Bayesian Linear Regression — Different Conjugate Models and Their (In)Sensitivity to Prior-Data Conflict

Gero Walter and Thomas Augustin

Abstract The paper is concerned with Bayesian analysis under prior-data conflict, i.e. the situation when observed data are rather unexpected under the prior (and the sample size is not large enough to eliminate the influence of the prior). Two approaches for Bayesian linear regression modeling based on conjugate priors are considered in detail, namely the standard approach also described in Fahrmeir et al. (2007) and an alternative adoption of the general construction procedure for exponential family sampling models. We recognize that – in contrast to some standard i.i.d. models like the scaled normal model and the Beta-Binomial / Dirichlet-Multinomial model, where prior-data conflict is completely ignored – the models may show some reaction to prior-data conflict, however in a rather unspecific way. Finally we briefly sketch the extension to a corresponding imprecise probability model, where, by considering sets of prior distributions instead of a single prior, prior-data conflict can be handled in a very appealing and intuitive way.

Key words: Linear regression; conjugate analysis; prior-data conflict; imprecise probability

1 Introduction

Regression analysis is a central tool in applied statistics that aims to answer the omnipresent question how certain variables (called covariates / confounders, regressors, stimulus or independent variables, here denoted by x) influence a certain outcome (called response or dependent variable, here denoted by z). Due to the complexity of real-life data situations, basic linear regression models, where the

Gero Walter and Thomas Augustin
Institut für Statistik, Ludwig-Maximilians-Universität München, D-80539 München, Germany, e-mail: gero.walter@stat.uni-muenchen.de, thomas.augustin@stat.uni-muenchen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_4,

expectation of the outcome z_i simply equals the linear predictor $x_i^T\beta$, have been generalized in numerous ways, ranging from generalized linear models (Fahrmeir & Tutz (2001), see also Fahrmeir & Kaufmann (1985) for classical work on asymptotics) for non-normal distributions of $z_i \mid x_i$, or linear mixed models allowing the inclusion of clustered observations, over semi- and nonparametric models (Kauermann et al. 2009, Fahrmeir & Raach 2007, Scheipl & Kneib 2009), up to generalized additive (mixed) models and structured additive regression (Fahrmeir & Kneib 2009, Fahrmeir & Kneib 2006, Kneib & Fahrmeir 2007).

Estimation in such highly complex models may be based on different estimation techniques such as (quasi-) likelihood, general estimation equations (GEE) or Bayesian methods. Especially the latter offer in some cases the only way to attain a reasonable estimate of the model parameters, due to the possibility to include some sort of prior knowledge about these parameters, for instance by "borrowing strength" (e.g., Higgins & Whitehead 1996).

The tractability of large scale models with their ever increasing complexity of the underlying models and data sets should not obscure that still many methodological issues are a matter of debate. Since the early days of modern Bayesian inference one central issue has, of course, been the potentially strong dependence of the inferences on the prior. In particular in situations where data is scarce or unreliable, the actual estimate obtained by Bayesian techniques may rely heavily on the shape of prior knowledge, expressed as prior probability distributions on the model parameters. Recently, new arguments came into this debate by new methods for detecting and investigating *prior-data conflict* (Evans & Moshonov 2006, Bousquet 2008), i.e. situations where "... the observed data is surprising in the light of the sampling model and the prior, [so that] ... we must be at least suspicious about the validity of inferences drawn." (Evans & Moshonov 2006, p. 893)

The present contribution investigates the sensitivity of inferences on potential prior-data conflict: What happens in detail to the posterior distribution and the estimates derived from it if prior knowledge and what the data indicates are severely conflicting? If the sample size n is not sufficiently large to discard the possibly erroneous prior knowledge and thus to rely on data only, prior-data conflict should affect the inference and should – intuitively and informally – result in an increased degree of uncertainty in posterior inference. Probably most statisticians would thus expect a higher variance of the posterior distribution in situations of prior-data conflict.

However, this is by no means automatically the case, in particular when adopting conjugate prior models, which are often used when data are scarce, where only strong prior beliefs allow for a reasonably precise answer in inference. Two simple and prominent examples of complete insensitivity to prior-data conflict are recalled in Section 2: i.i.d. inferences on the mean of a scaled normal distribution and on the probability distribution of a categorical variable by the Dirichlet-Multinomial model.

Sections 3 and 4 extend the question of (in)sensitivity to prior-data to regression models. We confine attention to linear regression analysis with conjugate priors, because – contrary to the more advanced regression model classes – the linear model still allows a fully analytical access, making it possible to understand potential re-

strictions imposed by the model in detail. We discuss and compare two different conjugate models:

(i) the standard conjugate prior (SCP, Section 3) as described in Fahrmeir et al. (2007) or, in more detail, in O'Hagan (1994); and

(ii) a conjugate prior, called "canonically constructed conjugate prior" (CCCP, Section 4) in the following, which is derived by a general method used to construct conjugate priors to sample distributions that belong to a certain class of exponential families, described, e.g., in Bernardo & Smith (1994).

Whereas the former is the more general prior model, allowing for a very flexible modeling of prior information (which might be welcome or not), the latter allows only a strongly restricted covariance structure for β, however offering a clearer insight in some aspects of the update process.

In a nutshell, the result is that both conjugate models do react to prior-data conflict by an enlarged factor to the variance-covariance matrix of the distribution on the regression coefficients β; however, this reaction is unspecific, as it affects the variance and covariances of all components of β in a uniform way – even if the conflict occurs only in one single component.

Probably such an unspecific reaction of the variance is the most a (classical) Bayesian statistician can hope for, and traditional probability theory based on precise probabilities can offer. Indeed, Kyburg (1987) notes, that

> [...] there appears to be no way, within the theory, of distinguishing between the cases in which there are good statistical grounds for accepting a prior distribution, and cases in which the prior distribution reflects merely ungrounded personal opinion.

and the same applies, in essence, to the posterior distribution.

A more sophisticated modeling would need a more elaborated concept of imprecision than is actually provided by looking at the variance (or other characteristics) of a (precise) probability distribution. Indeed, recently the theory of imprecise probabilities (Walley 1991, Weichselberger 2001) is gaining strong momentum. It emerged as a general methodology to cope with the multidimensional character of uncertainty, also reacting to recent insights and developments in decision theory (see Hsu et al. (2005) for a neuro science corroboration of the constitutive difference of stochastic and non-stochastic aspects of uncertainty in human decision making, in the tradition of Ellsberg's (1961) seminal experiments) and artificial intelligence, where the exclusive role of probability as a methodology for handling uncertainty has eloquently been rejected (Klir & Wierman 1999):

> For three hundred years [...] uncertainty was conceived solely in terms of probability theory. This seemingly unique connection between uncertainty and probability is now challenged [... by several other] theories, which are demonstrably capable of characterizing situations under uncertainty. [...]
>
> [...] it has become clear that there are several distinct types of uncertainty. That is, it was realized that uncertainty is a multidimensional concept. [... That] multidimensional nature of uncertainty was obscured when uncertainty was conceived solely in terms of probability theory, in which it is manifested by only one of its dimensions.

Current applications include, among many other, risk analysis, reliability modeling and decision theory, see de Cooman et al. (2007), Augustin et al. (2009) and

Coolen-Schrijner et al. (2009) for recent collections on the subject. As a welcome byproduct imprecise probability models also provide a formal superstructure on models considered in robust Bayesian analysis (Ríos Insua & Ruggeri 2000) and frequentist robust statistic in the tradition of Huber & Strassen (1973), see also Augustin & Hable (2009) for a review.

By considering *sets* of distributions, and corresponding interval-valued probabilities for events, imprecise probability models allow to express the quality of the underlying knowledge in an elegant way. The higher the ambiguity, the larger c.p. the sets. The traditional concept of probability is contained as a special case, appropriate if and only if there is perfect stochastic information. This methodology allows also for a natural handling of prior-data conflict. If prior and data are in conflict, the set of posterior distributions are enlarged, and inferences become more cautious.

In Section 5 we briefly report that the CCCP model has a structure that allows a direct extension to an imprecise probability model along the lines of Quaeghebeur & de Cooman's (2005) imprecise probability models for i.i.d. exponential family models. Extending the models further by applying arguments from Walter & Augustin (2009) yields a powerful generalization of the linear regression model that is also capable of a component-specific reaction to prior-data conflict.

2 Prior-data Conflict in the i.i.d. Case

As a simple demonstration that conjugate models might not react to prior-data conflict reasonably, inference on the mean of data from a scaled normal distribution and inference on the category probabilities in multinomial sampling will be described in the following examples 1 and 2.

Example 1 (Samples from a scaled Normal distribution $\mathrm{N}(\mu,1)$*).* The conjugate distribution to an i.i.d.-sample x of size n from a scaled normal distribution with mean μ, denoted by $\mathrm{N}(\mu,1)$ is a normal distribution with mean $\mu^{(0)}$ and variance ${\sigma^{(0)}}^2$[1]. The posterior is then again a normal distribution with the following updated parameters:

$$\mu^{(1)} = \frac{\frac{1}{n}}{\frac{1}{n}+{\sigma^{(0)}}^2}\mu^{(0)} + \frac{{\sigma^{(0)}}^2}{\frac{1}{n}+{\sigma^{(0)}}^2}\bar{x} = \frac{\frac{1}{{\sigma^{(0)}}^2}}{\frac{1}{{\sigma^{(0)}}^2}+n}\mu^{(0)} + \frac{n}{\frac{1}{{\sigma^{(0)}}^2}+n}\bar{x} \qquad (1)$$

$${\sigma^{(1)}}^2 = \frac{{\sigma^{(0)}}^2 \cdot \frac{1}{n}}{{\sigma^{(0)}}^2+\frac{1}{n}} = \frac{1}{\frac{1}{{\sigma^{(0)}}^2}+n}. \qquad (2)$$

[1] Here, and in the following, parameters of a prior distribution will be denoted by an upper index $^{(0)}$, whereas parameters of the respective posterior distribution by an upper index $^{(1)}$.

The posterior expectation (and mode) is thus a simple weighted average of the prior mean $\mu^{(0)}$ and the estimation from data $\bar{x}$, with weights $\frac{1}{{\sigma^{(0)}}^2}$ and n, respectively.[2] The variance of the posterior distribution is getting smaller automatically.

Now, in a situation where data is scarce but with prior information one is very confident about, one would choose a low value for ${\sigma^{(0)}}^2$, thus resulting in a high weight for the prior mean $\mu^{(0)}$ in the calculation of $\mu^{(1)}$. The posterior distribution will be centered around a mean between $\mu^{(0)}$ and $\bar{x}$, and it will be even more pointed as the prior, because ${\sigma^{(1)}}^2$ is considerably smaller than ${\sigma^{(0)}}^2$ as the factor to ${\sigma^{(0)}}^2$ in (2) is quite smaller than one.

The posterior basically would thus say that one can be quite sure that the mean μ is around $\mu^{(1)}$, regardless if $\mu^{(0)}$ and $\bar{x}$ were near to each other or not, where the latter would be a strong hint on prior-data conflict. The posterior variance does not depend on this; the posterior distribution is thus insensitive to prior-data conflict.

Even if one is not so confident about one's prior knowledge and thus assigning a relatively large variance to the prior, the posterior mean is less strongly influenced by the prior mean, but the posterior variance still is getting smaller no matter if the data support the prior information or not.

The same insensitivity appears also in the widely used Dirichlet-Multinomial model:

Example 2 (Samples from a Multinomial distribution $M(\theta)$). Given a sample of size n from a multinomial distribution with probabilities θ_j for categories / classes $j = 1,\ldots,k$, subsumed in the vectorial parameter θ (with $\sum_{j=1}^{k}\theta_j = 1$), the conjugate prior on θ is a Dirichlet distribution $\mathrm{Dir}(\alpha^{(0)})$. Written in terms of a reparameterization used e.g. in Walley (1996), $\alpha_j^{(0)} = s^{(0)} \cdot t_j^{(0)}$ such that $\sum_{j=1}^{k} t_j^{(0)} = 1$, $(t_1^{(0)},\ldots,t_k^{(0)})^\top =: t^{(0)}$, it holds that the components of $t^{(0)}$ have a direct interpretation as prior class probabilities, whereas $s^{(0)}$ is a parameter indicating the confidence in the values of $t^{(0)}$, similar to the inverse variance as in Example 1, and the quantity $n^{(0)}$ in Section 4.[3]

The posterior distribution, obtained after updating via Bayes' rule with a sample vector $(n_1,\ldots,n_k)$, $\sum_{j=1}^{k} n_j = n$ collecting the observed counts in each category, is a Dirichlet distribution with parameters

$$t_j^{(1)} = \frac{s^{(0)}}{s^{(0)}+n} t_j^{(0)} + \frac{n}{s^{(0)}+n} \cdot \frac{n_j}{n}, \qquad s^{(1)} = s^{(0)} + n.$$

The posterior class probabilities $t^{(1)}$ are calculated as a weighted mean of the prior class probabilities and $\frac{n_j}{n}$, the proportion in the sample, with weights $s^{(0)}$ and n, respectively; the confidence parameter is incremented by the sample size n.

Also here, there is no systematic reaction to prior-data conflict. The posterior variance for each class probability θ_j calculates as

[2] The reason for using these seemingly strange weights will become clear later.

[3] If $\theta \sim \mathrm{Dir}(s,t)$, then $\mathbb{V}(\theta_j) = \frac{t_j(1-t_j)}{s+1}$. If s is high, then the variances of θ will become low, thus indication high confidence in the chosen values of t.

$$\mathbb{V}(\theta_j \mid n) = \frac{t_j^{(1)}(1-t_j^{(1)})}{s^{(1)}+1} = \frac{t_j^{(1)}(1-t_j^{(1)})}{s^{(0)}+n+1}\,.$$

The posterior variance depends heavily on $t_j^{(1)}(1-t_j^{(1)})$, having values between 0 and $\frac{1}{4}$, which do not change specifically to prior data conflict. The denominator increases from $s^{(0)}+1$ to $s^{(0)}+n+1$. Imagine a situation with strong prior information suggesting a value of $t_j^{(0)} = 0.25$, so one could choose $s^{(0)} = 5$, resulting in a prior class variance of $\frac{1}{32}$. When observing a sample of size $n = 10$ all belonging to class j (thus $n_j = 10$), being in clear contrast to the prior information, the posterior class probability is $t_j^{(1)} = 0.75$, resulting the enumerator value of the class variance to remain constant. Therefore, due to the increasing denominator, the variance decreases to $\frac{3}{256}$, in spite of the clear conflict between prior and sample information. Of course, one can construct situations where the variance increases, but this happens only in case of an update of $t_j^{(0)}$ towards $\frac{1}{2}$. If $t_j^{(0)} = \frac{1}{2}$, the variance will decrease for any degree of prior-data conflict.

3 The Standard Approach for Bayesian Linear Regression (SCP)

The regression model is noted as follows:

$$z_i = x_i^{\mathsf{T}}\beta + \varepsilon_i\,, \quad x_i \in \mathbb{R}^p\,,\ \beta \in \mathbb{R}^p\,,\ \varepsilon_i \sim \mathrm{N}(0,\sigma^2)\,,$$

where z_i is the response, x_i the vector of the p covariates for observation i, and β is the p-dimensional vector of adjacent regression coefficients.

The vector of regressors x_i for each observation i is generally considered to be non-stochastic, thus it holds that $z_i \sim \mathrm{N}(x_i^{\mathsf{T}}\beta,\sigma^2)$, or, for n i.i.d. samples, $z \sim \mathrm{N}(\mathbf{X}\beta,\sigma^2\mathbf{I})$, where $z \in \mathbb{R}^n$ is the column vector of the responses z_i, and $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the *design matrix*. Without loss of generality, one can either assume $x_{i1} = 1\ \forall i$ such that the first component of β is the intercept parameter[4], or consider only centered responses z and standardized covariates to make the estimation of an intercept unnecessary.

In Bayesian linear regression analysis, the distribution of the response z is interpreted as a distribution of z given the parameters β and σ^2, and prior distributions on β and σ^2 must be considered. For this, it is convenient to split the joint prior on β and σ^2 as $p(\beta,\sigma^2) = p(\beta \mid \sigma^2)p(\sigma^2)$ and to consider conjugate distributions for both parts, respectively.

In the literature, the proposed conjugate prior for $\beta \mid \sigma^2$ is a normal distribution with expectation vector $m^{(0)} \in \mathbb{R}^p$ and variance-covariance matrix $\sigma^2\boldsymbol{M}^{(0)}$, where $\boldsymbol{M}^{(0)}$ is a symmetric positive definite matrix of size $p \times p$. The prior on σ^2 is an inverse gamma distribution (i.e., $1/\sigma^2$ is gamma distributed) with parameters $a^{(0)}$ and $b^{(0)}$, in the sense that $p(\sigma^{-2}) \propto (\sigma^{-2})^{a^{(0)}+1}\exp\{-b^{(0)}\sigma^{-2}\}$. The joint prior on

[4] usually denoted by β_0; however, we stay with the numbering $1,\ldots,p$ for the components of β.

$\theta = (\beta, \sigma^2)^\mathsf{T}$ is then denoted as a normal – inverse gamma (NIG) distribution. The derivation of this prior and the proof of its conjugacy can be found, e.g., in Fahrmeir et al. (2007) or in O'Hagan (1994), the latter using a different parameterization of the inverse gamma part, where $a^{(0)} = \frac{d}{2}$ and $b^{(0)} = \frac{a}{2}$.

For the prior model, it holds thus that (if $a^{(0)} > 1$ resp. $a^{(0)} > 2$)

$$\begin{aligned} \mathbb{E}[\beta \mid \sigma^2] &= m^{(0)}\,, \qquad & \mathbb{V}(\beta \mid \sigma^2) &= \sigma^2 \boldsymbol{M}^{(0)}\,, \\ \mathbb{E}[\sigma^2] &= \frac{b^{(0)}}{a^{(0)}-1}\,, \qquad & \mathbb{V}(\sigma^2) &= \frac{(b^{(0)})^2}{(a^{(0)}-1)^2(a^{(0)}-2)}\,. \end{aligned} \tag{3}$$

As σ^2 is considered as nuisance parameter, the unconditional distribution on β is of central interest because it subsumes the shape of prior knowledge on β as expressed by the choice of parameters $m^{(0)}$, $\boldsymbol{M}^{(0)}$, $a^{(0)}$ and $b^{(0)}$. It can be shown that $p(\beta)$ is a multivariate noncentral t distribution with $2a^{(0)}$ degrees of freedom, location parameter $m^{(0)}$ and dispersion parameter $\frac{b^{(0)}}{a^{(0)}}\boldsymbol{M}^{(0)}$, such that

$$\mathbb{E}[\beta] = m^{(0)}\,, \qquad \mathbb{V}(\beta) = \frac{b^{(0)}}{a^{(0)}-1}\boldsymbol{M}^{(0)} = \mathbb{E}[\sigma^2]\boldsymbol{M}^{(0)}\,. \tag{4}$$

The joint posterior distribution $p(\theta \mid z)$, due to conjugacy, is then again a normal – inverse gamma distribution with the updated parameters

$$\begin{aligned} m^{(1)} &= \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\left(\boldsymbol{M}^{(0)^{-1}} m^{(0)} + \mathbf{X}^\mathsf{T} z\right), \\ \boldsymbol{M}^{(1)} &= \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}, \\ a^{(1)} &= a^{(0)} + \frac{n}{2}\,, \qquad b^{(1)} = b^{(0)} + \frac{1}{2}\left(z^\mathsf{T} z + m^{(0)^\mathsf{T}} \boldsymbol{M}^{(0)^{-1}} m^{(0)} - m^{(1)^\mathsf{T}} \boldsymbol{M}^{(1)^{-1}} m^{(1)}\right). \end{aligned}$$

The properties of the posterior distributions can thus be analyzed by inserting the updated parameters into (3) and (4).

3.1 Update of $\beta \mid \sigma^2$

The normal distribution part of the joint prior is updated as follows:

$$\mathbb{E}[\beta \mid \sigma^2, z] = m^{(1)} = \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\left(\boldsymbol{M}^{(0)^{-1}} m^{(0)} + \mathbf{X}^\mathsf{T} z\right) = (\mathbf{I} - \boldsymbol{A})\, m^{(0)} + \boldsymbol{A}\,\hat{\beta}_{LS}\,,$$

where $\boldsymbol{A} = \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$. The posterior estimate of $\beta \mid \sigma^2$ thus calculates as a matrix-weighted mean of the prior guess and the least-squares estimate. The larger the diagonal elements of $\boldsymbol{M}^{(0)}$ (i.e., the weaker the prior information), the smaller the elements of $\boldsymbol{M}^{(0)^{-1}}$ and thus the 'nearer' is A to the identity matrix, so that the posterior estimate is nearer to the least-squares estimate.

The posterior variance of $\beta \mid \sigma^2$ calculates as

$$\mathbb{V}(\beta \mid \sigma^2, z) = \sigma^2 \boldsymbol{M}^{(1)} = \sigma^2 \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\intercal \mathbf{X} \right)^{-1}.$$

As the elements of $\boldsymbol{M}^{(1)^{-1}}$ get larger with respect to $\boldsymbol{M}^{(0)^{-1}}$, the elements of $\boldsymbol{M}^{(1)}$ will, roughly speaking, become smaller than those of $\boldsymbol{M}^{(0)}$, so that the variance of $\beta \mid \sigma^2$ decreases.

Therefore, the updating of $\beta \mid \sigma^2$ is obviously insensitive to prior-data conflict, because the posterior distribution will not become flatter in case of a large distance between $\mathbb{E}[\beta]$ and $\hat{\beta}_{LS}$. Actually, as O'Hagan (1994) derives, for any $\phi = a^\intercal \beta$, i.e., any linear combination of elements of β, it holds that $\mathbb{V}(\phi \mid \sigma^2, z) \leq \mathbb{V}(\phi \mid \sigma^2)$, becoming a strict inequality if $\mathbf{X}$ has full rank. In particular, the variance of each β_i decreases automatically with the update step.

3.2 Update of σ^2

It can be shown (O'Hagan 1994) that

$$\mathbb{E}[\sigma^2 \mid z] = \frac{2a^{(0)} - 2}{2a^{(0)} + n - 2} \mathbb{E}[\sigma^2] + \frac{n-p}{2a^{(0)} + n - 2} \hat{\sigma}^2_{LS} + \frac{p}{2a^{(0)} + n - 2} \hat{\sigma}^2_{PDC}, \qquad (5)$$

where $\hat{\sigma}^2_{LS} = \frac{1}{n-p}(z - \mathbf{X}\hat{\beta}_{LS})^\intercal (z - \mathbf{X}\hat{\beta}_{LS})$ is the least-squares based estimate for σ^2, and $\hat{\sigma}^2_{PDC} = \frac{1}{p}(m^{(0)} - \hat{\beta}_{LS})^\intercal \left(\boldsymbol{M}^{(0)} + (\mathbf{X}^\intercal \mathbf{X})^{-1} \right)^{-1} (m^{(0)} - \hat{\beta}_{LS})$. For the latter it holds that $\mathbb{E}[\hat{\sigma}^2_{PDC} \mid \sigma^2] = \sigma^2$; the posterior expectation of σ^2 calculates thus as a weighted mean of three estimates:

(i) the prior expectation for σ^2,

(ii) the least-squares estimate, and

(iii) an estimate based on a weighted squared difference of the prior mean and the least-squares estimate for β.

The weights depend on $a^{(0)}$ (one prior parameter for the inverse gamma part), the sample size n, and the dimension of β, respectively. The role of the first weight gets more plausible when remembering the formula for the prior variance of σ^2 in (3), where $a^{(0)}$ appears in the denominator. A larger value of $a^{(0)}$ means thus smaller prior variance, in turn giving a higher weight for $\mathbb{E}[\sigma^2]$ in the calculation of $\mathbb{E}[\sigma^2 \mid z]$. The weight to $\hat{\sigma}^2_{LS}$ corresponds to the classical degrees of freedom, $n - p$. With the the sample size approaching infinity, this weight will dominate the others, such that $\mathbb{E}[\sigma^2 \mid z]$ approaches $\hat{\sigma}^2_{LS}$.

Similar results hold for the posterior mode instead of the posterior expectation.

Here, the estimate $\hat{\sigma}^2_{PDC}$ allows some reaction to prior-data conflict: it measures the distance between $m^{(0)}$ (prior) and $\hat{\beta}_{LS}$ (data) estimates for β, with a large distance resulting basically in a large value of $\hat{\sigma}^2_{PDC}$ and thus an enlarged posterior estimate for σ^2. The weighting matrix for the distances is playing an important role as well. The influence of $\boldsymbol{M}^{(0)}$ is as follows: for components of β one is quite certain about the assignment of $m^{(0)}$, the respective diagonal elements of $\boldsymbol{M}^{(0)}$ will be low, so that these

diagonal elements of the weighting matrix will be high. Therefore, large distances in these dimensions will increase t strongly. An erroneously high confidence in the prior assumptions on β is thus penalized by an increasing posterior estimate for σ^2. The influence of $\mathbf{X}^\mathsf{T}\mathbf{X}$ interprets as follows: covariates with a low spread in x-values, giving an unstable base for the estimate $\hat{\beta}_{LS}$, will result in low diagonal elements of $\mathbf{X}^\mathsf{T}\mathbf{X}$. Via the double inverting, those diagonal elements of the weighting matrix will remain low and thus give the difference a low weight. Therefore, $\hat{\sigma}^2_{PDC}$ will not excessively increase due to a large difference in dimensions where the location of $\hat{\beta}_{LS}$ is to be taken cum grano salis. As to be seen in the following subsection, the behavior of $\mathbb{E}[\sigma \mid z]$ is of high importance for posterior inferences on β.

3.3 Update of β

The posterior distribution of β is again a multivariate t, with expectation $\mathbb{E}[\beta \mid z] = \mathbb{E}\big[\mathbb{E}[\beta \mid \sigma^2, z] \mid z\big] = m^{(1)}$ (as described in Section 3.1) and variance

$$\mathbb{V}[\beta \mid z] = \frac{b^{(1)}}{a^{(1)}-1}\boldsymbol{M}^{(1)} = \mathbb{E}[\sigma^2 \mid z]\,\boldsymbol{M}^{(1)} \tag{6}$$

$$= \left(\frac{2a^{(0)}-2}{2a^{(0)}+n-2}\mathbb{E}[\sigma^2] + \frac{n-p}{2a^{(0)}+n-2}\hat{\sigma}^2_{LS} + \frac{p}{2a^{(0)}+n-2}\hat{\sigma}^2_{PDC}\right)\left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$$

$$= \left(\frac{2a^{(0)}-2}{2a^{(0)}+n-2}\mathbb{E}[\sigma^2] + \frac{n-p}{2a^{(0)}+n-2}\hat{\sigma}^2_{LS} + \frac{p}{2a^{(0)}+n-2}\hat{\sigma}^2_{PDC}\right)$$
$$\cdot \left(\boldsymbol{M}^{(0)} - \boldsymbol{M}^{(0)}\mathbf{X}^\mathsf{T}(\mathbf{I} + \mathbf{X}\boldsymbol{M}^{(0)}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}\boldsymbol{M}^{(0)}\right),$$

not being directly expressible as a function of $\mathbb{E}[\sigma^2]\boldsymbol{M}^{(0)}$, the prior variance of β.

Due to the effect of $\mathbb{E}[\sigma^2 \mid z]$, the posterior variance-covariance matrix of β can increase in case of prior data conflict, if the rise of $\mathbb{E}[\beta \mid z]$ (due to an even stronger rise of t) can overcompensate the decrease in the elements of $\boldsymbol{M}^{(1)}$. However, we see that the effect of prior-data conflict on the posterior variance of β is *globally* and not component-specific; it influences the variances for *all* components of β to the same amount even if the conflict was confined only to some or even just one single component. Taking it to the extremes, if the prior assignment $m^{(0)}$ was (more or less) correct in all but one component, with that one being far out, the posterior variances will increase for all components, also for the ones with prior assignments that have turned out to be basically correct.

4 An Alternative Approach for Conjugate Priors in Bayesian Linear Regression (CCCP)

In this section, a prior model for $\theta = (\beta, \sigma^2)$ will be constructed along the general construction method for sample distributions that form a linear, canonical exponential family (see, e.g., Bernardo & Smith 1994). The method is typically used for the i.i.d. case, but the likelihood arising from $z \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$ will be shown to follow the specific exponential family form as well. The canonically constructed conjugate prior (CCCP) model will also result in a normal - inverse gamma distribution, but with a fixed variance - covariance structure. The CCCP model is thus a special case of the SCP model, which – as will be detailed in this section – offers some interesting further insights into the structure of the update step.

The likelihood arising from the distribution of z,

$$
\begin{aligned}
f(z \mid \beta, \sigma^2) &= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp\Big\{ -\frac{1}{2\sigma^2}(z-\mathbf{X}\beta)^{\mathsf{T}}(z-\mathbf{X}\beta)\Big\} \\
&= \underbrace{\frac{1}{(2\pi)^{\frac{n}{2}}}}_{\mathbf{a}(z)} \exp\Big\{ \underbrace{\Big(\frac{\beta}{\sigma^2}\Big)}_{\psi_1}^{\mathsf{T}} \underbrace{\mathbf{X}^{\mathsf{T}}z}_{\tau_1(z)} \underbrace{-\frac{1}{\sigma^2}}_{\psi_2} \underbrace{\frac{1}{2}z^{\mathsf{T}}z}_{\tau_2(z)} \underbrace{-\Big(\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta + \frac{n}{2}\log(\sigma^2)\Big)}_{n\mathbf{b}(\psi)} \Big\},
\end{aligned}
$$

indeed corresponds to the linear, canonical exponential family form

$$
f(z \mid \psi) = \mathbf{a}(z) \cdot \exp\{\langle \psi, \tau(z)\rangle - n \cdot \mathbf{b}(\psi)\},
$$

where $\psi = \psi(\beta, \sigma^2)$ is a certain function of β and σ^2, the parameters on which one wishes to learn. $\tau(z)$ is a sufficient statistic of z used in the update step. Here, we have

$$
\psi = \begin{pmatrix} \frac{\beta}{\sigma^2} \\ -\frac{1}{\sigma^2} \end{pmatrix}, \quad \tau(z) = \begin{pmatrix} \mathbf{X}^{\mathsf{T}}z \\ \frac{1}{2}z^{\mathsf{T}}z \end{pmatrix}, \quad \mathbf{b}(\psi) = \frac{1}{2n\sigma^2}\beta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta + \frac{1}{2}\log(\sigma^2). \tag{7}
$$

According to the general construction method, a conjugate prior for ψ can be obtained from these ingredients by the following equation:

$$
p(\psi) = \mathbf{c}(n^{(0)}, y^{(0)}) \cdot \exp\Big\{ n^{(0)} \cdot [\langle \psi, y^{(0)}\rangle - \mathbf{b}(\psi)] \Big\},
$$

where $n^{(0)}$ and $y^{(0)}$ are the parameters that define the concrete prior distribution of its distribution family; whereas ψ and $\mathbf{b}(\psi)$ were identified in (7). $\mathbf{c}$ corresponds to a normalization factor for the prior. When applying the general construction method to the two examples from Section 2, the very same priors as presented there will result, where $y^{(0)} = \mu^{(0)}$ and $n^{(0)} = 1/{\sigma^{(0)}}^2$ for the prior to the scaled normal model, and $y^{(0)} = t^{(0)}$ and $n^{(0)} = s^{(0)}$ for the prior to the multinomial model.

Here, the conjugate prior writes as

$$p(\psi)d\psi = \mathbf{c}(n^{(0)}, y^{(0)}) \exp\left\{ n^{(0)} \left[y^{(0)\mathsf{T}} \begin{pmatrix} \frac{\beta}{\sigma^2} \\ -\frac{1}{\sigma^2} \end{pmatrix} - \frac{1}{2n\sigma^2} \beta^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \beta - \frac{1}{2} \log(\sigma^2) \right] \right\} d\psi .$$

As this is a prior on ψ, but we want to arrive at a prior on $\theta = (\beta, \sigma^2)^{\mathsf{T}}$, we must transform the density $p(\psi)$. For the transformation, we need the determinant of the Jacobian matrix $\frac{d\psi}{d\theta}$:

$$\left| \det\left(\frac{d\psi}{d\theta} \right) \right| = \left| \det \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{I}_p & -\frac{\beta}{(\sigma^2)^2} \\ \mathbf{0} & \frac{1}{(\sigma^2)^2} \end{pmatrix} \right| = \frac{1}{(\sigma^2)^{p+2}} .$$

Therefore, the prior on $\theta = (\beta, \sigma^2)^{\mathsf{T}}$ is

$$p(\theta)d\theta = p(\psi)d\psi \cdot \left| \det\left(\frac{d\psi}{d\theta} \right) \right| = \mathbf{c}(n^{(0)}, y^{(0)}) \cdot \tag{8}$$

$$\exp\left\{ n^{(0)} y_1^{(0)\mathsf{T}} \frac{\beta}{\sigma^2} - n^{(0)} y_2^{(0)} \frac{1}{\sigma^2} - \frac{n^{(0)}}{2n\sigma^2} \beta^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \beta - \frac{n^{(0)}}{2} \log(\sigma^2) - (p+2) \log(\sigma^2) \right\} .$$

θ can now be shown to follow a normal – inverse gamma distribution by comparing coefficients. In doing that, some attention must be paid to the terms proportional to $-1/\sigma^2$ (appearing as $-\log(\sigma^2)$ in the exponent) because the normal $p(\beta \mid \sigma^2)$ and the inverse gamma $p(\sigma^2)$ will have to 'share' it. Furthermore, it is necessary to complete the square for the normal part, resulting in an additional term for the inverse gamma part.

The density of a normal distribution on $\beta \mid \sigma^2$ with a mean vector $\overline{m}^{(0)} = \overline{m}(n^{(0)}, y^{(0)})$ and a variance-covariance matrix $\sigma^2 \overline{\boldsymbol{M}}^{(0)} = \sigma^2 \overline{\boldsymbol{M}}(n^{(0)}, y^{(0)})$, both to be seen as functions of the canonical parameters $n^{(0)}$ and $y^{(0)}$, has the following form:

$$\begin{aligned} p(\beta \mid \sigma^2) &= \frac{1}{(2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \overline{m}^{(0)} \right)^{\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \left(\beta - \overline{m}^{(0)} \right) \right\} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left\{ \overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2} \beta^{\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \beta \right. \\ &\qquad \left. - \frac{1}{2\sigma^2} \overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \overline{m}^{(0)} - \frac{p}{2} \log(\sigma^2) \right\} . \end{aligned}$$

Comparing coefficients with the terms from (8) depending on β, we get

$$\overline{\boldsymbol{M}}^{(0)^{-1}} = \overline{\boldsymbol{M}}(n^{(0)})^{-1} = \frac{n^{(0)}}{n} \mathbf{X}^{\mathsf{T}} \mathbf{X} , \qquad \overline{m}^{(0)} = \overline{m}(y^{(0)}) = n \left(\mathbf{X}^{\mathsf{T}} \mathbf{X} \right)^{-1} y^{(0)} .$$

With the square completed, the joint density of β and σ^2 reads as

$$p(\beta,\sigma^2) = \mathbf{c}(n^{(0)},y^{(0)})\cdot$$
$$\exp\Big\{\underbrace{n^{(0)}y_1^{(0)\mathsf{T}}\frac{\beta}{\sigma^2} - \frac{n^{(0)}}{2n\sigma^2}\beta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta - \frac{1}{2\sigma^2}\Big(n\cdot n^{(0)}y_1^{(0)\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)}\Big) - \frac{p}{2}\log(\sigma^2)}_{\text{to } p(\beta|\sigma^2) \text{ (normal distribution)}}$$
$$\underbrace{-\frac{1}{\sigma^2}\Big(-\frac{n^{(0)}n}{2}y_1^{(0)\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)}\Big) - n^{(0)}y_2^{(0)}\frac{1}{\sigma^2} - \Big(\frac{n^{(0)}+p}{2}+2\Big)\log(\sigma^2)}_{\text{to } p(\sigma^2) \text{ (inverse gamma distribution)}}\Big\}. \quad (9)$$

Therefore, one part of the conjugate prior (9) reveals as a multivariate normal distribution with mean vector $\overline{m}^{(0)} = \overline{m}(y_1^{(0)}) = n\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\,y_1^{(0)}$ and covariance matrix $\sigma^2\overline{\boldsymbol{M}}^{(0)} = \sigma^2\overline{\boldsymbol{M}}(n^{(0)}) = \frac{n\sigma^2}{n^{(0)}}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, i.e.

$$\beta \mid \sigma^2 \sim \mathrm{N}_p\left(n\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\,y_1^{(0)},\ \frac{n\sigma^2}{n^{(0)}}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\right). \quad (10)$$

The other terms of (9) can be directly identified with the core of an inverse gamma distribution with parameters

$$\overline{a}^{(0)} = \frac{n^{(0)}+p}{2}+1 \qquad \text{and}$$
$$\overline{b}^{(0)} = n^{(0)}y_2^{(0)} - \frac{n^{(0)}}{2}y_1^{(0)\mathsf{T}}n(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)} = n^{(0)}y_2^{(0)} - \frac{1}{2}\overline{m}^{(0)\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)},$$

$$\text{i.e.,}\quad \sigma^2 \sim \mathrm{IG}\left(\frac{n^{(0)}+p+2}{2},\ n^{(0)}y_2^{(0)} - \frac{n^{(0)}}{2}y_1^{(0)\mathsf{T}}n(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)}\right). \quad (11)$$

We have thus derived the CCCP distribution on (β,σ^2), which can be expressed either in terms of the canonical prior parameters $n^{(0)}$ and $y^{(0)}$ or in terms of the prior parameters from Section 3, $\overline{m}^{(0)}$, $\overline{\boldsymbol{M}}^{(0)}$, $\overline{a}^{(0)}$ and $\overline{b}^{(0)}$. As already noted, $\overline{\boldsymbol{M}}^{(0)} = \frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ can be seen as a restricted version of $\boldsymbol{M}^{(0)}$. $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ is known as a variance-covariance structure from the least squares estimate $\mathbb{V}(\beta) = \hat{\sigma}^2_{LS}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, and is here the fixed prior variance-covariance structure for $\beta \mid \sigma^2$. Confidence in the prior assignment is expressed by the choice of $n^{(0)}$: With $n^{(0)}$ chosen large relative to n, strong confidence in the prior assignment of $\overline{m}^{(0)}$ can be expressed, whereas a low value of $n^{(0)}$ will result in a less pointed prior distribution on $\beta \mid \sigma^2$.

The update step for a canonically constructed prior, expressed in terms of $n^{(0)}$ and $y^{(0)}$, possesses a convenient form: In the prior, the parameters $n^{(0)}$ and $y^{(0)}$ must simply be replaced by their updated versions $n^{(1)}$ and $y^{(1)}$, which calculate as

$$y^{(1)} = \frac{n^{(0)}y^{(0)} + \tau(z)}{n^{(0)}+n}, \qquad\qquad n^{(1)} = n^{(0)}+n.$$

4.1 Update of $\beta \mid \sigma^2$

As $y^{(0)}$ and $y^{(1)}$ are not directly interpretable, it is certainly easier to express prior beliefs on β via the mean vector $\overline{m}^{(0)}$ of the prior distribution of $\beta \mid \sigma^2$ just as in the SCP model. As the transformation $\overline{m}^{(0)} \mapsto y^{(0)}$ is linear, this poses no problem:

$$\begin{aligned}\mathbb{E}[\beta \mid \sigma^2, z] = \overline{m}^{(1)} &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} y_1^{(1)} = n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\left(\frac{n^{(0)}}{n^{(0)}+n} y_1^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{1}{n}(\mathbf{X}^\mathsf{T} z)\right) \\ &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \frac{n^{(0)}}{n^{(0)}+n} \cdot \frac{1}{n}(\mathbf{X}^\mathsf{T}\mathbf{X})\overline{m}^{(0)} + n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \frac{n}{n^{(0)}+n} \cdot \frac{1}{n}(\mathbf{X}^\mathsf{T} z) \\ &= \frac{n^{(0)}}{n^{(0)}+n}\mathbb{E}[\beta \mid \sigma^2] + \frac{n}{n^{(0)}+n}\hat{\beta}_{LS}\,. \end{aligned} \tag{12}$$

The posterior expectation for $\beta \mid \sigma^2$ is here a scalar-weighted mean of the prior expectation and the least squares estimate, with weights $n^{(0)}$ and n, respectively. The role of $n^{(0)}$ in the prior variance of $\beta \mid \sigma^2$ is directly mirrored here. As described for the generalized setting in Walter & Augustin (2009, p. 258) in more detail, $n^{(0)}$ can be seen as a parameter describing the "prior strength" or expressing "pseudocounts". In line with this interpretation, high values of $n^{(0)}$ as compared to n result here in a strong influence of $\overline{m}^{(0)}$ for the calculation of $\overline{m}^{(1)}$, whereas for small values of $n^{(0)}$, $\mathbb{E}[\beta \mid \sigma^2, z]$ will be dominated by the value of $\hat{\beta}_{LS}$.

The variance of $\beta \mid \sigma^2$ is updated as follows:

$$\mathbb{V}(\beta \mid \sigma^2, z) = \frac{n\sigma^2}{n^{(1)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \frac{n\sigma^2}{n^{(0)}+n}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,.$$

Here, $n^{(0)}$ is updated to $n^{(1)}$, and thus the posterior variances are automatically smaller than the prior variances, just as in the SCP model.

4.2 Update of σ^2

For the assignment of the parameters $\overline{a}^{(0)}$ and $\overline{b}^{(0)}$ to define the inverse gamma part of the joint prior, only $y_2^{(0)}$ is left to choose, as $n^{(0)}$ and $y_1^{(0)}$ are already assigned via the choice of $\overline{m}^{(0)}$ and $\overline{\boldsymbol{M}}^{(0)}$. To choose $y_2^{(0)}$, it is convenient to consider the prior expectation of σ^2 (alternatively, the prior mode of σ^2 could be considered as well):

$$\mathbb{E}[\sigma^2] = \frac{\overline{b}^{(0)}}{\overline{a}^{(0)}-1} = \frac{2n^{(0)}}{n^{(0)}+p} y_2^{(0)} - \frac{1}{n^{(0)}+p}\overline{m}^{(0)\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)}\,.$$

A value of $y_2^{(0)}$ dependent on the value of $\mathbb{E}[\sigma^2]$ can thus be chosen by the linear mapping

$$y_2^{(0)} = \frac{n^{(0)}+p}{2n^{(0)}}\mathbb{E}[\sigma^2] + \frac{1}{2n^{(0)}}\overline{m}^{(0)\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)}\,.$$

For the posterior expected value of σ^2, there is a similar decomposition as for the SCP model, and furthermore two other possible decompositions offering interesting interpretations of the update step of σ^2. The three decompositions are presented in the following.

4.2.1 Decomposition Including an Estimate of σ^2 Through the Null Model

The posterior variance of σ^2 calculates firstly as:

$$\begin{aligned}
\mathbb{E}[\sigma^2 \mid z] &= \frac{\overline{b}^{(1)}}{\overline{a}^{(1)} - 1} = \frac{2n^{(1)}}{n^{(1)} + p} y_2^{(1)} - \frac{1}{n^{(1)} + p} \overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)} \\
&= \frac{2n^{(0)}}{n^{(0)} + n + p} y_2^{(0)} + \frac{1}{n^{(0)} + n + p} z^{\mathsf{T}} z - \frac{1}{n^{(0)} + n + p} \overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)} \\
&= \frac{n^{(0)} + p}{n^{(0)} + n + p} \mathbb{E}[\sigma^2] + \frac{n - 1}{n^{(0)} + n + p} \frac{1}{n - 1} z^{\mathsf{T}} z \\
&\quad + \frac{1}{n^{(0)} + n + p} \left(\overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \overline{m}^{(0)} - \overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)} \right), \qquad (13)
\end{aligned}$$

and so displays as a weighted average of the prior expected value, $\frac{1}{n-1} z^{\mathsf{T}} z$, and a term depending on prior and posterior estimates for β, with weights $n^{(0)} + p$, $n - 1$ and 1, respectively. When adopting the centered z, standardized $\mathbf{X}$ approach, $\frac{1}{n-1} z^{\mathsf{T}} z$ is the estimate for σ^2 under the null model, that is, if $\beta = 0$. Contrary to what a cursory inspection might suggest, the third term's influence, having the constant weight of 1, will not vanish for $n \to \infty$, as the third term does not approach a constant.[5]

The third term reflects the change in information about β:

If we are very uncertain about the prior beliefs on β expressed in $\overline{m}^{(0)}$ and thus assign a small value for $n^{(0)}$ with respect to n, we will get relatively large variances and covariances in $\overline{\boldsymbol{M}}^{(0)}$ by a factor $\frac{n}{n^{(0)}} > 1$ to $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, resulting in a small term $\overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \overline{m}^{(0)}$. After updating, the elements in $\overline{\boldsymbol{M}}^{(1)}$ become smaller automatically due to the updated factor $\frac{n}{n^{(0)}+n}$ to $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$. If the values of $\overline{m}^{(1)}$ do not differ much from the values in $\overline{m}^{(0)}$, the term $\overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)}$ would be larger than its prior counterpart, ultimately reducing the posterior expectation for σ^2 through the third term being negative. If $\overline{m}^{(1)}$ does significantly differ from $\overline{m}^{(0)}$, then the term $\overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)}$ can actually result smaller than its prior counterpart and thus give a larger value of $\mathbb{E}[\sigma^2 \mid z]$ as compared with the situation $\overline{m}^{(1)} \approx \overline{m}^{(0)}$.

On the contrary, large values for $n^{(0)}$ with respect to n indicating high trust in prior beliefs on β lead to small variances and covariances in $\overline{\boldsymbol{M}}^{(0)}$ by the factor $\frac{n}{n^{(0)}} < 1$

[5] Although $\overline{m}^{(1)}$ approaches $\hat{\beta}_{LS}$, and $\overline{m}^{(0)}$ is a constant, $\overline{\boldsymbol{M}}^{(0)^{-1}}$ and $\overline{\boldsymbol{M}}^{(1)^{-1}}$ are increasing for growing n, with $\overline{\boldsymbol{M}}^{(1)^{-1}}$ increasing faster than $\overline{\boldsymbol{M}}^{(0)^{-1}}$. The third term will thus eventually turn negative, reducing the null model variance that has weight $n - 1$.

to $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, resulting in a larger term $\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)}$ as compared to the case with low $n^{(0)}$. After updating, variances and covariances in $\overline{\boldsymbol{M}}^{(1)}$ will become even smaller, amplifying the term $\overline{m}^{(1)^\mathsf{T}}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)}$ even more if $\overline{m}^{(1)} \approx \overline{m}^{(0)}$, ultimately reducing the posterior expectation for σ^2 more than in the situation with low $n^{(0)}$. If, however, the values of $\overline{m}^{(1)}$ do differ significantly from the values in $\overline{m}^{(0)}$, the term $\overline{m}^{(1)^\mathsf{T}}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)}$ can result smaller than its prior counterpart also here and even more so as compared to the situation with low $n^{(0)}$, giving eventually an even larger posterior expectation for σ^2.

4.2.2 Decomposition Similar to the SCP Model

A decomposition similar to the one in Section 3.2 can be derived by considering the third term from (13) in more detail:

$$
\begin{aligned}
&\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)} - \overline{m}^{(1)^\mathsf{T}}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)} \\
&\quad = n^{(0)} \cdot n \cdot y_1^{(0)^\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}y_1^{(0)} - (n^{(0)}+n)\cdot n\frac{n^{(0)}y^{(0)^\mathsf{T}} + z^\mathsf{T}\mathbf{X}}{n^{(0)}+n}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\frac{n^{(0)}y^{(0)} + \mathbf{X}^\mathsf{T}z}{n^{(0)}+n} \\
&\quad = \frac{n}{n^{(0)}+n}\left[\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)} - 2\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\hat{\beta}_{LS} - \frac{n}{n^{(0)}}\hat{\beta}_{LS}^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}\hat{\beta}_{LS}\right] \\
&\quad = \frac{n}{n^{(0)}+n}\left(\overline{m}^{(0)} - \hat{\beta}_{LS}\right)^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}\left(\overline{m}^{(0)} - \hat{\beta}_{LS}\right) - z^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}z.
\end{aligned}
$$

Thus, we get

$$
\begin{aligned}
\mathbb{E}[\sigma^2 \mid \boldsymbol{z}] &= \frac{n^{(0)}+p}{n^{(0)}+n+p}\mathbb{E}[\sigma^2] + \frac{1}{n^{(0)}+n+p}\left(z^\mathsf{T}z - z^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}z\right) \\
&\quad + \frac{1}{n^{(0)}+n+p}\cdot\frac{n}{n^{(0)}+n}(\overline{m}^{(0)} - \hat{\beta}_{LS})^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}(\overline{m}^{(0)} - \hat{\beta}_{LS}) \\
&= \frac{n^{(0)}+p}{n^{(0)}+n+p}\mathbb{E}[\sigma^2] + \frac{n-p}{n^{(0)}+n+p}\cdot\underbrace{\frac{1}{n-p}(z - \mathbf{X}\hat{\beta}_{LS})^\mathsf{T}(z - \mathbf{X}\hat{\beta}_{LS})}_{\hat{\sigma}_{LS}^2} \\
&\quad + \frac{p}{n^{(0)}+n+p}\cdot\underbrace{\frac{n}{n^{(0)}+n}\frac{1}{p}(\overline{m}^{(0)} - \hat{\beta}_{LS})^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}(\overline{m}^{(0)} - \hat{\beta}_{LS})}_{=:\overline{\sigma}_{PDC}^2}. \qquad (14)
\end{aligned}
$$

The posterior expectation for σ^2 can therefore be seen also here as a weighted average of the prior expected value, the estimation $\hat{\sigma}_{LS}^2$ resulting from least squares methods, and $\overline{\sigma}_{PDC}^2$,[6] with weights $n^{(0)}+p$, $n-p$ and p, respectively. As in the update step

[6] $\mathbb{E}[\overline{\sigma}_{PDC}^2 \mid \sigma^2] = \sigma^2$ computes very similar to the calculations given in O'Hagan (1994, p. 249).

for $\beta \mid \sigma^2$, $n^{(0)}$ is guarding the influence of the prior expectation on the posterior expectation. Just as in the decomposition for the SCP model, the weight for $\hat{\sigma}^2_{LS}$ will dominate the others when the sample size approaches infinity. Also for the CCCP model, $\overline{\sigma}^2_{PDC}$ is getting large if prior beliefs on β are skewed with respect to "what the data says", eventually inflating the posterior expectation of σ^2. The weighting of the differences is similar as well: High prior confidence in the chosen value of $\overline{m}^{(0)}$ as expressed by a high value of $n^{(0)}$ will give a large $\overline{\boldsymbol{M}}^{(0)^{-1}}$ and thus penalizing erroneous assignments stronger as compared to a lower value of $n^{(0)}$. Again, $\mathbf{X}^\intercal\mathbf{X}$ weighs the differences for components with covariates having a low spread weaker due to the instability of the respective component of $\hat{\beta}_{LS}$ under such conditions.

4.2.3 Decomposition with Estimates of σ^2 Through Prior and Posterior Residuals

A third interpretation of $\mathbb{E}[\sigma^2 \mid z]$ can be derived by another reformulation of the third term in (13):

$$\begin{aligned}
&\overline{m}^{(0)\intercal}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)} - \overline{m}^{(1)\intercal}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)} = \frac{n^{(0)}}{n}\overline{m}^{(0)\intercal}\mathbf{X}^\intercal\mathbf{X}\overline{m}^{(0)} - \frac{n^{(1)}}{n}\overline{m}^{(1)\intercal}\mathbf{X}^\intercal\mathbf{X}\overline{m}^{(1)} \\
&= \frac{n^{(0)}}{n}(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)}) - \frac{n^{(1)}}{n}(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)}) \\
&\quad + \frac{n^{(1)}}{n}z^\intercal z - \frac{n^{(0)}}{n}z^\intercal z + \frac{n^{(0)}}{n}2z^\intercal\mathbf{X}\overline{m}^{(0)} - \frac{n^{(1)}}{n}2z^\intercal\mathbf{X}\overline{m}^{(1)} \\
&= \frac{n^{(0)}}{n}(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)}) - \frac{n^{(1)}}{n}(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)}) + z^\intercal z - 2z^\intercal\mathbf{X}\hat{\beta}_{LS}\,.
\end{aligned}$$

With this, we get

$$\begin{aligned}
\mathbb{E}[\sigma^2 \mid z] = \frac{n^{(0)}+p}{n^{(0)}+n+p}\mathbb{E}[\sigma^2] + \frac{n^{(0)}+p}{n^{(0)}+n+p}\underbrace{\frac{n^{(0)}}{n}\cdot\frac{1}{n^{(0)}+p}(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)})}_{=:\sigma^{(0)^2},\ \text{as}\ \mathbb{E}[\sigma^{(0)^2}\mid\sigma^2]=\sigma^2} \\
+ \frac{2(n-p)}{n^{(0)}+n+p}\hat{\sigma}^2_{LS} - \frac{n^{(1)}+p}{n^{(0)}+n+p}\underbrace{\frac{n^{(1)}}{n}\cdot\frac{1}{n^{(1)}+p}(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)})}_{=:\sigma^{(1)^2},\ \text{as}\ \mathbb{E}[\sigma^{(1)^2}\mid\sigma^2,z]=\mathbb{E}[\sigma^{(1)^2}\mid\sigma^2]=\sigma^2}\,.
\end{aligned} \tag{15}$$

Here, the calculation of $\mathbb{E}[\sigma^2 \mid z]$ is based again on $\mathbb{E}[\sigma^2]$ and $\hat{\sigma}^2_{LS}$, but nowe complemented with two special estimates: $\sigma^{(0)^2}$, an estimate based on the prior residuals $(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)})$, and a respective posterior version $\sigma^{(1)^2}$, based on $(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)})$. However, $\mathbb{E}[\sigma^2 \mid z]$ is only "almost" a weighted average of these ingredients, as the weights sum up to $n^{(0)}-p+n$ instead of $n^{(0)}+p+n$.

Especially strange is the negative weight for ${\sigma^{(1)}}^2$, actually making the factor to ${\sigma^{(1)}}^2$ result to -1. A possible interpretation would be to group $\mathbb{E}[\sigma^2]$ and ${\sigma^{(0)}}^2$ as prior-based estimations with joint weight $2(n^{(0)}+p)$, and $\hat{\sigma}^2_{LS}$ as data-based estimation with weight $2(n-p)$. Together, these estimations have a weight of $2(n^{(0)}+n)$, being almost (neglecting the missing $2p$) a "double estimate" that is corrected back to a "single" estimate with the posterior-based estimate ${\sigma^{(1)}}^2$.

4.3 Update of β

As for the SCP model, the posterior on β, being the most important distribution for inference, is a multivariate t with expectation $\overline{m}^{(1)}$ as described in Section 4.1. For $\mathbb{V}(\beta \mid z)$, one gets different formulations depending on the formula for $\mathbb{E}[\sigma^2 \mid z]$:

$$
\begin{aligned}
\mathbb{V}(\beta \mid z) &= \frac{\overline{b}^{(1)}}{\overline{a}^{(1)}-1}\overline{\boldsymbol{M}}^{(1)} = \mathbb{E}[\sigma^2 \mid z]\,\frac{n}{n^{(1)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \qquad (16)\\
&\overset{(13)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{\mathbb{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\beta)} + \frac{n-1}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\frac{1}{n-1}z^{\mathsf{T}}z(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\\
&\quad + \frac{1}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\left(\overline{m}^{(0)\mathsf{T}}{\overline{\boldsymbol{M}}^{(0)}}^{-1}\overline{m}^{(0)} - \overline{m}^{(1)\mathsf{T}}{\overline{\boldsymbol{M}}^{(1)}}^{-1}\overline{m}^{(1)}\right)(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\\
&\overset{(14)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{\mathbb{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\beta)} + \frac{n-p}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\underbrace{\hat{\sigma}^2_{LS}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\hat{\beta}_{LS})}\\
&\quad + \frac{p}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\overline{\sigma}^2_{PDC}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\\
&\overset{(15)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{\mathbb{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\beta)} + \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{{\sigma^{(0)}}^2\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{=:\mathbb{V}^{(0)}(\beta)}\\
&\quad + \frac{2(n-p)}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\underbrace{\hat{\sigma}^2_{LS}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\hat{\beta}_{LS})} - \frac{n^{(1)}+p}{n^{(0)}+n+p}\underbrace{{\sigma^{(1)}}^2\frac{n}{n^{(1)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{=:\mathbb{V}^{(1)}(\beta)}.
\end{aligned}
$$

In these equations, it is possible to isolate $\mathbb{V}(\beta)$, $\mathbb{V}(\hat{\beta}_{LS})$ and, in the formulation with (15), the newly defined $\mathbb{V}^{(0)}(\beta)$ and $\mathbb{V}^{(1)}(\beta)$. However, all three versions do not constitute a weighted average, even when the formula for $\mathbb{E}[\sigma^2 \mid z]$ did have this property. Just as in the SCP model, $\mathbb{V}(\beta \mid z)$ can increase if the automatic abatement of the elements in $\overline{\boldsymbol{M}}^{(1)}$ is overcompensated by a strong increase of $\mathbb{E}[\sigma^2]$. Again, this reaction to prior-data conflict is unspecific because it depends on $\mathbb{E}[\sigma^2 \mid z]$ alone.

5 Discussion and Outlook

For both the SCP and CCCP model, $\mathbb{E}[\beta \mid z]$ results as a weighted average of $\mathbb{E}[\beta]$ and $\hat{\beta}_{LS}$, such that the posterior distribution on β will be centered around a mean somewhere between $\mathbb{E}[\beta]$ and $\hat{\beta}_{LS}$, with the location depending on the respective weights. The weights for the CCCP model appear especially intuitive: $\hat{\beta}_{LS}$ is weighted with the sample size n, whereas $\mathbb{E}[\beta]$ has the weight $n^{(0)}$ reflecting the "prior strength" or "pseudocounts". Due to this, prior-data conflict may at most affect the variances only. Indeed, for both prior models, $\mathbb{E}[\sigma^2 \mid z]$ can increase in the presence of prior-data conflict, as shown by the decompositions in Sections 3.2 and 4.2. Through the formulations (6) and (16) for $\mathbb{V}(\beta \mid z)$, respectively, it can be seen that the posterior distribution on β can in fact become less pointed than the prior when prior-data conflict is at hand. Nevertheless, the effect might be not be as strong as desired: In the formulations (5) and (14), respectively, the effect is based only on one term of the decomposition, and furthermore may be foiled through the automatic decrease of $\boldsymbol{M}^{(1)}$ and $\overline{\boldsymbol{M}}^{(1)}$.

Probably the most problematic finding is that this (possibly weak) reaction affects the whole variance-covariance matrix uniformally, and thus, in both models, the reaction to prior-data conflict is by no means component-specific.

Therefore, the prior models lack the capability to mirror the appropriateness of the prior assignments for each covariate separately. As the SCP model is already the most general approach in the class of conjugate priors, this non-specificity feature seems inevitable in Bayesian linear regression based on precise conjugate priors.

In fact, as argued in Section 1, a more sophisticated and specific reaction to prior-data conflict is only possible by extending considerations beyond the traditional concept of probability. Imprecise probabilities, as a general methodology to cope with the multidimensional nature of uncertainty, appears promising here. For generalized Bayesian approaches, the possibility to mirror the quality of prior knowledge is one of the main reasons for the paradigmatic skip from classical probability to interval / imprecise probability. In this framework ambiguity in the prior specification can be modeled by considering sets $\mathscr{M}_\vartheta$ of prior distributions. In the most common approach based on Walley's Generalized Bayes Rule (Walley 1991), posterior inference is then based on a set of posterior distributions $\mathscr{M}_{\vartheta|z}$, resulting from updating the distributions in the prior set element by element.

Of particular computational convenience are again models based on conjugate priors, as developed for the Dirichlet-Multinomial model by Walley (1996), see also Bernard (2009), and for i.i.d. exponential family sampling models by Quaeghebeur & de Cooman (2005), which were extended by Walter & Augustin (2009) to allow an elegant handling of prior-data conflict: With the magnitude of the set $\mathscr{M}_{\vartheta|z}$ mapping the posterior ambiguity, high prior-data conflict leads, ceteris paribus, to a large $\mathscr{M}_{\vartheta|z}$, resulting in high imprecision in the posterior probabilities, and cautious inferences based on it, while in the case of no prior-data conflict $\mathscr{M}_{\vartheta|x}$, and thus the imprecision, is much smaller.

The essential technical ingredient to derive this class of models is the general construction principle also underlying the CCCP model from Section 4, and thus that model can be extended directly to a powerful corresponding imprecise probability model.[7] A detailed development is beyond the scope of this contribution.

Acknowledgements We are very grateful to Erik Quaeghebeur and Frank Coolen for intensive discussions on foundations of generalized Bayesian inference, and to Thomas Kneib for help at several stages of writing this paper.

References

Augustin, T., Coolen, F. P., Moral, S. & Troffaes, M. C. (eds) (2009). *ISIPTA'09: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, Durham University, Durham, UK, July 2009*, SIPTA.

Augustin, T. & Hable, R. (2009). On the impact of robust statistics on imprecise probability models: a review, *ICOSSAR'09: The 10th International Conference on Structural Safety and Reliability, Osaka*. To appear.

Bernard, J.-M. (2009). Special issue on the Imprecise Dirichlet Model. *International Journal of Approximate Reasoning*.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*, Wiley, Chichester.

Bousquet, N. (2008). Diagnostic of prior-data agreement in applied bayesian analysis, **35**: 1011–1029.

Coolen-Schrijner, P., Coolen, F., Troffaes, M. & Augustin, T. (2009). Special Issue on Statistical Theory and Practice with Imprecision, *Journal of Statistical Theory and Practice* **3**.

de Cooman, G., Vejnarová, J. & Zaffalon, M. (eds) (2007). *ISIPTA'07: Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications, Charles University, Prague, Czech Republic, July 2007*, SIPTA.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms, *Q. J. Econ.* pp. 643–669.

Evans, M. & Moshonov, H. (2006). Checking for prior-data conflict, *Bayesian Analysis* **1**: 893–914.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum-likelihood estimator in generalized linear-models, *Annals of Statistics* **13**: 342–368.

Fahrmeir, L. & Kneib, T. (2006). Structured additive regression for categorial space-time data: A mixed model approach, *Biometrics* **62**: 109–118.

Fahrmeir, L. & Kneib, T. (2009). Propriety of posteriors in structured additive regression models: Theory and empirical evidence, *Journal of Statistical Planning and Inference* **139**: 843–859.

Fahrmeir, L., Kneib, T. & Lang, S. (2007). *Regression. Modelle, Methoden und Anwendungen*, Springer, New York.

Fahrmeir, L. & Raach, A. (2007). A Bayesian semiparametric latent variable model für mixed responses, *Psychometrika* **72**: 327–346.

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

Higgins, J. P. T. & Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis, *Statistics in Medicine* **15**: 2733–2749.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making, *Science* **310**: 1680–1683.

Huber, P. J. & Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities, *The Annals of Statistics* **1**: 251–263.

[7] For σ^2 fixed, the model from Section 3 can be comprised under a more general structure that also can be extended to imprecise probabilities, see Walter et al. (2007) and Walter (2006) for details.

Kauermann, R., Krivobokova, T. & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smooting, *J. Roy. Statist. Soc. Ser. B* **71**: 487–503.

Klir, G. J. & Wierman, M. J. (1999). *Uncertainty-based Information. Elements of Generalized Information Theory*, Physika, Heidelberg.

Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression for interval-censored survival times, **34**: 207–228.

Kyburg, H. (1987). Logic of statistical reasoning, *in* S. Kotz, N. L. Johnson & C. B. Read (eds), *Encyclopedia of Statistical Sciences*, Vol. 5, Wiley-Interscience, New York, pp. 117–122.

O'Hagan, A. (1994). *Bayesian Inference, Vol. 2B of Kendall´s Advanced Theory of Statistics*, Arnold, London.

Quaeghebeur, E. & de Cooman, G. (2005). Imprecise probability models for inference in exponential families, *in* F. G. Cozman, R. Nau & T. Seidenfeld (eds), *ISIPTA '05: Proc. 4th Int. Symp. on Imprecise Probabilities and Their Applications*, pp. 287–296.

Ríos Insua, D. & Ruggeri, F. (eds) (2000). *Robust Bayesian Analysis*, Springer, New York.

Scheipl, F. & Kneib, T. (2009). Locally adaptive Bayesian P-splines with a normal-exponential-gamma prior, *Computational Statistics & Data Analysis* **53**: 3533–3552.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles, *Journal of the Royal Statistical Society. Series B. Methodological* **58**: 3–57.

Walter, G. (2006). *Robuste Bayes-Regression mit Mengen von Prioris — Ein Beitrag zur Statistik unter komplexer Unsicherheit*, Master's thesis, Department of Statistics, LMU Munich. Diploma thesis. `http://www.stat.uni-muenchen.de/~walter`.

Walter, G. & Augustin, T. (2009). Imprecision and prior-data conflict in generalized Bayesian inference., *Journal of Statistical Theory and Practice* **3**: 255–271.

Walter, G., Augustin, T. & Peters, A. (2007). Linear regression analysis under sets of conjugate priors, *in* G. de Cooman, J. Vejnarová & M. Zaffalon (eds), *ISIPTA '07: Proc. 5th Int. Symp. on Imprecise Probabilities and Their Applications*, pp. 445–455.

Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*, Physika, Heidelberg.