.
`http://www.statistik.lmu.de/~thomas/team/diplomathesis_GeroWalter.pdf`

[24] G. Walter. The Normal Regression Model as a LUCK-model. *Discussion Paper*, 2007.
`http://www.statistik.lmu.de/~thomas/team/isipta07_proof.pdf`

[25] G. Walter. Sketch of an Alternative Approach to Linear Regression Analysis under Sets of Conjugate Priors. *Discussion Paper*, 2007.
`http://www.statistik.lmu.de/~thomas/team/isipta07_conjugate_prior.pdf`

[26] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung, volume I: Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, 2001.

[27] K. Weichselberger. The Logical Concept of Probability and Statistical Inference. In: F. G. Cozman, R. Nau, and T. Seidenfeld (eds.) *Proc. of the 4th Intern. Symp. on Imprecise Probabilities and their Applications (ISIPTA'05)*, 396–405, 2005.

[28] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung, volume II: Symmetrische Wahrscheinlichkeitstheorie*, in cooperation with A. Wallner. In preparation, 2008.

[29] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley & Sons, 1990.

[30] M. Zaffalon, K. Wesnes, and O. Petrini. Reliable Diagnoses of Dementia by the Naive Credal Classifier Inferred from Incomplete Cognitive Data. *Artif. Intell. Med.*, 29: 61–79, 2003.