A credibility approach for combining likelihoods of generalized linear models

A CREDIBILITY APPROACH FOR COMBINING LIKELIHOODS OF GENERALIZED LINEAR MODELS

MARCUS CHRISTIANSEN & EDO SCHINZINGER

Abstract. Generalized linear models (GLM) are a popular tool for the modelling of insurance claims data. Problems arise with the model fitting if little statistical information is available. In case that related statistics are available, statistical inference can be improved with the help of the borrowing-strength principle. We present a credibility approach that combines the maximum likelihood estimators of individual canonical GLMs in a meta-analytic way to an improved credibility estimator. We follow the concept of linear empirical Bayes estimation, which reduces the necessary parametric assumptions to a minimum. The concept is illustrated by a simulation study and an application example from mortality modeling.

Keywords: linear Bayes estimator; canonical generalized linear model; meta-analysis; pseudo likelihood estimator; multi-population mortality modelling

1. Introduction

Greatest accuracy credibility theory is a set of prediction techniques for random effects models that use the borrowing-strength principle, i.e. individual predictions are improved by using also information on the group that the individual belongs to. It started with Whitney (1918), who investigated the insurance problem of how to estimate individual expected claims from both, individual claims experience and portfolio claims data. Whitney proposed that the predictor be a linear combination of individual experience and group average, and this linearity assumption is fundamental to credibility theory since then. Many papers following afterwards focussed on models with parametric distribution assumptions, until Bühlmann (1967) emphasized that the linearity assumption combined with least square estimation allows for distribution free credibility formulas. That non-parametric notion of Bühlmann was a big step forward for greatest accuracy credibility theory as in insurance applications detailed knowledge on appropriate parametrisations is often missing. Following Bühlmann's perspective on credibility theory, which is also called linear (empirical) Bayes estimation, we use a linear representation assumption for the credibility estimator and least squares estimation in order to develop a credibility approach for the statistical analysis of groups of random effects GLMs. In particular, we avoid parametric distribution assumptions for the random effects, just using moments of first and second order. In the credibility theory literature, only the contribution by De Vylder (1985) combines non-linear regression and non-parametric random effects modelling. Unlike De Vylder, we take the perspective of a meta-analysis and base our credibility estimator on the conditional maximum likelihood estimators of the individual models rather than (a transformation of) the full observed sample. We think that our estimator is more intuitive, and it has advantages when access to the full original data is costly or restricted.

Let us consider the following example: For a group of different populations $i = 1, \dots, n$ a survival model shall be fitted to empirically observed mortality data. Most of the survival models in the actuarial literature can be written in the form of a generalized linear model (GLM), so we

posit here that

$$g(\mathbb{E}[Y_i]) = X_i\beta_i, \quad i = 1, \dots, n, \tag{1.1}$$

where the random vector Y_i is the observed central death rate for each considered age in population i. The link function g and the matrix X_i characterize the family of surival models, and the fixed effect vector β_i is the parameter vector that shall be estimated for population i. Let $\hat{\beta}_i$ be the classical maximum likelihood estimator (MLE) for β_i based only on the individual data. The borrowing-strengths principle suggests that $\hat{\beta}_i$ can be improved by involving also the data of the other populations, and finding such an improved estimator is exactly the aim of this paper.

In order to get the individual models (1.1) into a joint framework, the fundamental (Bayesian) idea of credibility theory is to replace unknown fixed effects by random effects (the credibility parameter), cf. chapter 1.2.4 in Bühlmann and Gisler (2005). In our example the deterministic vectors β_i are replaced by independent and identically distributed random vectors B_i. The redefined and combined models form then a multi-population model,

$$\begin{pmatrix} g(\mathbb{E}[Y_1 \mid B_1]) \\ \vdots \\ g(\mathbb{E}[Y_n \mid B_n]) \end{pmatrix} = \begin{pmatrix} X_1B_1 \\ \vdots \\ X_nB_n \end{pmatrix}. \tag{1.2}$$

By adding parametric assumptions for the distribution of B_i, we could interpret this multi-population model as a generalized linear mixed model (GLMM), a hierarchical generalized linear model (HGLM), or a hierarchical Bayesian model. The statistical literature offers a multitude of well established techniques for GLMMs, HGLMs, and hierarchical Bayesian frameworks. However, in our example we know hardly anything about the distribution of B_i, so that specific parametric assumptions should be avoided. This paper shows how parametric assumptions for B_i can indeed be avoided by using concepts from credibility theory.

After redefining the unknown individual fixed effects as i.i.d. random effects, the second fundamental idea of greatest accuracy credibility theory in terms of Bühlmann is to choose the estimator $\hat{B}_i$ of B_i as the best linear unbiased estimator in a mean squared error sense, cf. chapter 3 in Bühlmann and Gisler (2005). Without the linearity assumption, the best estimator would always be the conditional expectation of B_i given the sample, also called the Bayes estimator. Adding the linearity assumption has two useful consequences: First, the resulting predictor has always an easy and intuitive interpretation, and second, estimations can be solely based on first and second order moments. Moreover, the linearity assumption tremendously simplifies computations. In statistical inference theory similar concepts are used, there known as linear Bayes estimation, see Norberg (1980) and the overview paper Norberg (2004) for further references.

The linearity assumptions in the credibility theory literature are not throughout consistent. Most authors assume linearity with respect to the sample. In our example that would mean that

$$\hat{B}_i \in \operatorname{span}\{1, Y_1, \dots, Y_n\}.$$

The non-linear regression credibility estimator as suggested by De Vylder (1985) is of the form

$$\hat{B}_i \in \operatorname{span}\{1, h_1(Y_1), \dots, h_n(Y_n)\},$$

where the non-linear functions h_i are inverses of the mappings $b \mapsto g^{-1}(X_ib)$ in order to retrieve linearity with respect to the sample. De Vylder (1985) alternatively suggests to replace the non-linear regression problem by an approximating linear regression problem. In our example that would mean that we replace g in (1.2) by a linear approximation and then apply credibility

theory for linear regression. This paper uses neither of De Vylder's approaches but assumes that

$$\hat{B}_i \in \operatorname{span}\{1, \hat{\beta}_1, \dots, \hat{\beta}_n\}. \tag{1.3}$$

The MLEs $\hat{\beta}_1, \dots, \hat{\beta}_n$ can have irregularities, so we will replace them by asymptotically equivalent estimators, but for the moment we can simply think of the classical MLEs.

In order to motivate assumption (1.3), we have to go back to the original aim of credibility theory: The usual quantity of interest in credibility theory is the expected claims of an individual risk i. The credibility estimator for this quantity of interest is typically linear in two ways:

(a) it is a linear combination of the observed sample,
(b) it is a linear combination of the predictors that we would obtain from individual fixed effect models.

In our opinion it is rather the second property that makes credibility formulas so appealing, because it allows the perspective of a meta-analysis: separate investigations for each risk lead to individual empirical estimators, and the credibility formula is a meta step that linearly combines the results of the separate investigations in order to achieve a higher statistical power. In our mortality example, the two perspectives translate to

(a) $\hat{B}_i$ is a linear combination of the observed sample $Y_1, \dots, Y_n$,
(b) $\hat{B}_i$ is a linear combination of the predictors $\hat{\beta}_1, \dots, \hat{\beta}_n$ that we obtain from the individual fixed effect models (1.1).

In contrast to most classical credibility models, in our non-linear regression case the two perspectives (a) and (b) are not equivalent anymore but are mutually exclusive. Which kind of linearity should be maintained? Theoretically, from a mean squared error perspective it is optimal to let the credibility estimator be linear in some sufficient statistic of B_i, cf. Taylor (1977). The present paper takes a rather practical perspective and chooses (b) for two reasons:

- The individual predictors $\hat{\beta}_1, \dots, \hat{\beta}_n$ relate stronger to the quantity of interest B_i than the sample $Y_1, \dots, Y_n$ (or the transformed sample of De Vylder (1985)), yielding a more intuitive interpretation of the linear credibility formula.
- The meta analysis structure is a big advantage whenever the access to the original sample is restricted or costly.

Credibility estimators are also called linear Bayes estimators, and the question on what statistics to base a linear Bayes estimator is generally discussed in Neuhaus (1985) from the perspective of efficiency. We focus here on interpretability and data availability rather than efficiency. Our concept of linearly combining individual maximum likelihood estimators to an improved estimator has similarities with Efron (1996), where individual likelihoods are combined to define empirical Bayes estimators. Contrary to Efron (1996), our credibility estimators are *linear* empirical Bayes estimators, which allows us to to further weaken the a priori assumptions. While we directly combine the maximum likelihood estimators of the credibility parameters, Efron assumes a parametric distribution for the credibility parameters and works with the maximum likelihood estimators for the hyperparameter.

Credibility estimators for *linear* regression models were first introduced by Hachemeister (1975), followed by many further contributions in this field. Taylor (1977) extended Hachemeister's linear regression concept to general Hilbert spaces. Norberg (1980) showed the general link between credibility theory and linear Bayes estimation by discussing linear regression examples. Lo et al. (2007) explored the approach of generalized estimating equations (GEE) for linear regression credibility models. For further references see the overview paper by Brazauskas et al.

(2014), who emphasize that many common linear regression credibility models can be represented as Linear Mixed Models. A first attempt to extend credibility theory to general *non-linear* regression has been made by De Vylder (1985), who proposes two different approaches. The first approach bases on a transform of the observed sample which reverses the non-linearity. The resulting credibility estimator is linear in the transformed sample (instead of the original sample), see the comments above. The second approach linearises – and thereby approximates – the non-linear regression problem itself. Pitselis (2004) showed how to make De Vylder's credibility estimators more robust. Ohlsson and Johansson (2006) and Ohlsson (2008) study credibility estimation for non-linear regression within the parametric framework of Tweedie-GLMs. The random effects are assumed to have natural conjugate distributions and have the form of multiplicative factors that equally affect all observations and GLM-coefficients. Generally, we can interpret *fully parametric* regression credibility models as Hierarchical Generalized Linear Models, c.f. Nelder and Verrall (1997), or Generalized Linear Mixed Models, c.f. Antonio and Beirlant (2007), for which the statistical literature provides a multitude of inference methods. Differing from theses concepts, the present paper assumes only that the conditional individual models are parametric (namely GLMs) but refrains from making parametric assumptions for the credibility parameters. This takes account of the fact that in many applications there is hardly any information available on the distributional properties of the credibility parameters. *Nonparametric* regression estimation appears in the context of credibility theory in Qian (2000), who derives credibility premiums with the help of Kernel estimators. The focus is on the estimation of expectations, variances, and quantiles of insurance claims. A broader study of nonparametric credibility regression models is still missing.

This paper is organized as follows. Section 2 recalls the linear regression credibility model of Hachemeister (1975) and develops a fundamental framework for credibility GLMs. In Section 3 the classical maximum likelihood estimators for GLMs are embedded in the extended credibility GLM framework, providing a mathematical rigorous basis for the following sections. Section 4 calculates the theoretical credibility estimator and studies its asymptotic properties. In Section 4 empirical estimators for the hyperparameters in the theoretical credibility estimator are discussed, leading to the empirical credibility estimator. Section 6 studies the performance of the credibility estimator on a theoretical basis and by a simulation example. Section 7 demonstrates the application of the credibility estimator in mortality modeling. Section 8 concludes. Section 9 provides proofs of results from the previous sections.

2. The credibility regression model

Throughout the paper, we consider a portfolio of N clusters with observation vectors $Y_i = (Y_{i1}, \dots, Y_{in})$ and random effects B_i, $i = 1, \dots, N$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The linear regression credibility model. In the linear regression credibility model, the Y_{ij} are univariate random variables that, conditional on the random effects B_i, satisfy a linear regression equation. For a given design matrix $X_i \in \mathbb{R}^{n,p}$ with full rank p, the precise model assumptions according to Hachemeister (1975) are as follows.

(i) Conditionally, given B_i, the Y_{ij}, $j = 1, \dots, n$, are independent and fulfill

$$\mathbb{E}[Y_{ij} \mid B_i] = X_{ij} B_i, \tag{2.1}$$

where $X_{ij} \in \mathbb{R}^{1,p}$ is the j-th row of X_i. Furthermore, the clusters satisfy

$$\mathrm{Cov}(Y_i \mid B_i) = \Sigma_i(B_i). \tag{2.2}$$

(ii) The pairs $(B_1, Y_1), \ldots, (B_N, Y_N)$ are independent and $B_1, \ldots, B_N$ are independent and identically distributed (iid).

The credibility estimator $\hat{B}_i$ for the random regression parameter B_i is defined as the orthogonal projection of B_i on $L(1, \mathbf{Y})$, where the expectation is taken with respect to the full probability measure $\mathbb{P}$ and

$$L(1, \mathbf{Y}) = \left\{ a + \sum_{i=1}^{N} \sum_{j=1}^{n} A_{ij} Y_{ij} : a \in \mathbb{R}^p, A_{ij} \in \mathbb{R}^{p,p} \right\}. \tag{2.3}$$

Thus, every component of the vector $\hat{B}_i$ is an affine function of all observations. An equivalent formulation can be provided by writing $\hat{B}_i$ as the solution of the optimization problem

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\mathbf{Y})}{\arg\min}\ \mathbb{E}\left[(\tilde{B}_i - B_i)'(\tilde{B}_i - B_i) \right],$$

where B_i' denotes the matrix transposition of B_i. Hachemeister (1975) showed that the credibility estimator has the form

$$\hat{B}_i = A_i \hat{\beta}_i + (I - A_i)\mathbb{E}[B_i] \tag{2.4}$$

with

$$\hat{\beta}_i = (X_i' \Sigma_i(B_i) X_i)^{-1} X_i' \Sigma_i(B_i) Y_i \tag{2.5}$$

and credibility matrix

$$A_i = \mathrm{Cov}(B_i) \left(\mathrm{Cov}(B_i) + \left(X_i' \mathbb{E}[\Sigma_i(B_i)]^{-1} X_i \right)^{-1} \right)^{-1}.$$

As the best individual estimators (2.5) are linear in $\mathbf{Y}$, the credibility estimator also satisfies

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\boldsymbol{\beta})}{\arg\min}\ \mathbb{E}\left[(\tilde{B}_i - B_i)'(\tilde{B}_i - B_i) \right], \tag{2.6}$$

where the estimator is now from the linear space

$$L(1, \boldsymbol{\beta}) = \left\{ a + \sum_{k=1}^{N} A_k \hat{\beta}_k : a \in \mathbb{R}^p, A_k \in \mathbb{R}^{p,p} \right\}. \tag{2.7}$$

That means that in the linear regression case it does not make a difference whether we assume linearity of the credibility estimator with respect to (a) the observed sample or (b) the maximum likelihood estimators from individual fixed effect models. Both variants are equivalent. However, this equivalence property gets lost in the non-linear regression case. As motivated in the introduction of this paper, we aim to keep the second kind of linearity in the non-linear regression case.

The extended regression credibility model. The linear regression model can be naturally extended to canonical generalized linear models as follows.

(A1) Conditional on $B_i = \beta_i$, the components of the vector Y_i are independent and their distributions belong to a simple exponential family with natural parameters $\theta_i = (\theta_{ij})_{j=1}^n \subset \Theta$ and weights $w_i = (w_{ij})_{j=1}^n \subset \mathbb{R}^+$. The conditional joint pdf f_{β_i} takes the form

$$f_{\beta_i}(y) = \prod_{j=1}^{n} c(y_j, w_{ij}) \exp\left(\sum_{j=1}^{n} w_{ij} \left(\theta_{ij} y_j - b(\theta_{ij}) \right) \right), \quad y \in \mathbb{R}^n. \tag{2.8}$$

(A2) The natural parameters are linked to a linear predictor by the identity

$$\theta_{ij} = g(\mathbb{E}[Y_{ij} \mid B_i]) = \xi_{ij} + X_{ij}B_i, \quad \text{a.s.}, \tag{2.9}$$

where g is the canonical (natural) link function and $\xi_i = (\xi_{ij})_{j=1}^n \subset \mathbb{R}$ are offset terms.

(A3) The random effects $B_1, \dots, B_N$ are iid. The pairs $(B_1, Y_1), \dots, (B_N, Y_N)$ are independent but not necessarily identically distributed (as the parameters w_i, ξ_i, X_i may differ).

The distribution of $\mathbf{Y} = (Y_1, \dots, Y_N)$ is not specified until we condition on the outcome $\boldsymbol{\beta} = (\beta_1, \dots, \beta_N)$ of $\mathbf{B} = (B_1, \dots, B_N)$. Then, under the conditional measure $\mathbb{P}_\beta$, assumptions (A1) and (A2) state that the Y_{ij} follow a GLM of a univariate simple exponential family with parameter β_i. The linear predictor describes the first two conditional moments of Y_{ij} through the mean and variance functions

$$\mu_{ij}(B_i) = \mathbb{E}[Y_{ij} \mid B_i] = b'(\xi_{ij} + X_{ij}B_i), \tag{2.10}$$

$$w_{ij}v_{ij}(B_i) = \mathrm{Var}(Y_{ij} \mid B_i) = w_{ij}b''(\xi_{ij} + X_{ij}B_i), \tag{2.11}$$

respectively. Notice that $b' = g^{-1}$. The B_i thus characterize the individual distributions of the clusters, but without conditioning on $\mathbf{B}$ the clusters are homogeneous up to known parameters as declared by assumption (A3). The weights in (2.8) may describe known nuisance parameters λ_{ij} by $w_{ij} = 1/\lambda_{ij}$, e.g. the number of trials in a Binomial model. Also clusters with different sample sizes can be incorporated by choosing binary weights denoting the presence or absence of an observation. As we combine ideas from credibility theory and GLMs, we refer to the model as a credibility-GLM (CGLM). We concentrate on the univariate case since it is of main interest in practice and greatly simplifies the notation. Comments on CGLM for q-variate exponential families will follow later on.

Our aim is to establish credibility estimation for this model and a first step is to find a proper definition of the credibility estimator for B_i. In the linear regression case, the credibility estimator has been defined as the orthogonal projection of B_i on the space $L(1, \mathbf{Y})$, but can be equivalently defined as the orthogonal projection of B_i on the space $L(1, \boldsymbol{\beta})$, cf. (2.6). In the extended model, this equivalence is not true anymore, because the conditional expectations of the Y_{ij} are in general not linear in B_i. Estimators which are linear functions of the observation vector cannot capture the effects of the link g so that choosing $L(1, \mathbf{Y})$ as the admissible class of estimators is too restrictive. We suggest to select the credibility estimator within the class of the best individual solutions instead, and this requires us to follow a semi-parametric approach.

What do the best individual solutions look like? If only data for one specific cluster i is available and if we condition on $B_i = \beta_i$, the natural choice will be the best estimator of a GLM, i.e. the maximum likelihood estimator (MLE) $\hat{\beta}_i$. We therefore define $\hat{\beta}_i$ as the maximizer of the function

$$l_{in}(\beta) = \sum_{j=1}^{n} w_{ij} \left(\theta_{ij}Y_{ij} - b(\theta_{ij})\right), \tag{2.12}$$

which is the true likelihood function for Y_i when conditioned on $B_i = \beta$. Also of great importance are the score functions and Fisher information matrices

$$s_{in}(\beta) = \frac{\partial l_{in}(\beta)}{\partial \beta} = \sum_{j=1}^{n} w_{ij}X'_{ij} \left(Y_{ij} - \mu_{ij}(\beta)\right), \tag{2.13}$$

$$F_{in}(\beta) = \frac{\partial^2 l_{in}(\beta)}{\partial \beta \partial \beta'} = \sum_{j=1}^{n} w_{ij}v_{ij}(\beta)X'_{ij}X_{ij}. \tag{2.14}$$

In particular, the MLE solves $s_{in}(\hat{\beta}_i) = 0$. Special caution is needed as the credibility approach assumes that the GLM parameter is a random effect. The map l_{in} represents the true log-likelihood function under the conditional measure $\mathbb{P}_\beta$ but not under the unconditional measure $\mathbb{P}$. The latter requires integration involving the distribution of the random effect B_i, which we have not specified. The reader should keep in mind that the $\hat{\beta}_i$ are, to be precise, conditional MLEs.

The presence of N similar clusters should improve the estimation of the B_i. The idea of credibility estimation is to compose the best individual solutions into a mixing estimator which benefits from the shrinkage effect. As we will later see, $\hat{\beta}_i$ is already, in a proper sense, a good estimator but one should be worried if all other $\hat{\beta}_l$, $l \neq i$, clearly differed. Following the considerations we have made so far, we select the admissible class of estimators as

$$L(1, \boldsymbol{\beta}) = \left\{ a + \sum_{i=1}^{N} A_i \hat{\beta}_i : a \in \mathbb{R}^p, A_i \in \mathbb{R}^{p,p} \right\}. \tag{2.15}$$

The credibility estimator for B_i is then defined as the orthogonal projection of B_i on $L(1, \boldsymbol{\beta})$, or equivalently, as the minimizer of the quadratic loss function

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\boldsymbol{\beta})}{\arg\min}\ \mathbb{E}\left[(\tilde{B}_i - B_i)'(\tilde{B}_i - B_i) \right],$$

where the expectation is taken with respect to the full probability measure $\mathbb{P}$. However, a necessary condition is that $L(1, \boldsymbol{\beta}) \subset \mathcal{L}^2$, i.e. the $\hat{\beta}_i$ must be square integrable. This is in general not satisfied as the next example reveals.

Example 2.1. We consider the Poisson case with the simple design matrix $X = (1 \dots 1)' \in \mathbb{R}^{n,1}$ giving that, conditional on $B_i = \beta_i$,

$$Y_{ij} \sim \text{Poi}\left(\exp(\beta_i)\right), \quad j = 1, \dots, n.$$

Then,

$$s_{in}(\beta_i) = \sum_{j=1}^{n} (Y_{ij} - \exp(\beta_i))$$

and it follows that

$$\hat{\beta}_i = \log\left(\frac{1}{n} \sum_{j=1}^{n} Y_{ij} \right). \tag{2.16}$$

This expression is not well-defined if $\sum_{j=1}^{n} Y_j = 0$ which occurs with positive probability. Thus, $\hat{\beta}_i \notin \mathcal{L}^2$ and also not with respect to the conditional measure $\mathbb{P}_\beta$.

We have to modify the MLEs in order to ensure square integrability. Structure (2.16) is not only to a counterexample for square integrability of $\hat{\beta}_i$ but also a general problem in maximum likelihood estimation. Indeed, MLEs are meant to be defined on some measurable set contained in the whole sample space, cf. Witting and Nölle (1970) and Fahrmeir and Kaufmann (1983). In many papers of nowadays, this aspect of a MLE is often not mentioned. The absence is justified by the asymptotic existence of the estimator, that is the probability of existence converges to one as sample size n increases. For the particular case of (2.16), one can easily check that

$$\mathbb{P}\left(\sum_{j=1}^{n} Y_{ij} = 0 \right) \overset{n\to\infty}{\longrightarrow} 0$$

and we say that

$$\hat{\beta}_i \mathbb{1}_{\{\sum_{j=1}^n Y_{ij} > 0\}} \tag{2.17}$$

is a MLE on $\{\sum_{j=1}^n Y_{ij} > 0\}$. Such defining sets play a crucial role in credibility estimation and must be constructed properly.

3. Construction of a Pseudo Maximum Likelihood Estimator

To stress the role of the defining set of an MLE and to clearly distinguish from the unrestricted MLE, we introduce an explicit notation for estimators of type (2.17). For some family of measurable sets $(M_{in})_{n\in\mathbb{N}} \subset \mathcal{F}$, we call

$$\tilde{\beta}_{in} := \hat{\beta}_{in} \mathbb{1}_{M_{in}}, \quad n \in \mathbb{N}, \tag{3.1}$$

the pseudo maximum likelihood estimator (PMLE) for β_i. The additional index n now emphasizes the quantities' dependence on the sample size. The aim of this section is to construct M_{in} such that $\tilde{\beta}_{in} \in \mathcal{L}^2$ and further properties that will follow. For that purpose, we will work from now on with the following regularity assumptions.

(R1) The random vectors B_i have a compact and convex support $\mathcal{B}$. Furthermore, B_i has no mass at $0 \in \mathbb{R}^p$ and on the boundary of $\mathcal{B}$, i.e.

$$\mathbb{P}(B_i = 0) = 0, \tag{3.2}$$

$$\mathbb{P}(B_i \in \partial\mathcal{B}) = 0. \tag{3.3}$$

(R2) The link function g is twice continuously differentiable with non-singular Jacobian.

(R3) The admissible set of covariates

$$\{X_j : j \in \mathbb{N}\} \subset \mathbb{R}^p$$

is bounded and all of its elements satisfy $X_j\beta \in \Theta^0$ for all $\beta \in \mathcal{B}$. Here, Θ^0 denotes the interior of Θ.

(R4) $\sum_{j=1}^n X_j' X_j$ has full rank p for sufficiently large n.

(R5) The scaled Fisher information matrix $n^{-1}F_{in}(\beta)$ converges pointwise to a positive definite limit $F_i(\beta)$ for all $\beta \in \mathcal{B}$.

(R6) The weights w_{ij} and offset terms ξ_{ij} are bounded.

Remark. The above assumptions have to ensure that the conditional GLM is pointwise well-defined for all realizations $\beta \in \mathcal{B}$ of B_i. Thus, they naturally coincide with common assumptions on classical GLMs. Only (R5) is a totally new assumption and addresses asymptotic properties of maximum likelihood theory. In fact, it generalizes the linear growth condition of the Fisher information matrix as used by McFadden (1973) and Andersen (1980) to the whole state space $\mathcal{B}$ of B_i. The compactness condition in (R1) seems to contradict the usual assumption of parameter spaces being open sets, but we can without loss of generality enlarge $\mathcal{B}$ to an open set, say $\tilde{\mathcal{B}}$, where $\tilde{\mathcal{B}} \setminus \mathcal{B}$ gets zero weight. The main purpose of (R1) is to make the B_i almost surely bounded. Compactness is required for Lemma 9.1, which follows soon. Condition (3.2) is just for technical reasons, excluding the trivial case where a regression model is redundant.

The idea for the explicit construction of the sets (M_{in}) goes back to Fahrmeir and Kaufmann (1985). For $\delta > 0$, we define a sequence of neighborhoods

$$N_n(\delta, B_i) := \{\beta \in \mathcal{B} : \sqrt{n}\|\beta - B_i\| \leq \delta\}, \quad n \in \mathbb{N}, \tag{3.4}$$

which are spheres with radius $\delta/\sqrt{n}$ and random central point B_i. In addition, let

$$M_{in}^{\delta} := \{l_{in}(\beta) - l_{in}(B_i) < 0, \quad \text{for all } \beta \in \partial N_n(\delta, B_i)\}. \tag{3.5}$$

If the event M_{in}^{δ} occurs, there exists a local maximum in the interior of $N_n(\delta, B_i)$. Since the log-likelihood function l_{in} is concave, the local maximum is also a unique global maximum which is attained by $\hat{\beta}_{in}$. Therefore, $\omega \in M_{in}^{\delta}$ implies that $\hat{\beta}_{in}(\omega) \in N_n(\delta, B_i(\omega))$, i.e.

$$\mathbb{1}_{M_{in}^{\delta}} \|\hat{\beta}_{in} - B_i\| \leq \frac{\delta}{\sqrt{n}}, \quad \text{a.s.} \tag{3.6}$$

The PMLE will be constructed along these sets with an appropriate choice for δ.

Theorem 3.1. *For all $\eta > 0$, there exist a $\delta > 0$ and an $n_\eta \in \mathbb{N}$ such that for all $n \geq n_\eta$,*

$$\mathbb{P}\left(M_{in}^{\delta}\right) \geq 1 - \eta, \quad i = 1, \ldots, N.$$

Moreover, there exist a null sequence $(\eta_n)_{n\in\mathbb{N}}$ with corresponding sequence $(\delta_n)_{n\in\mathbb{N}}$ such that

$$\mathbb{P}\left(M_{in}^{\delta_n}\right) \to 1$$

and $\delta_n/\sqrt{n} \to 0$, i.e.

$$N_n(\delta_n, B_i) \to \{B_i\} \quad \text{a.s.}$$

for all $i = 1, \ldots, N$ as $n \to \infty$.

Proof. See Section 9. □

Based on this theorem, we can finally define the PMLE as follows.

Definition 3.2 (PMLE)**.** Let (δ_n) be defined as in Theorem 3.1. Then, the sets

$$M_{in} := M_{in}^{\delta_n}$$

define the PMLE

$$\tilde{\beta}_{in} = \hat{\beta}_{in}\mathbb{1}_{M_{in}}, \quad i = 1, \ldots, N.$$

Theorem 3.1 is a very strong result as it provides asymptotic existence under the unconditional measure $\mathbb{P}$ even though the distribution of B_i has not been specified. The resulting PMLE is more comfortable to work with compared to the ordinary MLE. Especially, $\tilde{\beta}_{in}$ is now square integrable.

Remark (q-variate exponential families). We can easily extend the results from univariate to q-variate simple exponential families. Credibility estimation will be purely based on the PMLEs. Drawing observations from q-variate simple exponential families only affects the construction of these estimators. The extension concerns the likelihood functions, score functions and the Fisher information matrices, which now involve multivariate quantities. The proof of Theorem 3.1 remains valid, see the remark at the end of the proof of Theorem 3.1 in Section 9.

Proposition 3.3. *It holds that $\tilde{\beta}_{in} \in \mathcal{L}^2$ for all $n \in \mathbb{N}$.*

Proof. We have

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\beta}_{in}\|^2\right] &= \mathbb{E}\left[\|\hat{\beta}_{in}\mathbb{1}_{M_{in}}\|^2\right] \\ &= \mathbb{E}\left[\mathbb{1}_{M_{in}}\|B_i + (\hat{\beta}_{in} - B_i)\|^2\right] \\ &\leq \mathbb{E}\left[\mathbb{1}_{M_{in}}\left(\|B_i\|^2 + \|\hat{\beta}_{in} - B_i\|^2 + 2\|B_i\|\|\hat{\beta}_{in} - B_i\|\right)\right] \\ &\leq \left(c_B^2 + \frac{\delta_n^2}{n} + 2c_B\frac{\delta_n}{\sqrt{n}}\right)\mathbb{P}(M_{in}) < \infty.\end{aligned}$$

□

The proof demonstrates how the restriction to this particular M_{in} dramatically simplifies the calculation. Further asymptotic properties are given in the following theorem.

Theorem 3.4. *The PMLE satisfies the following asymptotic properties as the number of observations n grows to infinity.*

i) $\tilde{\beta}_{in}$ *is weakly consistent, i.e.*

$$\tilde{\beta}_{in} \xrightarrow{\mathbb{P}} B_i.$$

ii) $\tilde{\beta}_{in}$ *is asymptotically unbiased, i.e.*

$$\mathbb{E}[\tilde{\beta}_{in}] \to \mathbb{E}[B_i].$$

iii) The second moments converge, i.e.

$$\mathrm{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) \to \mathrm{Cov}(B_i)$$

and

$$\mathrm{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i], B_i\right) \to \mathrm{Cov}(B_i).$$

iv) The conditional second moments converge, i.e.

$$\mathrm{Cov}(\tilde{\beta}_{in} \mid B_i) \to 0$$

almost surely and in $\mathcal{L}^1$.

v) $\tilde{\beta}_{in}$ *is asymptotically Normal, i.e.*

$$F_{in}^{T/2}(\tilde{\beta}_{in})(\tilde{\beta}_{in} - B_i) \xrightarrow{d} \mathcal{N}(0, I).$$

In particular, $F_{in}^{-1}(\tilde{\beta}_{in})$ *is the asymptotic covariance matrix of* $\tilde{\beta}_{in} - B_i$.

Proof. See Section 9. □

All properties except iii) correspond to classical GLM theory, cf. Theorems 1 to 3 in Fahrmeir and Kaufmann (1985). If β_0 denotes the true parameter in a classical GLM , the convergence towards β_0 holds under $\mathbb{P}_{\beta_0}$. Theorem 3.4 generalizes the convergence results to the unconditional measure $\mathbb{P}$. Moreover, the modification to the PMLE does not disturb the convergences. All these properties will play a central role in the next section where we will address the credibility estimation.

4. The credibility estimator

We now redefine

$$L(1,\boldsymbol{\beta}) := \left\{ a + \sum_{k=1}^{N} A_k \tilde{\beta}_k : a \in \mathbb{R}^p, A_k \in \mathbb{R}^{p,p} \right\} \tag{4.1}$$

as the class of admissible estimators by mixing the PMLEs instead of the MLEs. The final credibility estimator only depends on these variables and it does not matter whether the PMLEs belong to a multivariate or a univariate exponential family. It directly follows from Proposition 3.3 that $L(1,\boldsymbol{\beta})$ is a subspace of $\mathcal{L}^2$. Its linearity obviously follows from construction, and since $L(1,\boldsymbol{\beta})$ has finite dimension it is also closed.

Definition 4.1 (GLM credibility estimator)**.** The credibility estimator for B_i is defined as the orthogonal projection

$$\hat{B}_i = \mathrm{Pro}(B_i \mid L(1,\boldsymbol{\beta})) \tag{4.2}$$

of B_i on $L(1,\boldsymbol{\beta})$, or equivalently as

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\boldsymbol{\beta})}{\arg\min}\ \mathbb{E}\left[\left(\tilde{B}_i - B_i \right)' \left(\tilde{B}_i - B_i \right) \right]. \tag{4.3}$$

By choosing $a = 0$ and $A_k = \delta_{ik} I_p$, with $I_p \in \mathbb{R}^{p,p}$ being the identity matrix, one can easily see that $\tilde{\beta}_i \in L(1,\boldsymbol{\beta})$. Therefore, the GLM credibility estimator performs at least as good as the PMLE. Since $\tilde{\beta}_i$ is a weakly consistent and asymptotically unbiased estimator, cf. Theorem 3.4, it is already a good estimator. Moreover, the Bayes estimator $\mathbb{E}[B_i \mid \mathbf{Y}_n]$ is a $\mathbf{Y}_n$-martingale, where $\mathbf{Y}_n = (Y_{i1}, \ldots, Y_{in})_{i=1}^N$, and it converges almost surely and in $\mathcal{L}^1$ to B_i due to the martingale convergence theorem. As $\tilde{\beta}_i$ converges in probability to the same limit B_i, both estimators agree in probability as $n \to \infty$. Thus, the restriction to the linear class $L(1,\boldsymbol{\beta})$ is not a big concern. All these n-asymptotic properties of the PMLE seem to make credibility estimation redundant at first glance. In fact, credibility models target situations where n is not very large. In these cases missing observations can be partially compensated by involving further clusters and this effect can be observed in the credibility formula for B_i, which follows now.

Theorem 4.2. *The credibility estimator is given by*

$$\hat{B}_i = \mathbb{E}[B_i] + A_i(\tilde{\beta}_i - \mathbb{E}[\tilde{\beta}_i]) \tag{4.4}$$

with credibility matrix

$$A_i = \mathrm{Cov}(B_i, \tilde{\beta}_i)\, \mathrm{Cov}(\tilde{\beta}_i)^{-1}. \tag{4.5}$$

Proof. By plugging the linear representation of $\tilde{B}_i$ according to (4.1) into the mean squared error (4.3), we get the objective function

$$f_i(a, A_1, \ldots, A_N) := \mathbb{E}\left[\left(a + \sum_{k=1}^{N} A_k \tilde{\beta}_k - B_i \right)' \left(a + \sum_{k=1}^{N} A_k \tilde{\beta}_k - B_i \right) \right].$$

Taking its partial derivatives with respect to the components of a and the A_l and setting them equal to zero leads to the equations

$$a = \mathbb{E}[B_i] - \sum_{k=1}^{N} A_k \mathbb{E}[\tilde{\beta}_k],$$

$$\mathbb{E}[B_i \tilde{\beta}_l'] = \mathbb{E}[a\tilde{\beta}_l'] + \sum_{k=1}^{N} A_k \mathbb{E}[\tilde{\beta}_k \tilde{\beta}_l'], \quad l = 1, \dots, N,$$

whereat the latter equation simplifies to

$$\operatorname{Cov}(B_i, \tilde{\beta}_l) = \sum_{k=1}^{N} A_k \operatorname{Cov}(\tilde{\beta}_k, \tilde{\beta}_l), \quad l = 1, \dots, N, \tag{4.6}$$

also called the orthogonality conditions. Notice that f_i is differentiable as it is a polynomial with respect to all components of a and A_l. Thus, we may interchange differentiation and integration. Since the function f_i is convex, the solution of these equations is indeed a minimizer. The stochastic components in $\tilde{\beta}_l$ are $\hat{\beta}_l$ and $\mathbb{1}_{M_{ln}}$. Both depend only on (B_l, Y_l), so by assumption (A3), $\tilde{\beta}_l$ and $\tilde{\beta}_k$ are independent for $l \neq k$. Thus, (4.6) simplifies to

$$\operatorname{Cov}(B_i, \tilde{\beta}_l) = A_l \operatorname{Cov}(\tilde{\beta}_l).$$

By the same argument, $\operatorname{Cov}(B_i, \tilde{\beta}_l) = 0$ for $i \neq l$ and it follows that $A_l = 0$ for $i \neq l$. Finally, we obtain

$$A_i = \operatorname{Cov}(B_i, \tilde{\beta}_i) \operatorname{Cov}(\tilde{\beta}_i)^{-1}$$

and putting together the pieces completes the proof. □

Notice that the covariance matrix $\operatorname{Cov}(\tilde{\beta}_i)$ may be singular, which happens if and only if a component of $\tilde{\beta}_i$ is almost surely a linear combination of the others. Such a component is redundant and should be avoided in the stage of modeling by choosing covariate vectors with $p-1$ components. Therefore, without loss of generality we assume that $\operatorname{Cov}(\tilde{\beta}_i)$ is positive definite and, thus, invertible so that the credibility matrix A_i always exists. The resulting credibility formula (4.4) slightly differs from that for the linear regression model (2.4) but an analogous structure can be established in an n-asymptotic meaning. Notice that A_i as well as $\hat{B}_i$ depend on n through the PMLE $\tilde{\beta}_{in}$ and its moments. The additional index n will be used whenever its role is stressed.

Recall that two multivariate sequences $(G_n), (H_n) \subset \mathbb{R}^q$ are said to be *asymptotically equivalent* as $n \to \infty$, written $G_n \overset{n}{\sim} H_n$, if

$$(G_n - H_n) \in o(H_n). \tag{4.7}$$

The little-o symbol describes asymptotic dominance, i.e.

$$\|G_n - H_n\| \leq c\|H_n\|$$

for all $c > 0$ and n large enough. As the name suggests, asymptotic equivalence is indeed an equivalence relation providing reflexivity, symmetry and transitivity. Especially, (4.7) is equivalent to $(G_n - H_n) \in o(G_n)$. We can generalize the notion of asymptotic equivalence to random vectors and matrices by interpreting (4.7) in a probabilistic manner. To this end, we use the $o_{\mathbb{P}}$-notation introduced by Pratt (1959).

Definition 4.3. Let (G_n) and (H_n) be sequences of multivariate random variables. They are *asymptotically equivalent in probability*, written $G_n \overset{n}{\sim} H_n$, if

$$G_n - H_n \in o_{\mathbb{P}}(H_n),$$

i.e. for every $c > 0$,

$$\mathbb{P}(\|G_n - H_n\| \leq c\|H_n\|) \overset{n\to\infty}{\longrightarrow} 1.$$

This is an intuitive generalization of the deterministic counterpart as the convergence of the fraction $\|G_n - H_n\|/\|H_n\|$ towards 0 must now hold in probability. We use the same symbol since its meaning is always clear from the context.

Theorem 4.4. *Two n-asymptotic credibility formulas are given by*

$$\hat{B}_{in} \overset{n}{\sim} A_{in}\tilde{\beta}_{in} + (I_p - A_{in})\mathbb{E}[B_i] \tag{4.8}$$

and

$$\hat{B}_{in} \overset{n}{\sim} A_{in}\tilde{\beta}_{in} + (I_p - A_{in})\mathbb{E}[\tilde{\beta}_{in}], \tag{4.9}$$

where A_{in} is defined as in (4.5).

The right hand side of (4.8) now has the familiar structure as we know it from the linear regression credibility formula (2.4). The credibility estimator is composed of the best individual estimator $\tilde{\beta}_{in}$ and an average value, both weighted according to their credibility. If the individual estimator is evaluated to be highly credible, i.e. $A_{in} \approx I_p$, then the credibility estimator will approximately equal the PMLE. As we will soon see, this is the case for large sample sizes n, where $\tilde{\beta}_{in}$ consistently estimates B_i. The average value part will compensate the lack of information if necessary. Thus, it can be interpreted as a learning effect which vanishes as n increases.

The right hand side of (4.9) can be seen as an empirical variant of (4.8). The empirical version is in particular useful from the meta-analysis perspective, as it involves only the PLME.

The proof of Theorem 4.4 requires some preparation.

Lemma 4.5. *We have*

$$A_{in} \overset{n\to\infty}{\longrightarrow} I_p$$

and consequently $\|\hat{B}_{in} - \tilde{\beta}_{in}\| \to 0$ almost surely and in $\mathcal{L}^1$.

Proof. By (4.5),

$$A_{in} = \operatorname{Cov}(B_i, \tilde{\beta}_{in}) \operatorname{Cov}(\tilde{\beta}_{in})^{-1}$$

so that

$$A_{in} - I_p = \Big(\operatorname{Cov}(B_i, \tilde{\beta}_{in}) - \operatorname{Cov}(\tilde{\beta}_{in})\Big) \operatorname{Cov}(\tilde{\beta}_{in})^{-1}. \tag{4.10}$$

The law of total covariance yields

$$\begin{aligned}\operatorname{Cov}(B_i, \tilde{\beta}_{in}) &= \mathbb{E}\left[\operatorname{Cov}(B_i, \tilde{\beta}_{in} \mid B_i)\right] + \operatorname{Cov}\left(\mathbb{E}[B_i \mid B_i], \mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) \\ &= \operatorname{Cov}\left(B_i, \mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right)\end{aligned}$$

and

$$\operatorname{Cov}(\tilde{\beta}_{in}) = \mathbb{E}\left[\operatorname{Cov}(\tilde{\beta}_{in} \mid B_i)\right] + \operatorname{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right).$$

Hence, (4.10) can be written as

$$A_{in} - I_p = \Big(\operatorname{Cov}\left(B_i, \mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) - \operatorname{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) - \mathbb{E}\left[\operatorname{Cov}(\tilde{\beta}_{in} \mid B_i)\right]\Big) \operatorname{Cov}(\tilde{\beta}_{in})^{-1}.$$

By claims iii) and iv) of Theorem 3.4, the first factor in brackets converges to the zero matrix almost surely and in $\mathcal{L}^1$. Furthermore, $\mathrm{Cov}(\tilde{\beta}_{in})$ converges to $\mathrm{Cov}(B_i)$, which is invertible, and so does $\mathrm{Cov}(\tilde{\beta}_{in})^{-1}$. We conclude that

$$\sup_n \left\| \mathrm{Cov}(\tilde{\beta}_{in})^{-1} \right\| < \infty$$

and finally

$$\|A_{in} - I_p\| \overset{n\to\infty}{\longrightarrow} 0.$$

The convergence of $\hat{B}_{in} - \tilde{\beta}_{in}$ follows since $\sup_n \|\tilde{\beta}_{in}\| < \infty$ almost surely. □

Lemma 4.6. *The credibility estimator $\hat{B}_{in}$ is weakly consistent.*

Proof. The claim directly follows by weak consistence of $\tilde{\beta}_i$ and Lemma 4.5. To be more precise,

$$\begin{aligned}\|\hat{B}_{in} - B_i\| &= \|\mathbb{E}[B_i] + A_{in}\tilde{\beta}_{in} - A_{in}\mathbb{E}[\tilde{\beta}_{in}] - A_{in}B_i - (I_p - A_{in})B_i\| \\ &\leq \|A_{in}(\tilde{\beta}_{in} - B_i)\| + \|A_{in}(\mathbb{E}[B_i] - \mathbb{E}[\tilde{\beta}_{in}])\| + \|(I_p - A_{in})(\mathbb{E}[B_i] - B_i)\|\end{aligned}$$

and the right hand side vanishes in $\mathbb{P}$. □

We can now prove the asymptotic credibility formula (4.9).

Proof of Theorem 4.4. We have to show that

$$\begin{aligned}\left(A_{in}\tilde{\beta}_{in} + (I_p - A_{in})\mathbb{E}[B_i]\right) &\overset{n}{\sim} \left(\mathbb{E}[B_i] + A_{in}(\tilde{\beta}_{in} - \mathbb{E}[\tilde{\beta}_{in}])\right), \\ \left(A_{in}\tilde{\beta}_{in} + (I_p - A_{in})\mathbb{E}[\tilde{\beta}_{in}]\right) &\overset{n}{\sim} \left(\mathbb{E}[B_i] + A_{in}(\tilde{\beta}_{in} - \mathbb{E}[\tilde{\beta}_{in}])\right)\end{aligned} \tag{4.11}$$

in the sense of Definition 4.3. In fact, the differences are

$$A_{in}(\mathbb{E}[\tilde{\beta}_{in}] - \mathbb{E}[B_i]), \qquad \mathbb{E}[\tilde{\beta}_{in}] - \mathbb{E}[B_i],$$

whose norms vanish according to ii) of Theorem 3.4 and Lemma 4.5. It suffices to show that they are also asymptotically dominated by the right hand side of (4.11), which is $\hat{B}_{in}$. Lemma 4.6 implies that $\|\hat{B}_{in}\|$ converges in probability to $\|B_i\|$. Therefore,

$$\frac{\|A_{in}(\mathbb{E}[\tilde{\beta}_{in}] - \mathbb{E}[B_i])\|}{\|\hat{B}_{in}\|}, \qquad \frac{\|\mathbb{E}[\tilde{\beta}_{in}] - \mathbb{E}[B_i]\|}{\|\hat{B}_{in}\|}$$

are products of converging sequences. They converge in probability to 0, which is the product of the limits, provided that $\mathbb{P}(\|B_i\| = 0) = 0$. The latter is guaranteed by condition (R1) so that the claim follows. □

5. Estimation of the hyperparameters

The structural parameters to be estimated in (4.5) and (4.9) are $\mathbb{E}[\tilde{\beta}_i]$, $\mathrm{Cov}(\tilde{\beta}_i)^{-1}$ and $\mathrm{Cov}(B_i, \tilde{\beta}_i)$. Difficulties arise since we only have one iid sample for each $\tilde{\beta}_i$. The PMLE have in common that each of them is a weakly consistent estimator of the corresponding B_i. These target variables B_i are iid. Thus, cluster specific effects will vanish as $n \to \infty$, and this property will be repeatedly used in estimation. For that purpose, we will use the variables

$$T := \mathrm{Cov}(B_i)$$

and

$$S_i := \mathbb{E}\left[\mathrm{Cov}(\tilde{\beta}_{in} \mid B_i)\right].$$

By Theorem 3.4, they allow for the asymptotic decomposition

$$\operatorname{Cov}(\tilde{\beta}_{in}) \overset{n}{\sim} T + S_i, \tag{5.1}$$

which consists of a cluster common and a cluster specific term. We first motivate the estimators using typical structures and then discuss the necessary changes for the particular setting. Readers who are mainly interested in the results may jump over to Theorem 5.2.

ad $\mathbb{E}[\tilde{\beta}_{in}]$. A simple estimator is given by the sample mean

$$\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} \tilde{\beta}_{in}.$$

The n-asymptotic unbiasedness of $\tilde{\beta}_{in}$ yields

$$\begin{aligned}
\lim_{n\to\infty}\lim_{N\to\infty}\left\|\hat{\beta}_0 - \mathbb{E}[\tilde{\beta}_{in}]\right\| &\leq \lim_{n\to\infty}\lim_{N\to\infty}\left\|\hat{\beta}_0 - \mathbb{E}[B_i]\right\| + \lim_{n\to\infty}\left\|\mathbb{E}[B_i] - \mathbb{E}[\tilde{\beta}_{in}]\right\| \\
&= \lim_{n\to\infty}\lim_{N\to\infty}\left\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\beta}_{in} - B_i) + (B_i - \mathbb{E}[B_i])\right\| \\
&\leq \lim_{n\to\infty}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\beta}_{in} - B_i\right\| + \lim_{N\to\infty}\left\|\frac{1}{N}\sum_{i=1}^{N} B_i - \mathbb{E}[B_i]\right\|.
\end{aligned}$$

The second sum almost surely vanishes by the strong law of large numbers. For the first sum, we use the dominated convergence theorem. All its summands are almost surely bounded such that

$$\lim_{N\to\infty}\sup_{n\in\mathbb{N}}\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\beta}_{in} - B_i\right\| \leq \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sup_{n\in\mathbb{N}}\left\|\tilde{\beta}_{in} - B_i\right\| < \infty$$

and therefore

$$\lim_{n\to\infty}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\beta}_{in} - B_i\right\| = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\lim_{n\to\infty}\left\|\tilde{\beta}_{in} - B_i\right\| = 0.$$

The last convergence holds in probability, cf. Theorem 3.4. Hence, $\hat{\beta}_0$ properly estimates all $\mathbb{E}[\tilde{\beta}_{in}]$. It is weakly consistent as $n \to \infty$ and $N \to \infty$ and also n-asymptotically unbiased. Nonetheless, we propose a weighted sample mean of the type

$$\hat{\beta}_0 = \left(\sum_{i=1}^{N} C_i\right)^{-1}\sum_{i=1}^{N} C_i\tilde{\beta}_i \tag{5.2}$$

with $C_i \in \mathbb{R}^{p,p}$. The sample mean can be written in form of structure (5.2) by selecting constant weights $C_i \equiv C$. The estimator obtained by choosing the C_i as the inverse covariance matrices of the $\tilde{\beta}_i$ has minimum MSE among all estimators of type (5.2). Thus, we choose

$$\hat{\beta}_0 := \left(\sum_{i=1}^{N}\operatorname{Cov}(\tilde{\beta}_{in})^{-1}\right)^{-1}\sum_{i=1}^{N}\operatorname{Cov}(\tilde{\beta}_{in})^{-1}\tilde{\beta}_{in}. \tag{5.3}$$

Notice that $\hat{\beta}_0$ does not depend on i although the target variable $\mathbb{E}[\tilde{\beta}_{in}]$ does. The idea behind this is the asymptotic equivalence of the first moments, i.e.

$$\mathbb{E}[\tilde{\beta}_{in}] \overset{n}{\sim} \mathbb{E}[B_i] =: \beta_0, \quad i = 1, \ldots, n. \tag{5.4}$$

For calculation of (5.3), the inverse covariance matrices in (5.3) have to be replaced by their estimators which follow soon. Besides that, structure (5.3) can be differently interpreted by means of the credibility matrices A_i. Since

$$\operatorname{Cov}(B_i, \tilde{\beta}_{in}) \overset{n}{\sim} \operatorname{Cov}(B_i) = T,$$

we have $A_i \overset{n}{\sim} T \operatorname{Cov}(\tilde{\beta}_{in})^{-1}$ so that

$$\hat{\beta}_0 = \left(\sum_{i=1}^{N} \operatorname{Cov}(\tilde{\beta}_i)^{-1} \right)^{-1} T^{-1} \sum_{i=1}^{N} T \operatorname{Cov}(\tilde{\beta}_i)^{-1} \tilde{\beta}_i \overset{n}{\sim} \left(\sum_{i=1}^{N} A_i \right)^{-1} \sum_{i=1}^{N} A_i \tilde{\beta}_i.$$

The right hand side is the credibility weighted sample mean and the asymptotic equivalence will become an equality if we plug in the final estimators $\hat{A}_i$ of A_i.

ad $\operatorname{Cov}(\tilde{\beta}_i)^{-1}$ *and* $\operatorname{Cov}(B_i, \tilde{\beta}_i)$. In a first step we analyze the sample covariance matrix

$$\hat{\tau} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\tilde{\beta}_i - \frac{1}{N} \sum_{l=1}^{N} \tilde{\beta}_l \right) \left(\tilde{\beta}_i - \frac{1}{N} \sum_{j=l}^{N} \tilde{\beta}_l \right).$$

Using the independence of the $\tilde{\beta}_i$, we get

$$\begin{aligned}
\mathbb{E}[\hat{\tau}] &= \frac{1}{N-1} \sum_{i=1}^{N} \mathbb{E} \left[\left(\tilde{\beta}_i - \frac{1}{N} \sum_{l=1}^{N} \tilde{\beta}_l \right) \left(\tilde{\beta}_i - \frac{1}{N} \sum_{l=1}^{N} \tilde{\beta}_l \right)' \right] \\
&= \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{N-2}{N} \mathbb{E} \left[\tilde{\beta}_i \tilde{\beta}_i' \right] - \frac{2}{N} \sum_{\substack{l=1 \\ l \neq i}}^{N} \mathbb{E}[\tilde{\beta}_i] \mathbb{E}[\tilde{\beta}_l'] \right. \\
&\qquad \left. + \frac{1}{N^2} \sum_{l=1}^{N} \mathbb{E} \left[\tilde{\beta}_l \tilde{\beta}_l' \right] + \frac{1}{N^2} \sum_{l=1}^{N} \sum_{\substack{k=1 \\ l \neq l}}^{N} \mathbb{E}[\tilde{\beta}_l] \mathbb{E}[\tilde{\beta}_k'] \right).
\end{aligned}$$

Further, by (5.4),

$$\begin{aligned}
\mathbb{E}[\hat{\tau}] &\overset{n}{\sim} \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{N-2}{N} \left(\operatorname{Cov}(\tilde{\beta}_i) + \beta_0 \beta_0' \right) - \frac{2(N-1)}{N} \beta_0 \beta_0' \right. \\
&\qquad \left. + \frac{1}{N^2} \sum_{l=1}^{N} \operatorname{Cov}(\tilde{\beta}_l) + \frac{1}{N} \beta_0 \beta_0' + \frac{N-1}{N} \beta_0 \beta_0' \right) \\
&= \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{N-2}{N} \operatorname{Cov}(\tilde{\beta}_i) + \frac{1}{N^2} \sum_{l=1}^{N} \operatorname{Cov}(\tilde{\beta}_l) \right) \\
&= \frac{N-2}{N(N-1)} \sum_{i=1}^{N} \operatorname{Cov}(\tilde{\beta}_i) + \frac{1}{N(N-1)} \sum_{i=1}^{N} \operatorname{Cov}(\tilde{\beta}_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} \operatorname{Cov}(\tilde{\beta}_i).
\end{aligned}$$

This simplification will not be possible if we use the weighted sample mean (5.3) instead of $\frac{1}{N} \sum_l \tilde{\beta}_{ln}$ in $\hat{\tau}$. The $\tilde{\beta}_{in}$ have different covariance matrices and $\hat{\tau}$ estimates their sample mean.

Hence, for some specific cluster i, $\hat{\tau}$ over- or underestimates $\mathrm{Cov}(\tilde{\beta}_i)$ depending on the constellation of the portfolio. In order to remove the systematic error, we use decomposition (5.1). We then obtain

$$\mathbb{E}[\hat{\tau}] \overset{n}{\sim} \mathrm{Cov}(B_i) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\mathrm{Cov}(\tilde{\beta}_i \mid B_i)] = T + \frac{1}{N}\sum_{i=1}^{N} S_i. \tag{5.5}$$

By (5.5), we set

$$\hat{T} := \hat{\tau} - \frac{1}{N}\sum_{i=1}^{N} \hat{S}_i \tag{5.6}$$

and by (5.1)

$$\widehat{\tau_i^{-1}} := \widehat{\mathrm{Cov}(\tilde{\beta}_{in})^{-1}} := \left(\hat{T} + \hat{S}_i\right)^{-1}. \tag{5.7}$$

It remains to find estimators for the S_i.

ad S_i. For the estimation of

$$S_i = \mathbb{E}[\mathrm{Cov}(\tilde{\beta}_i \mid B_i)],$$

we use that the inverse Fisher information matrix is the asymptotic covariance matrix of $\tilde{\beta}_{in}$ given B_i, cf. Theorem 3.4. Thus, we propose

$$\hat{S}_i := \frac{1}{N}\sum_{l=1}^{N} F_{in}^{-1}(\tilde{\beta}_l). \tag{5.8}$$

The idea behind this choice is that $F_{in}^{-1}(B_i)$ is, conditional on B_i, also the asymptotic covariance matrix of $\tilde{\beta}_i$. See Theorem 3 in Fahrmeir and Kaufmann (1985). Since $B_1, \ldots, B_N$ are iid,

$$\frac{1}{N}\sum_{l=1}^{N} F_{in}^{-1}(B_l)$$

consistently estimates $\mathbb{E}[F_{in}^{-1}(B_i)]$ as portfolio size N increases and (5.8) replaces the B_l by $\tilde{\beta}_l$. The following lemma justifies this procedure.

Lemma 5.1. *For all i and $l = 1, \ldots, N$, $F_{in}(B_l)$ and $F_{in}(\tilde{\beta}_l)$ as well as $F_{in}^{-1}(B_l)$ and $F_{in}^{-1}(\tilde{\beta}_l)$ are asymptotically equivalent in probability.*

Proof. First we show that $\frac{1}{n}F_{in}(B_l) - \frac{1}{n}F_{in}(\tilde{\beta}_l)$ converges in probability to zero. In fact, for $\epsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n}F_{in}(B_l) - \frac{1}{n}F_{in}(\tilde{\beta}_l)\right\| > \epsilon\right) \\
&= \mathbb{P}\left(\left\|\frac{1}{n}F_{in}(B_l) - \frac{1}{n}F_{in}(\tilde{\beta}_l)\right\| > \epsilon \mid M_{ln}\right)\mathbb{P}(M_{ln}) + \mathbb{P}\left(\left\|\frac{1}{n}F_{in}(B_l) - \frac{1}{n}F_{in}(\tilde{\beta}_l)\right\| > \epsilon \mid M_{ln}^c\right)\mathbb{P}(M_{ln}^c) \\
&\leq \mathbb{P}(M_{ln})\left(\mathbb{P}\left(\left\|\frac{1}{n}F_{in}(B_l) - \frac{1}{n}F_{in}(\tilde{\beta}_l)\right\| > \epsilon \mid M_{ln}, \tilde{\beta}_{ln} \in \mathcal{B}\right)\mathbb{P}(\tilde{\beta}_{ln} \in \mathcal{B} \mid M_{ln})\right. \\
&\qquad \left. + \mathbb{P}\left(\left\|\frac{1}{n}F_{in}(B_l) - \frac{1}{n}F_{in}(\tilde{\beta}_l)\right\| > \epsilon \mid M_{ln}, \tilde{\beta}_{ln} \notin \mathcal{B}\right)\mathbb{P}(\tilde{\beta}_{ln} \notin \mathcal{B} \mid M_{ln})\right) + \mathbb{P}(M_{ln}^c).
\end{aligned}$$

F_{in} is Lipschitz continuous on $\mathcal{B}$ with a Lipschitz constant $L > 0$ that does not depend on n. Furthermore, on M_{ln},

$$\tilde{\beta}_{ln} = B_l + (\tilde{\beta}_{ln} - B_l)$$

with $\|\tilde{\beta}_{ln} - B_l\| \leq \frac{\delta_n}{\sqrt{n}}$. By (3.3), B_l almost surely lies in the interior of $\mathcal{B}$, i.e. there exists an η-neighborhood around B_l that is completely included in $\mathcal{B}$. Thus, since $\frac{\delta_n}{\sqrt{n}} \to 0$,

$$P(\tilde{\beta}_{ln} \in \mathcal{B} \mid M_{ln}) \overset{n\to\infty}{\longrightarrow} 1.$$

Altogether, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n}F_{in}(B_l) - \frac{1}{n}F_{in}(\tilde{\beta}_l)\right\| > \epsilon\right) \\
&\leq \left(\mathbb{P}\left(L\left\|\tilde{\beta}_l - B_l\right\| > \epsilon \mid M_{ln}, \tilde{\beta}_{ln} \in \mathcal{B}\right)\mathbb{P}(\tilde{\beta}_{ln} \in \mathcal{B} \mid M_{ln}) + \mathbb{P}(\tilde{\beta}_{ln} \notin \mathcal{B} \mid M_{ln})\right)\mathbb{P}(M_{ln}) + \mathbb{P}(M_{ln}^c) \\
&\overset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

On the other hand, the asymptote $\frac{1}{n}F_{in}(B_l)$ almost surely converges to a positive definite matrix $F_i(B_l)$. The quotient

$$\frac{\left\|F_{in}(B_j) - F_{in}(\tilde{\beta}_j)\right\|}{\|F_{in}(B_j)\|}$$

converges in probability to zero so that asymptotic equivalence follows. The proof for the inverse sequences works similarly. □

We summarize the results in the following theorem.

Theorem 5.2. *The structural parameters can be estimated as follows.*

i) A weakly consistent estimator for $\mathbb{E}[\tilde{\beta}_{in}]$ as both $n \to \infty$ and $N \to \infty$ is given by

$$\hat{\beta}_0 = \left(\sum_{i=1}^{N}\widehat{\tau_i^{-1}}\right)^{-1}\sum_{i=1}^{N}\widehat{\tau_i^{-1}}\tilde{\beta}_{in}, \tag{5.9}$$

where $\widehat{\tau_i^{-1}}$ is defined in (5.7).

ii) The random matrix

$$\hat{S}_i = \frac{1}{N}\sum_{l=1}^{N}F_{in}^{-1}(\tilde{\beta}_l)$$

is an asymptotically unbiased and weakly consistent estimator for S_i as both $n \to \infty$ and $N \to \infty$.

iii) Let $\hat{\tau}$ be the sample covariance matrix of $(\tilde{\beta}_i)_{i=1}^N$. Then,

$$\hat{T} = \hat{\tau} - \frac{1}{N}\sum_{i=1}^{N}\hat{S}_i \tag{5.10}$$

is an asymptotically unbiased and weakly consistent estimator of $\mathrm{Cov}(B_i, \tilde{\beta}_i)^{-1}$ as both $n \to \infty$ and $N \to \infty$.

iv) A weakly consistent estimator of $\mathrm{Cov}(\tilde{\beta}_i)^{-1}$ as both $n \to \infty$ and $N \to \infty$ is given by

$$\widehat{\tau_i^{-1}} = \left(\hat{T} + \hat{S}_i\right)^{-1}.$$

v) The credibility matrix (4.5) *can be estimated by*

$$\hat{A}_i = \hat{T}\widehat{\tau_i^{-1}},$$

which is a weakly consistent estimator as both $n \to \infty$ and $N \to \infty$.

Finally, the credibility formula can be evaluated by replacing its hyperparameters (structural parameters) by their estimators.

Corollary 5.3. *The estimator*

$$\hat{\tilde{B}}_i = \hat{A}_i\tilde{\beta}_i + (I_p - \hat{A}_i)\hat{\beta}_0 \tag{5.11}$$

is a weakly consistent estimator for the exact credibility estimator (4.4) *as both* $n \to \infty$ *and* $N \to \infty$.

Proof. The claim is a direct consequence of Theorem 4.4 and Theorem 5.2. □

As $\hat{T}$ estimates a covariance matrix, it should be symmetric and positive semidefinite. By the nature of structure (5.10), $\hat{T}$ is symmetric but not necessarily positive semidefinite. If in practical applications $\hat{T}$ is indeed not positive semidefinite, we suggest to use the following modification: There exist an orthogonal matrix Q and a diagonal matrix D such that

$$D = Q'\hat{T}Q,$$

where the diagonal elements of D are the eigenvalues of $\hat{T}$. We construct a new matrix D^* by replacing all negative entries of D by zero. Then, a positive semidefinite alternative $\hat{T}^*$ for $\hat{T}$ is given by

$$\hat{T}^* = QD^*Q'.$$

Instability and lack of positive definiteness is a general problem in empirical covariance estimation. As an alternative, the literature offers various approaches that use the concept of shrinkage.

6. Relative gain in efficiency

To understand the benefits of credibility estimation, we take a closer look at the mean squared errors

$$\mathbb{E}[\|\hat{B}_i - B_i\|^2] \quad \text{and} \quad \mathbb{E}[\|\tilde{\beta}_i - B_i\|^2].$$

The effort of considering the whole portfolio instead of a single cluster will be worth if the improvement in terms of the mean squared errors is large, i.e. if

$$\mathrm{rMSE}_i = \frac{\mathbb{E}[\|\hat{B}_i - B_i\|^2]}{\mathbb{E}[\|\tilde{\beta}_i - B_i\|^2]} \tag{6.1}$$

is clearly smaller than 1. By definition of the credibility estimator, $\mathrm{rMSE}_i \leq 1$ always holds.

Simulation Study. We evaluate the performance of the credibility estimators by means of the ratios of the simulated mean squared errors (simulated rMSE)

$$\frac{\sum_{k=1}^m \|\hat{\tilde{B}}_i(\omega_k) - B_i(\omega_k)\|^2}{\sum_{k=1}^m \|\tilde{\beta}_i(\omega_k) - B_i(\omega_k)\|^2}, \quad i = 1, \dots N. \tag{6.2}$$

A value of (6.2) smaller than 1 means that the credibility estimator performs better than the PMLE. Its computation is based on $m = 10000$ scenarios denoted by $\omega_1, \dots, \omega_m$.

Poisson case. We consider a Poisson-CGLM for several constellations of the portfolio and sample sizes (N, n). The B_i are independently drawn from a Normal distribution with mean vector $(2, 1)$ and covariance matrix I_2. The covariate vectors are given by

$$X_{ij} = X_j = \begin{pmatrix} 1 & \frac{j}{n} \end{pmatrix}, \quad j = 1, \dots, n,$$

including an overall and a linear effect. Table 6.1 shows the simulated rMSEs for cluster $i = 1$. The columns labeled no.1 list the relative improvement of the credibility estimator without

$N\backslash n$	15		25		50		100	
	no.1	no.2	no.1	no.2	no.1	no.2	no.1	no.2
5	2.14	1.02	3.33	1.06	5.89	1.05	12.04	1.05
10	1.12	0.90	1.47	0.95	2.02	0.98	3.28	0.98
20	0.90	0.85	1.01	0.90	1.18	0.95	1.49	0.97
30	0.86	0.83	0.93	0.89	1.04	0.94	1.20	0.97
50	0.82	0.81	0.89	0.88	0.97	0.93	1.05	0.96
MSE	0.132		0.078		0.038		0.019	

TABLE 6.1. Relative improvement of the credibility estimator compared to the PMLE. The line MSE shows the simulated mean squared error of the PMLE.

making $\hat{T}$ positive semidefinite. The modification is applied in the estimation whose results are given in columns no.2. There are noticeable differences between these two estimators and the modification is in fact absolutely essential in the cases $n = 50$ and $n = 100$ to make the credibility estimator performing better than the PMLE. Generally, extremal behavior can be observed in the first row ($N = 5$) and in the last column ($n = 100$). If the portfolio contains only a small number of clusters, the estimation of the structural parameters as described in Theorem 5.2 will not work well. There are simply too few independent observations to estimate the empirical means and covariance matrices properly. When sample size n is large, the relative improvement is very small. That does not mean that credibility estimation performs badly. Rather, the opposite is the case: The credibility estimator is as good as the PMLE, which is already itself a good estimator. In all other constellations of (N, n), the credibility estimator shows considerable improvements in sense of the mean squared error. Lack of statistical information in single clusters can be compensated by the huge amount of information that the portfolio delivers. It is also remarkable that the n-asymptotic credibility formula delivers a good approximation for even a small number of n, e.g. $n = 15$.

Binomial case. We consider a portfolio of $N = 30$ independent clusters each with sample size $n = 25$. The assumptions on the B_i and X_{ij} remain the same as in the Poisson case. Conditionally, given B_i, the Y_{ij} follow a Binomial distribution with success probability characterized through the linear predictor $X_j B_i$ via the logit-link. The weights w_{ij} in (2.8) are equal to the nuisance parameters, i.e. the number of trials, which we choose as

$$w_{ij} = 10 + i$$

for all i and j. Figure 6.1 shows the values of (6.2) for several constellations and estimators. The top left plot belongs to the current case of $N = 30$ and includes the simulated rMSEs for two different estimators. The solid line (1) represents our proposed credibility estimator, i.e. $\hat{T}$ is made positive semidefinite and $\hat{\beta}_0$ is the weighted sample mean (5.9). Line (2) uses the unweighted sample mean of the PMLEs for the estimation of the $\mathbb{E}[\tilde{\beta}_i]$. The credibility estimators perform well for each cluster and we observe improvements in a range between 30% and 65%. Improvements compared to the PMLE are especially large for clusters with small weights, i.e. a small number of trials. The line (2) is hardly visible since it is almost identical to (1) but there is a minor advantage for (1) in the per thousand range. In fact, both estimators for $\mathbb{E}[\tilde{\beta}_i]$ and also the credibility estimators $\hat{\tilde{B}}_i$ do not show any noteworthy differences. However, it should be mentioned that the weighted sample mean $\hat{\beta}_0$ has, as originally motivated, a lower variance in both of its components. The last modification concerns $\hat{T}$ and we strongly recommend to use the positive semidefinite version of $\hat{T}$. Without this modification, the simulated ratio of mean

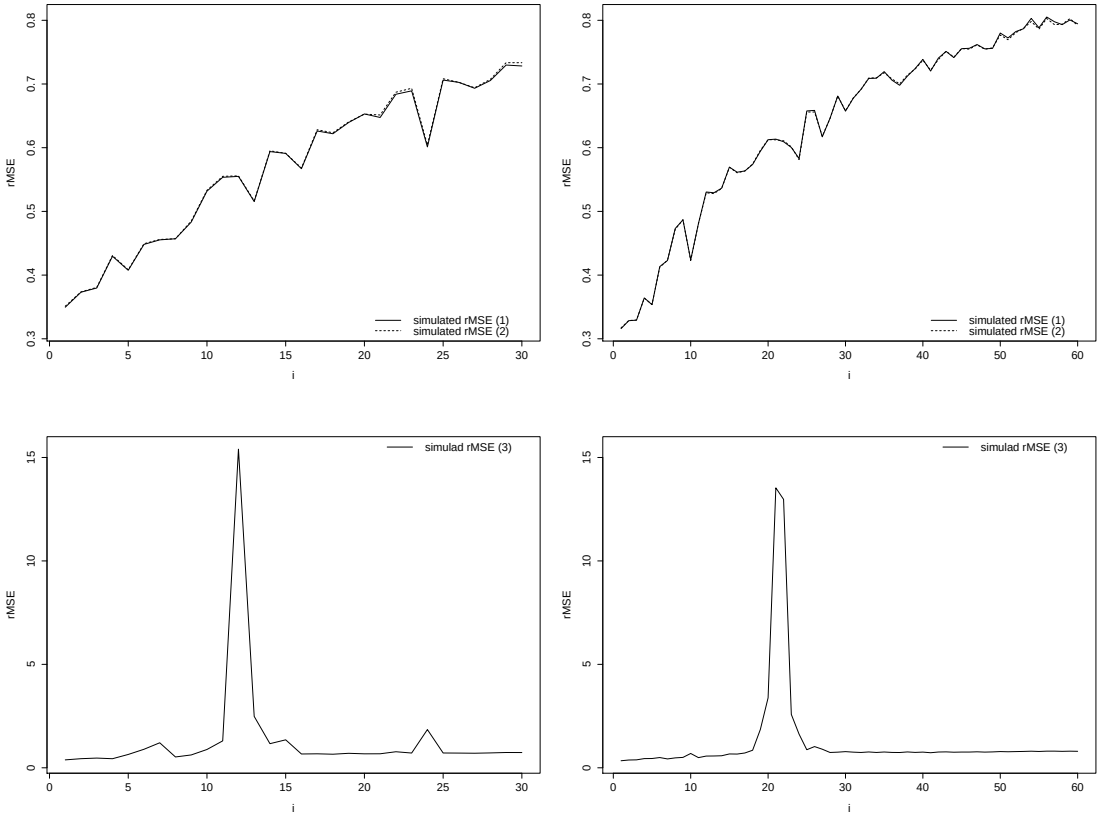

FIGURE 6.1. Plots of the rMSE for a Binomial-CGLM with $N = 30$ (top and bottom left) and $N = 60$ (top and bottom right). The lines correspond to the following estimators.

simulated rMSE (1): The proposed credibility estimator with positive semidefinite $\hat{T}$ and weighted sample mean $\hat{\beta}_0$.

simulated rMSE (2): The credibility estimator using the unweighted sample mean instead.

simulated rMSE (3): The credibility estimator using a non-modified version of $\hat{T}$ instead.

squared errors without are in some cases clearly greater than 1, see bottom left plot in Figure 6.1. In the worst case it reaches the value 15.40 and indicates a bad performance of the credibility estimator. The top right and bottom right plots in Figure 6.1 are the analogs of the left ones for the case $N = 60$. Estimation of the structural parameters should be more accurate and in fact, the simulated rMSEs of the first 30 clusters have improved by around 5%. When we take a look at the modified estimators, we observe the same behavior as in the case $N = 30$. Estimator (2) is almost identical to (1) but a non-modified $\hat{T}$ should be avoided, see bottom right plot.

7. Application to mortality data

GLMs are a popular tool for studying mortality data of single populations. If also mortality data for related populations is available, the CGLM approach can help to improve statistical estimates for single populations.

The model. Suppose that we observe mortality statistics for people of ages $x = x_1, \dots, x_m$ and calendar years $t = 1, \dots, T$, where m is the number of age groups. For example, the Human-Mortality-Database (2014) provides death counts $D_x(t)$ as well as the initial exposure-to-risk $E_x(t)$ for each age x and year t. Poisson models and Poisson-GLMs are typical tools for studying counted data and this also applies for the current case of mortality data. For instance, the Cairns-Blake-Dowd model (CBD model) by Cairns et al. (2006) is a Poisson model with the structure

$$\log \mathbb{E}[D_x(t)] = \log E_x(t) + \kappa_1(t) + \kappa_2(t)(x - \bar{x}) \tag{7.1}$$

for $x = x_1, \dots, x_m$ and $t = 1, \dots, T$. The two κ-terms denote the age independent and age dependent period effects, respectively. The latter describes the impact of linear age effects, where $\bar{x} = \frac{1}{2}(x_1 + x_m)$ is the central age in fit. Model (7.1) can be also written in form of a GLM by choosing the GLM parameter

$$\beta = \begin{pmatrix} \kappa_1(1) & \dots & \kappa_1(T) & \kappa_2(1) & \dots & \kappa_2(T) \end{pmatrix}' \in \mathbb{R}^{2T}$$

and the design matrix $X = (X^{(1)}, X^{(2)}) \in \mathbb{R}^{mT,2T}$ with

$$X^{(1)} = I_T \otimes 1_m = I_T \otimes \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and} \quad X^{(2)} = I_T \otimes \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_m - \bar{x} \end{pmatrix}. \tag{7.2}$$

Recall that the Kronecker product of matrices $G = (g_{ij}) \in \mathbb{R}^{a,b}$ and $H \in \mathbb{R}^{c,d}$ is given by

$$G \otimes H = \begin{pmatrix} g_{11}H & \dots & g_{1b}H \\ \vdots & \ddots & \vdots \\ g_{a1}H & \dots & g_{ab}H \end{pmatrix} \in \mathbb{R}^{ac,bd}.$$

It is easy to check that the linear predictor with offset $\log E_x(t)$ represents the right hand side of (7.1). The GLM can now be solved for the unknown period parameters $\kappa_1(t)$ and $\kappa_2(t)$.

We adjust the CBD model (7.1) to fit into the credibility framework described by (A1) to (A3) in Section 2. The observed death counts are assumed to be drawn from random variables $D_{ix}(t)$ that, conditionally on B_i, follow a Poisson distribution. The linear predictor is given by

$$\log \mathbb{E}[D_{ix}(t) \mid B_i] = \log E_{ix}(t) + K_{i1}(t) + K_{i2}(t)(x - \bar{x}), \tag{7.3}$$

where the first T entries of B_i correspond to the process $K_{i1} = (K_{i1}(t))_{t=1}^T$ and the last T to $K_{i2} = (K_{i2}(t))_{t=1}^T$. The two processes are assumed to be independent. We can reformulate structure (7.3) in the form of (2.9) by using the design matrix $X = (X^{(1)}, X^{(2)})$ as defined in (7.2).

Empirical results. Our portfolio consists of $N = 36$ countries which are Australia, Austria, Belarus, Belgium, Bulgaria, Canada, Czech, Denmark, Estonia, Finland, France, East Germany, West Germany, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Netherlands, New Zealand, Norway, Poland, Portugal, Russia, Slovakia, Scotland, Spain, Sweden, Switzerland, Taiwan, UK, USA and Ukraine. These are all countries of the Human-Mortality-Database (2014) for which data is available from 1980 on or earlier. Note that Germany is

separately considered in its eastern and western part. We fit the CBD model for the ages 90 to 100 and the calender years 1980 to 2009, thus $m = 11$ and $T = 30$. First, individual maximum likelihood estimators are obtained by fitting the conditional model, which is (7.1), separately for all countries. In doing so, we use the normed design matrix

$$\left(\frac{X^{(1)}}{\|X^{(1)}\|} \quad \frac{X^{(2)}}{\|X^{(2)}\|}\right)$$

to make the two period effects comparable. The estimators are then combined to the credibility estimators according to formula (5.11). We are especially interested in the countries where the credibility estimates clearly differ from the conditional maximum likelihood estimates. Figure 7.1 shows the quadratic deviations. The credibility and maximum likelihood estimators almost agree for most of the countries but Iceland and Luxembourg clearly stand out. Since Iceland and Luxembourg have the smallest population of all N countries, the result is plausible and agrees with the theoretical idea of credibility theory. The plots in Figure 7.2 display the estimators for K_{i1} of Iceland. An interesting behavior can be observed for Iceland in year 1983, when none of its three 100 years old citizens died. The MLE is strongly affected and exhibits a downward peak. The same incidence happens in Luxembourg, 1984. In that year, all 22 of the 96 year old people survived. If we look at the collective estimates (line beta0) for the years 1983 and 1984 we cannot recognize any unusual downward movements. The credibility estimators weaken the peaks of the MLE. There are alternative ways to smooth the estimators, for example using a penalty term in the likelihood function, see Currie et al. (2004). The advantage of the CGLM approach is that it relies much more on empirical observations than a apriori model assumptions.

The empirical example shows that the credibility approach becomes useful when populations are studied that have the size of Luxemburg's populaton or smaller. In recent years many life insurers in Europe started to use life tables for increasingly smaller sub-populations, differentiating by postcode areas, smoking habits, and so on. This trend towards smaller sub-populations makes the credibility approach increasingly relevant and useful.

8. Discussion

Theoretically every GLM can be extended to a CGLM as far as data is available and properly structured. Yet, additional regularity conditions are needed in order to verify that the credibility estimators really behave well. We suggested a set of regularity conditions that define a large class of canonical GLMs. For this class we calculated asymptotic properties for the corresponding credibility estimators, showing that the credibility estimators indeed behave well. The simulation study and the mortality example indicate that the theoretical gain in efficiency indeed materializes in practice whenever the number of clusters is large enough. Special caution is needed for the estimation of the hyperparameters (structural parameters). We presented and discussed several modifications for the estimators, but it seems that there is still room for improvement.

9. Proofs

The remainder of the paper is devoted for proofs and we begin with that of Theorem 3.1 which claims asymptotic existence of the PMLE.

Proof of Theorem 3.1.

Lemma 9.1. *The scaled Fisher information matrices converge uniformly on $\mathcal{B}$, i.e.*

$$\sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n} F_{in}(\beta) - F_i(\beta) \right\| \to 0$$

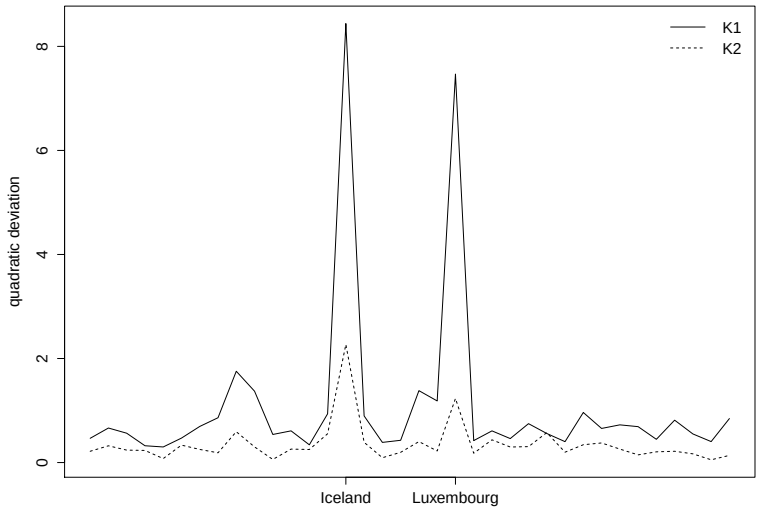

FIGURE 7.1. The quadratic deviations between the credibility and ML estimators for K_{i1} and K_{i2} respectively.

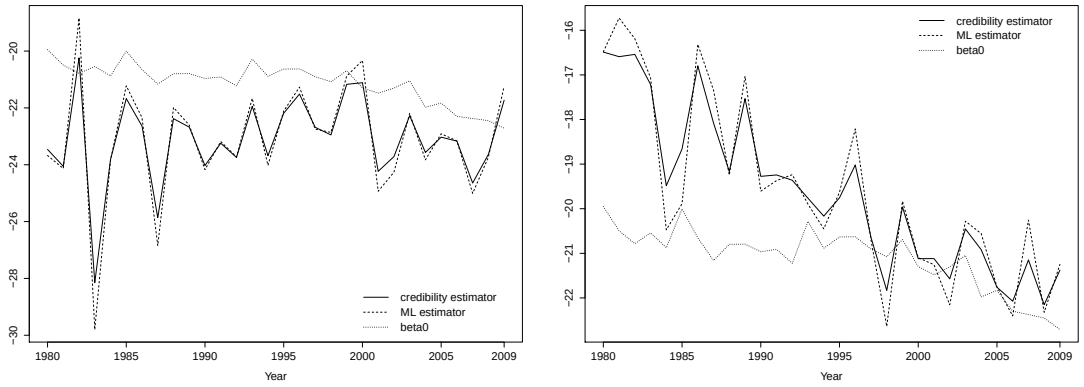

FIGURE 7.2. The credibility CBD model for high ages, estimators for K_{i1} of Iceland (left) and Luxembourg (right).

as $n \to \infty$ *for all* $i = 1, \ldots, N$.

Proof. Since $\frac{1}{n}F_{in}$ converges pointwise and its domain $\mathcal{B}$ is compact, uniform convergence is guaranteed if the sequence $(\frac{1}{n}F_{in})_n$ is equicontinuous. This is indeed the case. First recall that a sequence of functions (g_n) on D is said to be equicontinuous if for all $x \in D$ and $\epsilon > 0$, there exists a $\delta_{\epsilon,x} > 0$ such that for all $y \in D$ with $\|y - x\| < \delta_{\epsilon,x}$ and all $n \in \mathbb{N}$,

$$\|g_n(x) - g_n(y)\| < \epsilon.$$

In short, the δ depends only on ϵ and x but not on n. A sufficient condition for equicontinuity is that the family is Lipschitz continuous with the same Lipschitz constant. In fact, for all $\beta_1, \beta_2 \in \mathcal{B}$, the mean value theorem yields that

$$\begin{aligned}\left\|\frac{1}{n}F_{in}(\beta_1) - \frac{1}{n}F_{in}(\beta_2)/n\right\| &\leq \frac{1}{n}\sum_{j=1}^{n}\left\|w_{ij}X_{ij}'X_{ij}\left(v_{ij}(\beta_1) - v_{ij}(\beta_2)\right)\right\| \\ &\leq \frac{1}{n}\sum_{j=1}^{n}|w_{ij}|\left\|X_{ij}'X_{ij}\right\|\left\|\sup_{\gamma\in\overline{\beta_1\beta_2}}X_{ij}'b^{(3)}(\xi_{ij}+X_{ij}\gamma)\right\|\|\beta_1-\beta_2\| \\ &\leq L\|\beta_1-\beta_2\|.\end{aligned}$$

Such a bound $L > 0$ exists since $b' = g^{-1}$ is twice continuously differentiable by (R2) and since the domains of w_{ij}, ξ_{ij}, X_{ij} and γ are bounded. □

Throughout the following proofs the cluster index i is omitted for notational simplicity, e.g. a particular element of $\mathbf{B} = (B_1, \ldots, B_N)$ will be denoted by B instead of B_i.

The set M_n^δ. Recall the constructions (3.4)

$$N_n(\delta, B) = \{\beta \in \mathcal{B} : \sqrt{n}\|\beta - B\| \leq \delta\}, \quad n \in \mathbb{N},$$

for $\delta > 0$ and (3.5)

$$M_n^\delta = \{l_n(\beta) - l_n(B) < 0, \quad \text{for all } \beta \in \partial N_n(\delta, B)\}.$$

We have already seen that

$$\sqrt{n}\|\hat{\beta}_n - B\| \leq \delta$$

whenever the event M_n^δ occurs. What remains to show is that $1-\mathbb{P}(M_n^\delta)$ vanishes for a particular choice of δ. More precisely, it suffices to prove that for all $\eta > 0$ there exist a $\delta > 0$ and an $n_\eta \in \mathbb{N}$ such that

$$\mathbb{P}\big(\exists\beta \in \partial N_n(\delta, B) : l_n(\beta) - l_n(B) \geq 0\big) \leq \eta, \quad \text{for all } n \geq n_\eta.$$

Let $\delta > 0$ and $\beta \in \partial N_n(\delta, B)$. The Taylor expansion of l_n around B gives

$$l_n(\beta) - l_n(B) = (\beta - B)'s_n(B) - \frac{1}{2}(\beta - B)'F_n(\xi)(\beta - B), \tag{9.1}$$

with derivatives $s_n = \partial l_n/\partial\beta$ and $F_n = -\partial^2 l_n/(\partial\beta\partial\beta')$ and an intermediate point ξ which lies between β and B. By construction of $N_n(\delta, B)$,

$$\sqrt{n}\|\beta - B\| = \delta$$

and thus, the vector

$$v := \frac{\sqrt{n}}{\delta}(\beta - B)$$

fulfills $\|v\| = 1$. Substituting with v, the Taylor expansion (9.1) can be written as

$$l_n(\beta) - l_n(B) = \frac{\delta}{\sqrt{n}}v's_n(B) - \frac{1}{2}\delta^2 v'\frac{F_n(\xi)}{n}v,$$

where the two summands on the right hand side satisfy

$$v's_n(B) \leq \max_{\|z\|=1}|z's_n(B)| \leq \|s_n(B)\| \tag{9.2}$$

$$v'\frac{F_n(\xi)}{n}v \geq \min_{\|z\|=1} z'\frac{F_n(\xi)}{n}z. \tag{9.3}$$

The last inequality of (9.2) follows by the Cauchy-Schwarz inequality.

Lower bound for (9.3). Since $B \in N_n(\delta, B) \subset \mathcal{B}$ and $\beta \in \partial N_n(\delta, B)$, it directly follows from convexity of $\mathcal{B}$ that $\xi \in N_n(\delta, B)$. Now uniform convergence of the scaled Fisher information matrix, cf. Lemma 9.1, yields a lower bound for the expression

$$z' \frac{F_n(\xi)}{n} z, \quad \|z\| = 1.$$

Specifically, there exists for all $\epsilon > 0$ an $n_\epsilon \in \mathbb{N}$ which does not dependent on ξ such that for all $n \geq n_\epsilon$,

$$\left\| \frac{F_n(\xi)}{n} - F(\xi) \right\| \leq \epsilon, \quad \text{a.s.}$$

It follows for all $z \in \mathbb{R}^p$ with $\|z\| = 1$ that

$$\left| z' \frac{F_n(\xi)}{n} z - z' F(\xi) z \right| \leq \epsilon$$

and thus,

$$z' \frac{F_n(\xi)}{n} z \geq z' F(\xi) z - \epsilon.$$

By assumption the matrix $F(\xi)$ is positive definite for all ξ. Furthermore, as a uniform convergent limit of continuous functions F_n, F is also continuous. Boundedness of the domain therefore provides for sufficiently small ϵ and some $d > 0$ that

$$z' \frac{F_n(\xi)}{n} z \geq d > 0, \quad \text{for all } n \geq n_\epsilon.$$

Putting the pieces together. Altogether, we have

$$\begin{aligned}
&\mathbb{P}\left(\exists \beta \in \partial N_n(\delta, B) : l_n(\beta) - l_n(B) \geq 0\right) \\
&= \mathbb{P}\left(\exists \beta \in \partial N_n(\delta, B) : (\beta - B)' s_n(B) \geq \frac{1}{2} (\beta - B)' F_n(\xi) (\beta - B)\right) \\
&= \mathbb{P}\left(\exists \beta \in \partial N_n(\delta, B) : \frac{\delta}{\sqrt{n}} v' s_n(B) \geq \frac{1}{2} \delta^2 v' \frac{F_n(\xi)}{n} v\right) \\
&\leq \mathbb{P}\left(\frac{\delta}{\sqrt{n}} \|s_n(B)\| \geq \frac{1}{2} \delta^2 \min_{\|z\|=1} z' \frac{F_n(\xi)}{n} z\right) \\
&\leq \mathbb{P}\left(\|s_n(B)\| \geq \frac{1}{2} \sqrt{n} \delta d\right) \qquad (9.4)
\end{aligned}$$

for $n \geq n_\epsilon$. By Chebyshev's inequality, the last expression satisfies

$$\mathbb{P}\left(\|s_n(B)\| \geq \frac{1}{2} \sqrt{n} \delta d\right) \leq \frac{4}{n \delta^2 d^2} \mathbb{E}\left[\|s_n(B)\|^2\right], \qquad (9.5)$$

where the expectation linearly grows in n. More precisely,

$$\begin{aligned}
\mathbb{E}\left[\|s_n(B)\|^2\right] &= \sum_{k=1}^{p} \mathbb{E}\left[\left(\sum_{j=1}^{n} w_j X_{jk}\left(Y_j - \mu_j(B)\right)\right)^2\right] \\
&= \sum_{k=1}^{p} \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{j=1}^{n} w_j X_{jk}\left(Y_j - \mathbb{E}[Y_j \mid B]\right)\right)^2 \mid B\right]\right] \\
&= \sum_{k=1}^{p} \mathbb{E}\left[\sum_{j=1}^{n} w_j^2 X_{jk}^2 \mathrm{Var}(Y_j \mid B)\right] \\
&\leq pnV.
\end{aligned}$$

The covariate vectors and weights are bounded according to assumption (R3) and (R6). In addition, boundedness ensures that the conditional variances as continuous images $w_j b''(\xi_j + X_j B)$ are bounded. Hence, the above summands are bounded by some $V > 0$.

For a given $\eta > 0$, we can finally choose $n_\eta = n_\epsilon$ and

$$\delta := 2\sqrt{\frac{pV}{\eta d^2}}.$$

For these choices,

$$1 - \mathbb{P}(M_n^\delta) = \mathbb{P}\big(\exists \beta \in \partial N_n(\delta, B) : l_n(\beta) - l_n(B) \geq 0\big) \leq \eta$$

for all $n \geq n_\eta$, which shows the asymptotic occurrence of M_n.

For the second part of the theorem, we choose a null sequence (η_n) which converges to zero strictly slower than $1/n$. Then, since $\delta_n = \mathrm{const} \cdot \eta_n^{-1/2}$, $\delta_n/\sqrt{n}$ vanishes and the neighborhood shrinks to a singleton

$$N_n(\delta, B) \to \{B\}, \quad \text{a.s.}$$

as $n \to \infty$. This completes the proof.

Remark (q-variate exponential families). The proof can be easily extended from univariate to q-variate simple exponential families. We only need to verify the last step (9.5). Specifically, we have q-variate quantities X_{jk}, Y_j and $\mu_j(B)$ so that

$$\begin{aligned}
\mathbb{E}\left[\|s_n(B)\|^2\right] &= \sum_{k=1}^{p} \mathbb{E}\left[\left(\sum_{j=1}^{n} w_j X'_{jk}\left(Y_j - \mu_j(B)\right)\right)^2\right] \\
&= \sum_{k=1}^{p} \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{j=1}^{n} w_j X_{jk}\left(Y_j - \mathbb{E}[Y_j \mid B]\right)\right)^2 \mid B\right]\right] \\
&= \sum_{k=1}^{p} \mathbb{E}\left[\sum_{j=1}^{n} w_j X'_{jk} \Sigma_j(B) X_{jk}\right] \\
&\leq pnV.
\end{aligned}$$

We can find such a constant $V > 0$ since every covariate vector and every component of the conditional covariance matrix $\Sigma_j(B)$ is almost surely bounded.

Proof of Theorem 3.4. All proofs are based on (3.6) and Theorem 3.1, i.e.

$$\|\hat{\beta}_n - B\|\mathbb{1}_{M_n} \leq \frac{\delta_n}{\sqrt{n}}, \quad \text{a.s.}$$

with a vanishing upper bound and $\mathbb{P}(M_n) \to 1$ at the same time. Also recall that B is almost surely bounded by some constant $c_B > 0$.

ad i). Let $\epsilon > 0$. Then, $\tilde{\beta}_n$ is weakly consistent since

$$\begin{aligned}\mathbb{P}(\|\tilde{\beta}_n - B\| > \epsilon) &= \mathbb{P}(\|\hat{\beta}_n - B\| > \epsilon \mid M_n)\mathbb{P}(M_n) + \mathbb{P}(\|B\| > \epsilon \mid M_n^c)\mathbb{P}(M_n^c) \\ &\overset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

ad ii). Concerning asymptotic unbiasedness, we have

$$\begin{aligned}\|\mathbb{E}[\tilde{\beta}_n] - \mathbb{E}[B]\| &\leq \mathbb{E}[\|\tilde{\beta}_n - B\|] \\ &= \mathbb{E}[\|\hat{\beta}_n - B\|\mathbb{1}_{M_n}] + \mathbb{E}[\|B\|\mathbb{1}_{M_n^c}] \\ &\leq \frac{\delta_n}{\sqrt{n}}\mathbb{P}(M_n) + c_B\mathbb{P}(M_n^c) \\ &\overset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

ad iii). We claimed that

$$\mathrm{Cov}(\mathbb{E}[\tilde{\beta}_n \mid B]) \overset{n\to\infty}{\longrightarrow} \mathrm{Cov}(B).$$

In fact,

$$\begin{aligned}&\left\|\mathrm{Cov}(\mathbb{E}[\tilde{\beta}_n \mid B]) - \mathrm{Cov}(B)\right\| \\ =\ &\left\|\mathbb{E}\left[\left(\mathbb{E}[\tilde{\beta}_n \mid B] - B + B - \mathbb{E}[\tilde{\beta}_n]\right)\left(\mathbb{E}[\tilde{\beta}_n \mid B] - B + B - \mathbb{E}[\tilde{\beta}_n]\right)'\right] - \mathbb{E}[BB'] + \mathbb{E}[B]\mathbb{E}[B]'\right\| \\ \leq\ &\mathbb{E}\left[\left\|\mathbb{E}[\tilde{\beta}_n \mid B] - B\right\|^2\right] + 2\mathbb{E}\left[\left\|\mathbb{E}[\tilde{\beta}_n \mid B] - B\right\|\left\|B - \mathbb{E}[\tilde{\beta}_n]\right\|\right] \\ &\quad + \left\|\mathbb{E}\left[\left(B - \mathbb{E}[\tilde{\beta}_n]\right)\left(B - \mathbb{E}[\tilde{\beta}_n]\right)'\right] - \mathbb{E}[BB'] + \mathbb{E}[B]\mathbb{E}[B]'\right\| \\ =:\ &\mathrm{I} + 2\mathrm{II} + \mathrm{III},\end{aligned}$$

where all summands I to III turn out to be null sequences. Specifically, by claim ii),

$$\begin{aligned}\mathrm{III} &\leq \left\|\mathbb{E}[B]\left(\mathbb{E}[B] - \mathbb{E}[\tilde{\beta}_n]\right)'\right\| + \left\|\mathbb{E}[\tilde{\beta}_n]\left(\mathbb{E}[\tilde{\beta}_n] - \mathbb{E}[B]\right)'\right\| \\ &\leq \|\mathbb{E}[B]\|\|\mathbb{E}[B] - \mathbb{E}[\tilde{\beta}_n]\| + \|\mathbb{E}[\tilde{\beta}_n]\|\|\mathbb{E}[\tilde{\beta}_n] - \mathbb{E}[B]\| \to 0.\end{aligned}$$

Note that the $\mathcal{L}^1$-convergent sequence $(\tilde{\beta}_n)$ satisfies

$$\sup_n \|\mathbb{E}[\tilde{\beta}_n]\| \leq \sup_n \mathbb{E}[\|\tilde{\beta}_n\|] < \infty. \tag{9.6}$$

For the summand II,

$$\mathbb{E}\left[\left\|\mathbb{E}[\tilde{\beta}_n \mid B] - B\right\|\left\|B - \mathbb{E}[\tilde{\beta}_n]\right\|\right] \leq \frac{\delta_n}{\sqrt{n}}\mathbb{E}\big[\big\|B - \mathbb{E}[\tilde{\beta}_n]\big\|\mathbb{1}_{M_n}\big] + c_B\mathbb{E}\big[\big\|B - \mathbb{E}[\tilde{\beta}_n]\big\|\mathbb{1}_{M_n^c}\big].$$

By (9.6),

$$\mathbb{E}\big[\big\|B - \mathbb{E}[\tilde{\beta}_n]\big\|\mathbb{1}_{M_n}\big] \le c_B\mathbb{P}(M_n) + \sup_n \mathbb{E}[\|\tilde{\beta}_n\|]\mathbb{P}(M_n) < \infty,$$
$$\mathbb{E}\big[\big\|B - \mathbb{E}[\tilde{\beta}_n]\big\|\mathbb{1}_{M_n^c}\big] \le c_B\mathbb{P}(M_n^c) + \sup_n \mathbb{E}[\|\tilde{\beta}_n\|]\mathbb{P}(M_n^c) \to 0$$

so that II vanishes. At last, convergence of summand I easily follows as

$$\begin{aligned}\mathrm{I} &= \mathbb{E}\left[\left\|\mathbb{E}[\hat{\beta}_n \mid B] - B\right\|^2 (\mathbb{1}_{M_n} + \mathbb{1}_{M_n^c})\right] \\ &\le \frac{\delta_n^2}{n}\mathbb{P}(M_n) + c_B^2\mathbb{P}(M_n^c) \\ &\to 0.\end{aligned}$$

We similarly show the second part of the claim, which was

$$\mathrm{Cov}(\mathbb{E}[\tilde{\beta}_n \mid B], B) \to \mathrm{Cov}(B).$$

The proof works analogue to above, but we add $0 = -B + B$ only in the first factor of $\mathrm{Cov}(\mathbb{E}[\tilde{\beta}_n \mid B], B)$.

ad iv). Limiting behavior of the conditional covariance matrix is derived similarly to iii). Since

$$\begin{aligned}\|\,\mathrm{Cov}(\tilde{\beta}_n \mid B)\| &\le \mathbb{E}[\|\tilde{\beta}_n - \mathbb{E}[\tilde{\beta}_n \mid B]\|^2 \mid B] \\ &\le \mathbb{E}[\|\tilde{\beta}_n - B\|^2 \mid B] + 2\mathbb{E}[\|\tilde{\beta}_n - B\|\|B - \mathbb{E}[\tilde{\beta}_n \mid B]\| \mid B] + \mathbb{E}[\|B - \mathbb{E}[\tilde{\beta}_n \mid B]\|^2 \mid B] \\ &=: \mathrm{I} + 2\mathrm{II} + \mathrm{III},\end{aligned}$$

it suffices to prove that all three summands converge to zero in the proper senses. By using the monotonicity of the conditional expectation, we obtain

$$\begin{aligned}\mathrm{I} &= \mathbb{E}[\|\tilde{\beta}_n - B\|^2\mathbb{1}_{M_n} \mid B] + \mathbb{E}[\|\tilde{\beta}_n - B\|^2\mathbb{1}_{M_n^c} \mid B] \\ &\le \frac{\delta_n^2}{n}\mathbb{P}(M_n \mid B) + c_B^2\mathbb{P}(M_n^c \mid B).\end{aligned}$$

The last expression vanishes almost surely and in $\mathcal{L}^1$. We get the same upper bound for II and III as

$$\begin{aligned}\mathrm{II} &= \mathbb{E}[\|\tilde{\beta}_n - B\|\|B - \mathbb{E}[\tilde{\beta}_n \mid B]\|\mathbb{1}_{M_n} \mid B] + \mathbb{E}[\|\tilde{\beta}_n - B\|\|B - \mathbb{E}[\tilde{\beta}_n \mid B]\|\mathbb{1}_{M_n^c} \mid B] \\ &= \mathbb{E}[\|\tilde{\beta}_n - B\|\|\mathbb{E}[B - \tilde{\beta}_n \mid B]\|\mathbb{1}_{M_n} \mid B] + \mathbb{E}[\|\tilde{\beta}_n - B\|\|\mathbb{E}[B - \tilde{\beta}_n \mid B]\|\mathbb{1}_{M_n^c} \mid B] \\ &\le \frac{\delta_n^2}{n}\mathbb{P}(M_n \mid B) + c_B^2\mathbb{P}(M_n^c \mid B)\end{aligned}$$

and

$$\begin{aligned}\mathrm{III} &= \mathbb{E}[\|B - \mathbb{E}[\tilde{\beta}_n \mid B]\|^2\mathbb{1}_{M_n} \mid B] + \mathbb{E}[\|B - \mathbb{E}[\tilde{\beta}_n \mid B]\|^2\mathbb{1}_{M_n^c} \mid B] \\ &\le \frac{\delta_n^2}{n}\mathbb{P}(M_n \mid B) + c_B^2\mathbb{P}(M_n^c \mid B).\end{aligned}$$

ad v). Conditional on $B = \beta$, the relation

$$F_n^{T/2}(\hat{\beta}_n)(\hat{\beta}_n - B) \xrightarrow{d} \mathcal{N}(0, I)$$

holds under $\mathbb{P}_\beta$. It follows that $\hat{\beta}_n$ is also asymptotically Normal under the unconditional measure $\mathbb{P}$. In detail, let $Z \sim \mathcal{N}(0, I_p)$ and A be a Borel set in $\mathbb{R}^p$, then by the dominated convergence

theorem,

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left(F_n^{T/2}(\hat{\beta}_n)(\hat{\beta}_n - B)\in A\right) &= \int_{\mathcal{B}}\lim_{n\to\infty}\mathbb{P}_\beta\left(F_n^{T/2}(\hat{\beta}_n)(\hat{\beta}_n - B)\in A\right)\mathbb{P}(B\in d\beta)\\ &= \int_{\mathcal{B}}\mathbb{P}_\beta(Z\in A)\mathbb{P}(B\in d\beta)\\ &= \mathbb{P}(Z\in A).\end{aligned}$$

The asymptotic normality of the PMLE $\tilde{\beta}_n$ can be directly concluded. We have

$$F_n^{T/2}(\tilde{\beta}_n)(\tilde{\beta}_n - B) = F_n^{T/2}(\hat{\beta}_n)(\hat{\beta}_n - B)\mathbb{1}_{M_n} - F_n^{T/2}(0)B\mathbb{1}_{M_n^c}.$$

Since $\mathbb{1}_{M_n}$ converges in probability to 1 and $F_n^{T/2}(0)B\mathbb{1}_{M_n^c}$ converges in probability to 0, the claim follows by Slutsky's theorem, cf. Klenke (2006).

Acknowledgement

The authors thank the editor and the anonymous referees for useful comments which helped to improve the paper.

Both authors acknowledge financial support from the Deutsche Forschungsgemeinschaft (Research Training Group 1100 at the University of Ulm).

References

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. North-Holland Publishing Company Amsterdam.

Antonio, K. and J. Beirlant (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics 40*(1), 58–76.

Brazauskas, V., H. Dornheim, and P. Ratnam (2014). Credibility and regression modeling. *In: Predictive Modeling Applications in Actuarial Science, Vol. 1, Frees, Edward W., Richard A. Derrig, and Glenn Meyers (eds.)*.

Bühlmann, H. (1967). Experience rating and credibility. *Astin Bulletin 4*(03), 199–207.

Bühlmann, H. and A. Gisler (2005). *A course in credibility theory and its applications*. Springer.

Cairns, A. J., D. Blake, and K. Dowd (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance 73*(4), 687–718.

Currie, I. D., M. Durban, and P. H. Eilers (2004). Smoothing and forecasting mortality rates. *Statistical modelling 4*(4), 279–298.

De Vylder, F. (1985). Non-linear regression in credibility theory. *Insurance: mathematics and economics 4*(3), 163–172.

Efron, B. (1996). Empirical bayes methods for combining likelihoods. *Journal of the American Statistical Association 91*(434), 538–550.

Fahrmeir, L. and H. Kaufmann (1983). *Konsistenz und asymptotische Normalität des Maximum-Likelihood-Schätzers in verallgemeinerten linearen Modellen*. Regensburger Diskussionsbeiträge zur Wirtschaftswissenschaft. Univ., Fak. für Wirtschaftswiss.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics 13*(1), 342–368.

Hachemeister, C. A. (1975). Credibility for regression models with application to trend. In *Credibility, theory and applications, Proceedings of the berkeley Actuarial Research Conference on Credibility*, pp. 129–163.

Human-Mortality-Database (2014). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on December 29th, 2014).

Klenke, A. (2006). *Wahrscheinlichkeitstheorie*, Volume 1. Springer.

Lo, C. H., W. K. Fung, and Z. Y. Zhu (2007). Structural parameter estimation using generalized estimating equations for regression credibility models. *Astin Bulletin 37*(02), 323–343.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Nelder, J. and R. Verrall (1997). Credibility theory and generalized linear models. *Astin Bulletin 27*, 71–82.

Neuhaus, W. (1985). Choice of statistic in linear bayes estimation. *Scandinavian Actuarial Journal*, 1–26.

Norberg, R. (1980). Empirical bayes credibility. *Scandinavian Actuarial Journal 1980*(4), 177–194.

Norberg, R. (2004). *Credibility theory. In J. L. Teugels and B. Sundt, editors, Encyclopedia of Actuarial Science*. Wiley, Chichester, UK.

Ohlsson, E. (2008). Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal 2008*(4), 301–314.

Ohlsson, E. and B. Johansson (2006). Exact credibility and tweedie models. *Astin Bulletin 36*(1), 121.

Pitselis, G. (2004). De vylder's robust nonlinear regression credibility. *Beligan Actuarial Bulletin 1*(4), 44–49.

Pratt, J. W. (1959). On a general concept of" in probability". *The Annals of Mathematical Statistics*, 549–558.

Qian, W. (2000). An application of nonparametric regression estimation in credibility theory. *Insurance: Mathematics and Economics 27*(2), 169–176.

Taylor, G. C. (1977). Abstract credibility. *Scandinavian Actuarial Journal 1977*(3), 149–168.

Whitney, A. (1918). The theory of experience rating. *Proceedings of the Casualty Actuarial Society* (4), 274–292.

Witting, H. and G. Nölle (1970). *Angewandte mathematische Statistik: Optimale finite und asymptotische Verfahren*, Volume 14. Teubner.

Contact Details

Marcus C. Christiansen
Maxwell Institute for Mathematical Sciences, Edinburgh,
& Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh,
EH14 4AS, United Kingdom
e-mail: m.c.christiansen@hw.ac.uk

Edo Schinzinger (*corresponding author*)
Institute of Insurance Science
University of Ulm
89081 Ulm, Germany
e-mail: edo.schinzinger@uni-ulm.de