ulm university universität
uulm

Fakultät für Mathematik und Wirtschaftswissenschaften

Credibility estimation in insurance data: generalized linear models and evolutionary modeling

Dissertation

zur Erlangung des Doktorgrades Dr. rer. nat.
der Fakultät für Mathematik und Wirtschaftswissenschaften
der Universität Ulm

vorgelegt von

Edo Schinzinger

aus Kurume (Japan)

2015

Amtierender Dekan:	Prof. Dr. Werner Smolny
Erstgutachter:	Assoc. Prof. Dr. Marcus Christiansen
Zweitgutachter:	Prof. Dr. Ulrich Stadtmüller
Tag der Promotion:	22. Juni 2015

Abstract. Generalized linear models (GLM) have multiple applications, in particular they are a popular tool in insurance for fitting claims data. Insurance portfolios typically consist of heterogeneous clusters with similar but different risk characteristics. Problems arise when only limited statistical information is available for individual clusters. Credibility theory is a commonly used actuarial tool to improve statistical inference for small clusters, however credibility estimators have only been developed for a few specific models and a general theory remains lacking. In the present thesis we fill that gap, presenting a credibility estimator in a general GLM setting allowing all simple exponential families with natural link functions and cluster specific volume parameters. We study asymptotic properties of the estimator and illustrate our new concept with both a simulation study and an application to mortality data.

In the second part of the thesis we deal with an application of evolutionary credibility models to mortality data. Such a model correctly recognizes the random nature of the underlying time factor and further allows for the flexibility of time series modeling. The final model incorporates a smoothing procedure over time that ensures robustness over successive forecasts.

Zusammenfassung. Verallgemeinerte Lineare Modelle (GLM) sind ein beliebtes Werkzeug in der Regressionsanalyse und finden oft Verwendung in der Versicherungsmathematik. Das Portfolio eines Versicherungsunternehmens besteht typischerweise aus heterogenen Gruppen mit ähnlichen, aber dennoch unterschiedlichen Risikomerkmalen und versicherungstechnische Größen müssen für jede Gruppe einzeln geschätzt werden. Dabei treten Schwierigkeiten auf, falls bestimmte Gruppen nur wenig statistische Information enthalten. Die Credibility-Theorie wird in genau solchen Fällen eingesetzt um die Schätzung durch Berücksichtigung des gesamten Portfolios zu verbessern. Credibility-Ansätze bestehen jedoch nur für einige Spezialfälle des GLMs und ein allgemeingültiges Verfahren wurde bisher nicht entwickelt. In der vorliegenden Arbeit schließen wir diese Lücke, indem wir ein auf GLM angepasstes Credibility-Modell vorstellen. Wir erlauben alle natürlichen Exponentialfamilien mit kanonischer Linkfunktion und zudem die Berücksichtigung von gruppenspezifischen Mengenparametern. Die Vorteile des Modells werden anhand von Simulationstudien und Beispielen aus der Sterblichkeitsmodellierung verdeutlicht.

Im zweiten Teil der Arbeit entwickeln wir ein zeitdynamisches Credibility-Modell, das speziell auf Sterblichkeitsdaten zugeschnitten ist. Unser Modell behandelt die stochastischen Zeitfaktoren als solche und erlaubt ihre flexible Modellierung mit Techniken der Zeitreihenanalyse. Das finale Modell ermöglicht eine rekursive Aktualisierung der Prognosen anhand neuer Beobachtungen. Die auf diese Weise erzeugten Prognosen sind robust gegenüber aufeinanderfolgenden Beobachtungszeiträumen.

Contents

CHAPTER 1

Introduction

Credibility theory is a topic of Bayesian statistics that appears in many areas of actuarial science. Research in this field began with the pioneering work of Bichsel (1964), which was motivated from premium calculation in car insurance. Bichsel studied a method to incorporate individual claim statistics in the calculation to protect "good" risks from a forthcoming premium increase. Subsequently, credibility theory experienced a rapid growth and its methods are now established in many statistical models. New methods have been developed because of the practical necessity to respond to the continuously changing requirements of an actuary. Indeed, actuaries currently need to consider all types of underwriting risk that are categorized into risk of random fluctuations and risk of errors. In particular, the latter demands a proper reflection of the stochastic nature of the underlying problem. The main purpose of this thesis is to extend the actuarial toolbox by providing further credibility models.

Even with the passage of time, the fundamental idea to balance the individual risk experience with the collective risk experience still remains. While the former is more relevant in estimating future expectations, it is also more volatile than the latter. Uncertainty increases as less statistical information is contained in the individual cluster. Estimates obtained using the collective, on the other hand, may inadequately reflect the individual characteristics. The golden mean, which is the credibility estimator, is constructed by weighting both forms of information according to their *credibility*. From a more mathematical point of view, a portfolio of N random variables $(B_i, Y_i)_{i=1}^N$ is considered. Individual observations are assumed to be an outcome of the random vector $Y_i = (Y_{ij})_{j\in\mathbb{N}}$ whose distribution F_{β_i} is entirely or partly specified through the realization β_i of B_i. The great flexibility for the map $\beta_i \mapsto F_{\beta_i}$ allows a wide variety of models. We will, for instance, obtain the Bühlmann-Straub model, cf. Buhlmann and Straub (1972), when $Y_{ij} = B_i + \epsilon_{ij}$, and the linear regression model, cf. Hachemeister (1975), when $Y_{ij} = X_j B_i + \epsilon_{ij}$ with covariate vectors X_j and error terms ϵ_{ij}. More complex structures will be studied in the thesis. In all of the models, we seek a function $T(Y_1, \ldots, Y_N)$ that optimally estimates B_i with respect to the mean squared error. While classical Bayesian methods rely on some parametric families of conjugate prior distributions for the B_i, credibility follows a non-parametric approach established by Bühlmann (1967). In insurance applications it is untypical to impose much structure on the distributions; instead, structure is imposed on the class of admissible estimators. Indeed, credibility estimators are linear Bayes estimators, in that they are restricted to affine structures of the observation vectors. Seeking the optimal solution typically leads to the form

$$\hat{B}_i = \alpha_i \hat{\beta}_i + (1 - \alpha_i)\mathbb{E}[B_i]$$

of a weighted average of the best individual estimator $\hat{\beta}_i$ and best collective estimator $\mathbb{E}[B_i]$. Such structures, which may also be interpreted in a multidimensional context, are termed a credibility formula.

A credibility model unites the individual clusters, which have a certain structure in common, "under one roof". The clusters are a priori identical in the sense that their unobservable risk profiles B_i are identically distributed but variation in the portfolio arises once these variables are drawn. Many statistical models attempt to describe a collection of random variables, including a copula model that combines the individual clusters under a certain stochastic dependence structure. These models normally aim to correctly estimate the joint distribution of the entire portfolio, whereas the purpose of credibility models is fundamentally different. One is often interested only in certain clusters, where either the lack of individual information should be compensated or one wants to show that the information is sufficient on its own. Credibility theory provides a simple yet powerful way to make use of the portfolio structure in both cases.

1.1. Part 1

This thesis deals with two forms of credibility models each studied separately. Part 1 deals with credibility estimation for generalized linear models (GLM), which are a popular tool in insurance applications, among others. Poisson-GLMs, for example, are used for fitting counted data, including the claim frequency in a car insurance portfolio, cf. Ohlsson (2008), and the number of deaths in a life insurance portfolio, cf. Brillinger (1986). In both examples observed data are typically available for heterogeneous clusters referring to different countries, different groups of people and so forth, and insurance rates and actuarial reserves have to be calculated individually for each cluster. However, problems will arise when the clusters contain limited statistical information, for example, a small number of samples. Although a credibility approach would be the first choice, a proper framework for GLMs remains lacking in the literature.

Several papers studied similar types of models. Jewell (1974) demonstrated that for an exponential family of distributions that the credibility formula equals the exact Bayes estimator, meaning that the linear approximation and the exact formula agree when the conjugate prior distribution is assumed for the B_i. Generalized linear mixed models (GLMM), which were studied by Breslow and Clayton (1993), extends the classical GLM framework by random effects. While a GLM allows the first moment of the risks Y_{ij} to be related to a linear model of the form

$$g(\mathbb{E}[Y_{ij}]) = X_{ij}\beta_i, \quad j = 1, \ldots, n,$$

for some link function g, covariate vectors X_{ij} and parameter vectors β_i, a GLMM assumes that there is a joint fixed effect β but adds a Normal distributed random effects component B_i. More precisely,

$$g(\mathbb{E}[Y_{ij} \mid B_i]) = X_{ij}\beta + Z_{ij}B_i$$

with covariate vectors Z_{ij}. This model was extended by Lee and Nelder (1996) to a hierarchical GLM that allows the inclusion of non-Normal random effects $\nu(B_i)$ for some strictly monotonic function of B_i. Estimations for B_i are derived by maximizing the hierarchical likelihood under the assumption that the B_i are distributed according to the conjugate prior. The optimal solution then has the structure of a credibility formula. The first credibility type model that follows a semi-parametric approach can be found in Ohlsson and Johansson (2006) and Ohlsson (2008) for GLMs of Tweedie distributions. They consider a multiplicative model that in our terminology reads

$$\mathbb{E}[Y_{ij} \mid B_i] = g^{-1}(X_{ij}\beta)B_i,$$

such that B_i is considered as a multiplicative random effect with $\mathbb{E}[B_i] = 1$. Given the GLM parameter β, credibility estimators for the B_i are obtained by considering the scaled observations $\tilde{Y}_{ij} = Y_{ij}/g^{-1}(X_{ij}\beta)$. Although the Tweedie models cover practically relevant distributions like the Poisson and Gamma distributions, their model is in our opinion unsatisfactory. The random effect is scalar valued and equally affects all observations and covariates in the fit. However, it is more realistic to assume a GLMM-like structure with $X_{ij} = Z_{ij}$, where each component β_k of the p-variate GLM parameter includes its own random effect B_{ik}. Furthermore, from a theoretical point of view, the scaled model $\tilde{Y}_{ij}$ disregards the stochastic character of β, which is typically unknown and must be replaced by an estimator $\hat{\beta}$. Their credibility formula for B_i does not apply under a random scaling factor. De Vylder (1985) approaches the problem in a different way. The author considers a pure random effect model

$$\mathbb{E}[Y_{ij} \mid B_i] = g^{-1}(X_{ij}B_i)$$

and approximates the right hand side using a linear model $\tilde{x}_i + \tilde{X}_{ij}B_i$. Credibility estimators for the non-linear regression model are then derived using the linear structure of the approximating model. However, a convergence behavior towards the exact solution is not provided.

In this thesis we adopt the pure random effect model $g\left(\mathbb{E}[Y_{ij} \mid B_i]\right) = X_{ij}B_i$ but we consistently deal with the non-linear structure. Our model refrains from distributional assumptions on B_i but only demands mild regularity assumptions that do not restrict its applicability. A credibility formula is established for all simple exponential families with natural link functions and we further allow incorporation of volume parameters in terms of cluster specific weights and offset terms.

There are several papers dealing with random effect models that are not restricted to credibility type solutions but allow any kind of structure, see Fahrmeir et al. (1994) for an overview. Predictions for the random effects are typically performed in a two-step procedure. In a first step, the fixed parameters are estimated through maximizing the marginal likelihood, which is obtained by integrating over the random effects distribution. Because an analytic solution of the integral is only available in special cases, repeated numerical integration is required. Let us mention Gauss-Hermite quadrature, cf. Anderson and Aitkin (1985), Gibbs sampling, cf. Zeger and Karim (1991), Laplace approximation, cf. Breslow and Clayton (1993) and Markov chain Monte Carlo, cf. Gilks et al. (1996), to name some of the approximation methods. Santner and Duffy (1989) proposes an EM-type algorithm for indirect maximization of the marginal likelihood to avoid the numerical integration. Prediction of the random effects is then based on the posterior density, where parameters are replaced by their consistent estimates. The described methods are numerically intensive and their solutions are not readily traceable. The credibility formula, on the other hand, provides a closed form solution and each involved component has a practical interpretation.

Part 1 of the thesis is organized as follows. *Chapter 2* has a preparatory role and provides a brief introduction to credibility models and GLMs. Basic terminology and notation that are used throughout the work are established. We start with the multivariate and linear regression credibility models, because these provide a solid theoretical foundation on which we can develop our theory. Indeed, from a theoretical point of view, they are special cases of our credibility model below that is tailored for GLMs.

In *Chapter 3* the credibility approach for GLMs, abbreviated with CGLM, is established. We consider a simple setting, where the N clusters are assumed to be independent and identically distributed. The latter property will be relaxed in a later stage of analysis by allowing additional cluster specific parameters. The underlying model is semi-parametric with the prior distribution for the B_i having a compact support but not specified further. Following the structure of conventional credibility models, we seek an optimal solution within the affine class of the best individual estimators, namely the maximum likelihood estimators (MLE) for the GLM parameter with respect to the distribution of Y_i given B_i. The individual solutions, where the B_i are treated as constants, are then considered under the unconditional probability measure that takes the random nature of the B_i into account. Problems immediately arise because these estimators are not square integrable, meaning that mean squared error optimization is not possible or, to be precise, the mean squared error does not even exist. We overcome this problem by introducing what we term the pseudo maximum likelihood estimator (PMLE), which is a square integrable modification of the MLE. The PMLE restricts the MLE to a sequence of events where the behavior can be kept under control and we justify the procedure by showing that the probability of these events convergences to one. A detailed construction is given through Theorem 3.2 and Definition 3.3. Using these quantities instead, the credibility formula, Theorem 3.7, follows by solving the optimization problem. Because the PMLE is biased this formula does not have the typical structure of a weighted average and also contains structural parameters that are difficult to estimate. To solve these issues, we use the notion of asymptotic equivalence in probability, cf. Definition 3.8. Theorem 3.9 then provides an asymptotic equivalent credibility formula that is both easy to interpret and handle, and the involved structural parameters can all be consistently estimated as we prove in Theorem 3.12. The theory is completed with a simulation study that demonstrates the effectiveness of our approach. We compare the simulated mean squared error of the credibility estimator with that of the individual estimator. In accordance with our primary motivation, the improvement will be remarkable when clusters contain only a small number of observations, and when the sample size is large the clusters will be evaluated as being greatly credible.

Chapter 4 takes a closer look at the credibility model from a purely theoretical perspective. The special cases of a Poisson-CGLM and of grouped data are studied for possible impacts on the parameter estimation and the model assumptions. With the Poisson assumption, the PMLE and the MLE agree with high probability, and more precisely, the probability of the complementary event decays at an exponential rate instead of an inverse linear rate, cf. Theorem 4.1. Consequences on the estimation are discussed in Theorem 4.2. If data have a grouped structure, i.e. when there are only finitely many different covariate vectors, the assumption of the support of B_i being compact will reduce to mild integrability conditions for B_i, see Theorem 4.5.

Chapter 5 extends the credibility model with additional cluster specific volume parameters. These include offset terms in the linear predictor or weights, which may be the dispersion of a simple exponential family or a binary variable to describe missing observations. We allow these parameters to vary from cluster to cluster such that the portfolio consists of independent but no longer identically distributed pairs (B_i, Y_i). The revised model is more relevant for practical applications because clusters rarely have identical volumes. In the presented example of mortality data, for example, a proper model for the death counts must account for the different population sizes of each country, age

group and calendar year. Theorem 5.7 establishes the (asymptotic) credibility formula and thus makes a credibility estimation for such cases possible. The different volumes associated with the observed data are reflected in the formula in the form of cluster specific credibility weights. For the calculation, the structural parameters need to be estimated and these are derived in Theorem 5.9. The chapter ends with simulation studies that present the influence of different volume parameters. As one would intuitively expect, clusters with small volumes, for example a small number of samples or a high dispersion, benefit most from credibility estimation.

Finally, *Chapter 6* concludes with an application to real-world mortality data. Stochastic mortality modeling is a very topical subject and the focus in research is changing from single-country to multi-country models. However, many papers investigate the establishment of a joint model framework for the whole portfolio rather than using the data to improve the estimation for single countries, see, for example, Li and Lee (2005) and Hatzopoulos and Haberman (2013). A CGLM provides a natural way to extend GLM-based single-country models to a multi-country framework and we illustrate this procedure by means of the Cairns-Blake-Dowd model, cf. Cairns et al. (2006), which decomposes the death rates into age and time factors. The fit includes 36 European and first world countries and credibility estimators are calculated for their individual time factors. In accordance with our expectations, the benefits of the credibility approach are greatest for Iceland and Luxembourg, which are the countries with the smallest populations. The paths of the estimated time factors are smoother compared to those of the individual estimators and in fact, the credibility formula will smooth out humps that are incredible when compared to other calendar years and countries. This effect will become more apparent when we fit the model to the very advanced ages of 90 to 100 years. These ages are usually excluded from fitting because the data is very volatile due to the small exposure-to-risk. Most of the countries in the fit, for example France, Germany, Japan, UK and USA among others, are evaluated as being highly credible such that the credibility and individual estimators almost agree. Therefore, from a credibility point of view, a multi-country model is redundant for these countries.

1.2. Part 2

The second part of the thesis deals with credibility models of an evolutionary type. If we interpret i as a time index, then $(B_i, Y_i)_{i=1}^N$ will belong to a single cluster observed in successive periods $i = 1, \ldots, N$ and credibility methods can be used to obtain best predictions for future values. These predictions are successively updated based on recent observations that are weighted according to their credibility. In the thesis we investigate an evolutionary credibility framework for mortality modeling.

Mortality forecasts are used in a wide range of fields, including in making health policy, pharmaceutical research, social security, retirement fund planning and life insurance, to name just a few. In most countries, governmental agencies regularly publish mortality projections. In Belgium, for example, the Federal Planning Bureau now produces projected life tables on an annual basis, based on the most recent observations. However, standard forecasting approaches do not incorporate any smoothing procedures over time and this may cause some instability from one forecast to another. This is due to the model being entirely re-fitted based on an extended data set and that no association is made between the successive projections.

Following the elegant approach to mortality forecasting pioneered by Lee and Carter (1992) many projection models decompose the death rates (on a logarithmic scale) or the 1-year death probabilities (on a logit scale) into a linear combination of a limited number of time factors. See, for example, Hunt and Blake (2014). In a first step, regression techniques are used to extract the time factors from the available mortality data. In a second step, the time factors are intrinsically viewed as forming a time series to be projected to the future. The actual age-specific death rates are then derived from this forecast using the estimated age effects. This in turn yields projected life expectancies.

In the first step of the two-step model calibration procedure, the random nature of the unobservable time factor is disregarded, and this may bias the analysis. Because possible incoherence may arise from this two-step procedure, Czado et al. (2005) integrated both steps into a Bayesian version of the model developed by Lee and Carter (1992) to avoid this deficiency. After Czado et al. (2005), Pedroza (2006) formulated the Lee-Carter method as a state-space model, using Gaussian error terms and a random walk with drift for the mortality index. See also Girosi and King (2008), Kogure et al. (2009), Kogure and Kurachi (2010) and Li (2014) for related works. However, the practical implementation of Bayesian methods often requires computer-intensive Markov Chain Monte Carlo (MCMC) simulations. This is why we propose in this thesis a simple credibility model ensuring robustness over time while keeping the computational issues relatively easy and allowing for the flexibility of time series modeling. It is worth stressing that the time factor is treated here as such, and not as a parameter to be estimated from past mortality statistics using regression techniques before entering time series models. In this way, we recognize the hidden nature of the time factor and its intrinsic randomness. The credibility Cairns-Blake-Dowd model, which will be introduced in Chapter 6, can be also classified into this category of mortality models. It takes into account the stochastic character of the time factors but specific time dynamics are nevertheless not incorporated.

Whereas most mortality studies consider both genders separately, the model that we propose combines male and female mortality statistics. While ensuring model identification, this is particularly useful in practical applications when both genders are usually involved. In insurance applications, for example, separate analyses could lead to this strong dependence pattern being missed, which considerably reduces possible diversification effects between male and female policyholders within the portfolio. In demographic projections, combining male and female data is necessary to ensure consistency in a gender-specific mortality forecast. This problem has been considered by several authors in the literature. Let us mention Carter and Lee (1992) who fitted the Lee and Carter (1992) model to male and female populations separately and then determined the dependence between the two gender-specific time factors. These authors considered three models for the pair of time factors: a bivariate random walk with drift, a single time factor common to both genders and a co-integrated process where the male index follows a random walk with drift and there exists a stationary linear combination of both time factors. More recently, Yang and Wang (2013) assumed that the time factors followed a vector error correction model. See also Zhou et al. (2013). Other models incorporate a common factor for the combined population as a whole, as well as additional factors for each sub-population. The common factor describes the main long-term trend in mortality change while the additional factors depict the short-term discrepancy from the main trend within each sub-population. Li and Lee (2005) proposed applying the augmented

common factor model generalized by Li (2013) to several factors. The model structures proposed in Delwarde et al. (2006), by Debón et al. (2011), and by Russolillo et al. (2011) only include a single, common time factor. As argued in Carter and Lee (1992), this simple arrangement may enforce greater consistency and is a parsimonious way to model both populations jointly. However, it also implies that the death rates of the two populations are perfectly associated, an assumption with far-reaching consequences in risk management.

The thesis is innovative in that a new multi-population mortality projection model is proposed, based on mortality improvement rates instead of levels. Recently, several authors suggested targeting improvement rates to forecast future mortality instead of the death rates. While the time dependence structure of death rate models is dominated by the continuing downward trend, the improvement rates are already trend adjusted, cf. Aleksic and Börger (2011) or Mitchell et al. (2013). Furthermore, the model is fitted properly, recognizing the hidden nature of time factors which are not treated as unknown parameters to be estimated from the mortality data. Mortality projections are derived by means of the predictive distribution of the time index, i.e. it is a posterior distribution given past observations. This is the credibility feature of the proposed approach. New data feed this predictive distribution as they become available and so help to update mortality projections. This recognizes the dynamic aspect of mortality forecasting and avoids refitting the entire model based on new data. To the best of our knowledge, this dynamic updating approach has not been used as yet and our numerical illustrations demonstrate its advantages compared to classical frequentist approaches.

Part 2 of the thesis is organized as follows. Chapter 7 provides a brief introduction into evolutionary credibility models and establishes such a model for the mortality improvement rates. Considering the age-aggregate mortality improvement rates allows us to study the dynamics of the time factor in isolation from the detailed age structure. This results in a state-space model, where the state variable, i.e. the time factor, follows an autoregressive moving average process. However, this model possesses identifiability issues regarding the innovation variances of the observation and state processes. We consider a gender-combined model, which not only provides gender-consistent forecasts but also ensures identifiability of the covariance structure. By introducing a gender correlation parameter, we obtain a variance-covariance structure, cf. Lemma 7.3, that ensures identifiability, cf. Theorem 7.4. The remainder of the chapter is devoted to numerical illustrations based on Belgian data. The optimal model is selected and fitted to the observed general population mortality experience with regard to different degrees of homogeneity between the genders. Chapter 8 derives the predictive distribution, describes mortality forecasting obtained from that distribution, and discusses the numerical results obtained from the Belgian population. We conclude with comparisons with the Lee-Carter and the official mortality forecasts to demonstrate the advantage of our approach in the sense of its robustness.

Part 1

Credibility for Generalized Linear Models

CHAPTER 2

Preliminaries

2.1. Credibility Theory

The aim of this section is to give a brief overview of credibility theory. We will focus on the essential points which are necessary for establishing credibility theory for generalized linear models. The main reference for this section is Bühlmann and Gisler (2005). However, our notation will slightly differ from their convention to make it consistent with the upcoming chapters.

The multidimensional credibility model. To be considered is a portfolio of $N \in \mathbb{N}$ clusters numbered $i = 1, \dots, N$. The observation vector of cluster i is denoted by $Y_i = (Y_{ij})_{j=1}^n$ and its risk profile by B_i. Each Y_{ij} and B_i are p-variate random vectors on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and fulfill the following model assumptions.

(i) Conditionally, given B_i, the Y_{ij}, $j = 1, \dots, n$, are independent and identically distributed (iid) with conditional moments

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

$$\mathrm{Cov}(Y_{ij} \mid B_i) = \Sigma(B_i). \tag{2.2}$$

In (2.2), $\Sigma : \mathbb{R}^p \to \mathbb{R}^{p,p}$ stands for the covariance function.

(ii) The pairs $(B_1, Y_1), \dots, (B_N, Y_N)$ are iid.

Assumption (ii) ensures that the portfolio is homogeneous. However, once the B_i are drawn, the clusters become heterogeneous through the individual stochastic components B_i and $\Sigma(B_i)$. The aim is to find estimators $\hat{B}_i$ for the conditional expectations or, in an actuarial context, the fair premiums B_i. If B_i is a fixed but unknown constant, (2.1) and (2.2) will allow empirical estimation for B_i. This individual solution is then extended to a credibility solution where the random nature of B_i is taken into account. In the following, we will denote the entireties $\mathbf{B} = (B_1, \dots, B_N)$ and $\mathbf{Y} = (Y_1, \dots, Y_N)$ by bold symbols. Let

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

be the class of estimators which are affine functions of all available observations. The classes $L(1)$ and $L(\mathbf{Y})$ are similarly defined by taking the corresponding component in $L(1, \mathbf{Y})$. As we have already motivated, imposing structure on the class of admissible estimators is typical in credibility theory and the restriction compensates the lack of distributional assumptions for the B_i.

Notice that (2.2) of assumption (i) implicitly requires that all observations Y_{ij} are square-integrable, i.e. $Y_{ij} \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Thus, $L(1, \mathbf{Y})$ is a closed subspace of

$\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ or in short $\mathcal{L}^2$. Recall that $\mathcal{L}^2$ is a Hilbert space equipped with the dot product

$$\langle X_1, X_2 \rangle = \mathbb{E}[X_1' X_2], \quad X_1, X_2 \in \mathcal{L}^2,$$

where X_1' denotes the transpose of X_1.

Definition 2.1. The orthogonal projection

$$\hat{B}_i = \text{Pro}(B_i \mid L(1, \mathbf{Y})) \tag{2.3}$$

is called the *credibility estimator* for B_i.

Identity (2.3) is equivalent to

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\mathbf{Y})}{\arg\min} \left\| \tilde{B}_i - B_i \right\|_{\mathcal{L}^2}^2$$

and by using the induced norm

$$\|X\|_{\mathcal{L}^2} = \sqrt{\mathbb{E}[X'X]}, \quad X \in \mathcal{L}^2,$$

we get a more probabilistic interpretation for the credibility estimator. In fact, $\hat{B}_i$ is the minimizer of the expected quadratic loss function, which is also called the mean squared error, within the class $L(1, \mathbf{Y})$, i.e.

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\mathbf{Y})}{\arg\min} \; \mathbb{E}\left[(\tilde{B}_i - B_i)'(\tilde{B}_i - B_i) \right]. \tag{2.4}$$

Solving (2.4) with respect to the parameters a and A_{ij} provides an explicit solution for the credibility estimator. See Jewell (1973) for details. Let

$$\bar{Y}_i = \frac{1}{n} \sum_{j=1}^{n} Y_{ij},$$

$$\beta_0 = \mathbb{E}[B_i], \tag{2.5}$$

$$S = \mathbb{E}[\Sigma(B_i)] \tag{2.6}$$

$$\text{and} \quad T = \text{Cov}(B_i). \tag{2.7}$$

Theorem 2.2 (Jewell, 1973). *The multidimensional credibility estimator for cluster i is given by*

$$\hat{B}_i = A\bar{Y}_i + (I - A)\beta_0 \tag{2.8}$$

with credibility matrix

$$A = T \left(T + \frac{1}{n} S \right)^{-1}.$$

Since $\mathbb{E}[\hat{B}_i] = \mathbb{E}[B_i]$, the credibility estimator is the best linear unbiased estimator. Structure (2.8) also has a nice interpretation. The credibility estimator is a weighted average of the individual observed average $\bar{Y}_i$ and the overall expected value β_0. The credibility matrix A assigns weights according to the variation within and between clusters. By (2.6) and (2.7), we have that

$$\frac{1}{n} S = \frac{1}{n} \mathbb{E}[(Y_{ij} - B_i)(Y_{ij} - B_i)'] = \mathbb{E}[(\bar{Y}_i - B_i)(\bar{Y}_i - B_i)'],$$

$$T = \mathbb{E}[(\beta_0 - B_i)(\beta_0 - B_i)']$$

and they are the accuracies of $\bar{Y}_i$ and β_0 respectively. Hence, the weight of $\bar{Y}_i$ (resp. β_0) is the accuracy of the opposite term in relation to the total accuracy

$$T + \frac{1}{n}S = \mathrm{Cov}\left(\mathbb{E}[\bar{Y}_i \mid B_i]\right) + \mathbb{E}\left[\mathrm{Cov}(\bar{Y}_i \mid B_i)\right] = \mathrm{Cov}(\bar{Y}_i).$$

The univariate setting allows a simpler interpretation in which case the credibility weights read

$$A = \frac{\mathrm{Var}(B_i)}{\mathrm{Var}(\bar{Y}_i)} = \frac{\mathrm{Var}(B_i)}{\mathrm{Var}(B_i) + \mathbb{E}[\mathrm{Var}(\bar{Y}_i \mid B_i)]},$$
$$1 - A = \frac{\mathbb{E}[\mathrm{Var}(\bar{Y}_i \mid B_i)]}{\mathrm{Var}(B_i) + \mathbb{E}[\mathrm{Var}(\bar{Y}_i \mid B_i)]}.$$

Furthermore, it is easy to check that

$$\bar{Y}_i = \underset{\tilde{B}_i \in L(\mathbf{Y})}{\arg\min}\, \mathbb{E}\left[(\tilde{B}_i - B_i)'(\tilde{B}_i - B_i) \mid B_i\right] \quad \text{a.s.}$$
$$\text{and} \quad \beta_0 = \underset{\tilde{B}_i \in L(1)}{\arg\min}\, \mathbb{E}\left[(\tilde{B}_i - B_i)'(\tilde{B}_i - B_i)\right],$$

i.e. $\bar{Y}_i$ is the best estimator based on the individual observations without any prior knowledge and β_0 is in contrast the best estimator based only on the prior knowledge. The sample mean $\bar{Y}_i$ is sometimes also referred to as the best linear and individually unbiased estimator of B_i. One should also notice that the structure of the class $L(1, \mathbf{Y})$ allows $\hat{B}_i$ to depend on all observations of the collective but only data provided by its own cluster appear in formula (2.8). Other clusters nevertheless become indispensable when it comes to the estimation of the structural parameters β_0, S and T, e.g.

$$\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_i.$$

Equally important is the fact the credibility estimator (2.8) fulfills the identity

$$\hat{B}_i = \mathrm{Pro}(B_i \mid L(1, \mathbf{Y})) = \mathrm{Pro}(B_i \mid L(1, \bar{\mathbf{Y}})), \tag{2.9}$$

where

$$L(1, \bar{\mathbf{Y}}) = \left\{ a + \sum_{i=1}^{N} A_i \bar{Y}_i : a \in \mathbb{R}^p, A_i \in \mathbb{R}^{p,p} \right\}.$$

Hence, one can compress the class of admissible estimators without losing relevant information.

Credibility theory is strongly related to Bayesian statistics. Specifically, the Bayes estimator for B_i is, similar to (2.4), the minimum mean squared error estimator but within $\mathcal{L}^2$ instead of $L(1, \mathbf{Y})$. Thus, the credibility estimator is a linear Bayes estimator. The Bayes estimator has the particular form

$$\hat{B}_i^{(\mathrm{Bayes})} = \mathbb{E}[B_i \mid \mathbf{Y}] \tag{2.10}$$

so that both estimators will agree if the conditional expectation is linear in $\mathbf{Y}$. By construction,

$$\mathbb{E}\left[(\hat{B}_i^{(\mathrm{Bayes})} - B_i)'(\hat{B}_i^{(\mathrm{Bayes})} - B_i)\right] \leq \mathbb{E}\left[(\hat{B}_i - B_i)'(\hat{B}_i - B_i)\right]$$

and the purpose of the credibility estimator may be doubted. However, the Bayes estimator requires specification of the posterior distribution of B_i given $\mathbf{Y}$. This can

be determined for particular combinations of the prior and conditional distributions and in such cases,

$$\hat{B}_i^{(\text{Bayes})}(y) = \int_{\mathbb{R}^p} \beta \mathbb{P}(B_i \in d\beta \mid \mathbf{Y} = y).$$

An explicit solution of the integral is generally not given and numerically intensive integration methods have to be used. In contrast, the credibility formula (2.8) has a closed form that is easy to evaluate and its estimation will even work if none of the involved distributions are specified. Thus, credibility allows nonparametric modeling. The generalization to a broader class of distributions is bought at the cost of a possibly increased mean squared error. One should not also neglect the interpretability, which is highly appreciated by practitioners.

The regression credibility model. A portfolio of N clusters with observation vectors $Y_i = (Y_{i1}, \dots, Y_{in})$ and risk profiles B_i, $i = 1, \dots, N$, is given. In the regression credibility model, the Y_{ij} are univariate random variables that conditional on B_i satisfy a regression equation. Provided a known design matrix $X \in \mathbb{R}^{n,p}$ 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

$$\mathbb{E}[Y_{ij} \mid B_i] = X_j B_i, \tag{2.11}$$

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

$$\operatorname{Cov}(Y_i \mid B_i) = \Sigma(B_i). \tag{2.12}$$

(ii) The pairs $(B_1, Y_1), \dots, (B_N, Y_N)$ are iid.

Estimating the conditional expectations $\mathbb{E}[Y_{ij} \mid B_i]$ requires estimating the B_i, which are in terms of (2.11) random p-variate regression parameters. Analogously to the multidimensional credibility model, the credibility estimator $\hat{B}_i$ for B_i is defined as the orthogonal projection of B_i on $L(1, \mathbf{Y})$, where

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

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

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

The class $L(1, \mathbf{Y})$ especially includes the best individual solutions. Considering B_i as a fixed but unknown constant β_i, cluster i follows a linear regression model, i.e. the response vector Y_i can be written as

$$Y_i = X\beta_i + \epsilon_i,$$

where ϵ_i is an error term satisfying

$$\mathbb{E}_{\beta_i}[\epsilon_i] = 0 \quad \text{and} \quad \operatorname{Cov}_{\beta_i}(\epsilon_i) = \Sigma(\beta_i).$$

The moments are taken with respect to the conditional probability measure denoted by $\mathbb{P}_{\beta_i}$. A main result in linear regression analysis, e.g. Seber and Lee (2012), states that

$$\hat{\beta}_i = (X'\Sigma(\beta_i)X)^{-1} X'\Sigma(\beta_i) Y_i \tag{2.13}$$

is the minimum mean squared error estimator for β_i under $\mathbb{P}_{\beta_i}$. If the error terms are conditionally Normal distributed, $\hat{\beta}_i$ is also the maximizer of the conditional likelihood function of Y_i given B_i. When treating B_i as a random variable, expression (2.13) can be generally written as

$$\hat{\beta}_i = (X'\Sigma(B_i)X)^{-1}X'\Sigma(B_i)Y_i \tag{2.14}$$

and it is the best individual solution in the sense that

$$\hat{\beta}_i = \underset{\tilde{\beta}_i \in L(\mathbf{Y})}{\arg\min}\, \mathbb{E}\left[(\tilde{\beta}_i - B_i)'(\tilde{\beta}_i - B_i) \mid B_i\right] \quad \text{a.s.}$$

Its expected conditional covariance matrix or, in other words, its accuracy is given by

$$\mathbb{E}\left[\mathrm{Cov}(\hat{\beta}_i \mid B_i)\right] = \left(X'\mathbb{E}[\Sigma(B_i)]^{-1}X\right)^{-1}.$$

The credibility solution is then constructed within $L(1, \mathbf{Y})$ that contains $\hat{\beta}_i$ and optimization accounts for the stochastic character of the B_i. The credibility formula now follows. Let

$$\begin{aligned} \beta_0 &= \mathbb{E}[B_i], \\ S &= \mathbb{E}[\Sigma(B_i)], \\ T &= \mathrm{Cov}(B_i), \end{aligned}$$

which are the structural parameters of the model.

Theorem 2.3 (Hachemeister, 1975). *The regression credibility estimator for cluster i is given by*

$$\hat{B}_i = A\hat{\beta}_i + (I - A)\beta_0 \tag{2.15}$$

with credibility matrix

$$A = T\left(T + \left(X'S^{-1}X\right)^{-1}\right)^{-1}.$$

Formula (2.15) is very similar to that of the multidimensional credibility model (2.8). The credibility estimator is composed of the best individual estimator $\hat{\beta}_i$ and the best prior estimator β_0 both weighted according to their credibility. As the best individual solution $\hat{\beta}_i$ is linear in $\mathbf{Y}$, the credibility estimator also satisfies

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\boldsymbol{\beta})}{\arg\min}\, \mathbb{E}\left[(\tilde{B}_i - B_i)'(\tilde{B}_i - B_i)\right], \tag{2.16}$$

where

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

This structure will motivate the credibility estimator for generalized linear models in a later stage of analysis.

distribution	θ	$b(\theta)$	$\mathbb{E}_\theta[Y] = b'(\theta)$	$\mathrm{Var}_\theta(Y) = b''(\theta)$	natural link
$\mathrm{Nor}(\mu, 1)$	μ	$\frac{\theta^2}{2}$	$\mu = \theta$	1	identity
$\mathrm{Exp}(\lambda)$	$-\lambda$	$-\log(-\theta)$	$\frac{1}{\lambda} = -\frac{1}{\theta}$	$\frac{1}{\lambda^2}$	inverse
$\mathrm{Poi}(\lambda)$	$\log\lambda$	$\exp(\theta)$	$\lambda = \exp(\theta)$	λ	log
$\mathrm{Ber}(p)$	$\log\left(\frac{p}{1-p}\right)$	$\log(1+\exp(\theta))$	$p = \frac{\exp(\theta)}{1+\exp(\theta)}$	$p(1-p)$	logit

TABLE 2.1. Common simple exponential families of type (2.17).

2.2. Exponential family and GLM

Following Fahrmeir and Kaufmann (1985) and Fahrmeir et al. (1994), we give a short overview about exponential families and generalized linear models. The exponential family is a class of probability distributions which have a certain form in common. A q-variate random variable Y is said to follow a distribution of an exponential family if its probability density function (pdf) has the structure

$$f_\theta(y) = c(y)\exp\left(\eta(\theta)'t(y) - b(\theta)\right).$$

Furthermore, one speaks of a simple or natural exponential family with natural parameter $\theta \in \Theta$ if

$$f_\theta(y) = c(y)\exp\left(\theta' y - b(\theta)\right), \tag{2.17}$$

where c is a non-negative measurable function. The set $\Theta \subset \mathbb{R}^q$ is called the natural parameter space and it contains all θ satisfying

$$0 < \int c(y)\exp\left(\theta' y\right) dy < \infty.$$

The function f_θ is therefore a pdf for all $\theta \in \Theta$ and $b(\theta)$ defines the normalization factor. Moreover, the set Θ is convex and all derivatives of $b : \Theta \to \mathbb{R}$ and all moments of Y exist in its interior Θ^0. In particular, we have

$$\mathbb{E}_\theta[Y] = \frac{\partial b(\theta)}{\partial\theta} = \mu(\theta), \tag{2.18}$$

$$\mathrm{Cov}_\theta(Y) = \frac{\partial^2 b(\theta)}{\partial\theta\partial\theta'} = \Sigma(\theta) \tag{2.19}$$

with μ and Σ being called the mean and covariance functions respectively. The natural exponential family includes many of the common distributions. An overview is presented in Table 2.1. One may also include an additional nuisance parameter in (2.17), which will be introduced in Chapter 5.

Assume that observations are sampled from independent random variables $(Y_j)_{j\in\mathbb{N}}$ that belong to the same simple exponential family with parameters $(\theta_j)_{j\in\mathbb{N}}$. Generalized linear models (GLM), first introduced by Nelder and Wedderburn (1972), are characterized by the following structure. Known covariates $X_j \in \mathbb{R}^{1,p}$ describe the distribution of the Y_j through a linear predictor

$$\eta_j = X_j\beta,$$

where $\beta \in \mathcal{B} \subset \mathbb{R}^p$ is called the GLM parameter out of an admissible set $\mathcal{B}$. The linear predictor η_j itself is related to the mean $\mu_j = \mu(\theta_j)$ by a link function g. More precisely,

$$g(\mu_j) = \eta_j = X_j\beta. \tag{2.20}$$

This identity makes regression practicable even in cases where μ_j is constrained, e.g. $\mu_j \geq 0$ for Poisson and $\mu_j \in [0,1]$ for Bernoulli variables. Of special interest are so-called natural or canonical link functions $g = \mu^{-1} = (b')^{-1}$. Then,

$$g(\mu_j) = \theta_j = X_j\beta \tag{2.21}$$

so that the linear predictor is directly connected to the natural parameter of the distribution. In the sequel g will always be the natural link function if not otherwise stated. It is also common to use the matrix notation

$$\theta = X\beta$$

with design matrix $X = (X_1', X_2', \dots)'$. There is no consensus in literature concerning the notion of covariates. While we call X_j a covariate vector, the term "design vector" is also common. Design vectors are functions of covariates, for instance covariates a and b may be coded as

$$X_j = \begin{pmatrix} 1 & a_j & b_j \end{pmatrix}$$

to include a base effect which equally affects all $j \in \mathbb{N}$.

Estimation of the parameter β is based on likelihood methods and the following conditions are assumed to hold.

(L1) The admissible parameter space $\mathcal{B}$ is open and convex.
(L2) g is twice continuously differentiable with non-singular Jacobian.
(L3) $X_j\beta \in \Theta^0$ for all $\beta \in \mathcal{B}$ and $j \in \mathbb{N}$.
(L4) $\sum_{j=1}^n X_j' X_j$ has full rank p for sufficiently large n.

Considering (2.17), the log-likelihood function of sample $(Y_1, \dots, Y_n)$ is, up to a constant which does not depend on β, given by

$$l_n(\beta) = \sum_{j=1}^n \theta_j' Y_j - b(\theta_j). \tag{2.22}$$

In view of (2.21), the mean and covariance functions (2.18) and (2.19) can be expressed in terms of β and we write $\mu_j(\beta) = \mu(\theta_j)$ and $\Sigma_j(\beta) = \Sigma(\theta_j)$. Further quantities of importance are the score function s_n and the observed Fisher information matrix F_n which are

$$s_n(\beta) = \frac{\partial l_n(\beta)}{\partial \beta} = \sum_{j=1}^n X_j' \left(Y_j - \mu_j(\beta)\right), \tag{2.23}$$

$$F_n(\beta) = -\frac{\partial^2 l_n(\beta)}{\partial \beta \partial \beta'} = \sum_{j=1}^n X_j' \Sigma_j(\beta) X_j. \tag{2.24}$$

As $\mathbb{E}[F_n(\beta)] = F_n(\beta)$ the expected and observed Fisher information matrices agree. Conditions (L2) and (L4) imply that $F_n(\beta)$ is positive definite for large n so that the log-likelihood function is concave. Therefore, convexity of $\mathcal{B}$, cf. condition (L1), provides uniqueness of the maximum if it exists. In what follows, we will denote $F_n(\beta)$ simply as the Fisher information matrix without any prefixes. For non-natural link functions, $F_n(\beta)$ involves stochastic components and its positive definiteness is not guaranteed. Finally, the maximum likelihood estimator (MLE) $\hat{\beta}$ for β is given as the solution of

$$s_n(\hat{\beta}) = 0$$

and it uniquely maximizes l_n. Asymptotic properties of $\hat{\beta}$ require further assumptions on F_n, which will be discussed in Section 3.2. At this point, the reader should just keep in mind that the MLE is the best estimator as it is asymptotically efficient. Its asymptotic covariance matrix equals the inverse Fisher information matrix as sample size n increases.

2.3. Remarks on notation

The transpose of a matrix A is denoted by A'. In connection with matrix square roots defined through the identity $A = A^{1/2}A^{T/2}$, the transpose is addressed by a superscript T. The inverse matrices are written as $A^{-1} = A^{-T/2}A^{-1/2}$. The norm $\|\cdot\|$ without a subscript will always denote the (induced) 2-norm if not explicitly mentioned. Recall that

$$\begin{aligned} \|v\| &= \sqrt{v'v}, \quad v \in \mathbb{R}^p, \\ \|A\| &= \max_{\|v\|=1} \|Av\| = \sqrt{\lambda_{\max}(A'A)}, \quad A \in \mathbb{R}^{p,p}. \end{aligned}$$

This matrix norm is submultiplicative and consistent, i.e.

$$\begin{aligned} \|A_1A_2\| &\leq \|A_1\|\|A_2\| \\ \text{and} \quad \|A_1v\| &\leq \|A_1\|\|v\| \end{aligned}$$

for square matrices A_1, A_2 and column vector v respectively. If applied to a random variable Y, $\|Y\|$ will denote the ω-wise evaluation of the norm and $\|Y\|_{\mathcal{L}^2} = \sqrt{\mathbb{E}[\|Y\|^2]}$ will denote the norm in the Hilbert space $\mathcal{L}^2$. Notice that the former is again a random variable but the latter is deterministic.

CHAPTER 3

Credibility estimator for Generalized Linear Models

3.1. Model assumptions

We consider a portfolio of $N \in \mathbb{N}$ clusters (B_i, Y_i), $i = 1, \dots, N$, on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Each cluster consists of a random GLM parameter B_i and an observation vector $Y_i = (Y_{ij})_{j=1}^n$. We make the following assumptions for all $i = 1, \dots, N$.

(A1) Conditional on $B_i = \beta_i$, the Y_{ij}, $j = 1, \dots, n$, are independent and their distributions belong to a simple exponential family with natural parameters $\theta_i = (\theta_{ij})_{j=1}^n \subset \Theta$. The conditional joint pdf f_{β_i} takes the form

$$f_{\beta_i}(y) = \prod_{j=1}^n c(y_j) \exp\left(\sum_{j=1}^n \theta_{ij} y_j - b(\theta_{ij}) \right), \quad y \in \mathbb{R}^n, \tag{3.1}$$

where c and b have been defined in (2.17).

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

$$\theta_i = g\left(\mathbb{E}[Y_i \mid B_i]\right) = X B_i, \quad \text{a.s.}, \tag{3.2}$$

where g is the natural link function, $X \in \mathbb{R}^{n,p}$ is a known design matrix and B_i is a p-variate random GLM parameter.

(A3) The pairs $(B_1, Y_1), \dots, (B_N, Y_N)$ are iid.

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 an univariate simple exponential family with parameter β_i. As g is the natural link function, the linear predictor $X_j B_i$ describes the first two conditional moments of Y_{ij} through the mean and variance functions, respectively. Thus, we write

$$\mu_j(B_i) = b'(X_j B_i) = \mathbb{E}[Y_{ij} \mid B_i],$$
$$v_j(B_i) = b''(X_j B_i) = \mathrm{Var}(Y_{ij} \mid B_i).$$

The B_i characterize the individual distributions of the clusters but without conditioning on $\mathbf{B}$, the clusters are homogeneous as declared by assumption (A3). As we combine ideas from credibility theory and GLMs, we refer to the model as 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 in the end of Section 3.5.

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. Estimation again follows a two-step procedure. In the first step, best individual solutions are obtained by considering the local problem where the B_i are treated as unknown constants. The final

credibility estimator is then calculated as the optimal solution within an admissible class of estimators and optimization regards for the randomness of the B_i. Recall the multivariate and regression credibility models we have previously introduced. In both cases, the credibility estimator was defined as the orthogonal projection of B_i on the space

$$L(1,\mathbf{Y}) = \left\{ a + \sum_{i=1}^{N}\sum_{j=1}^{n} A_{ij}Y_{ij} : a \in \mathbb{R}^n, A_{ij} \in \mathbb{R}^{n,n} \right\} \subset \mathcal{L}^2.$$

As the best individual solutions were elements of $L(1,\mathbf{Y})$ we had $L(1,\mathbf{Y}) \supseteq L(1,\bar{\mathbf{Y}})$ (respectively $L(1,\boldsymbol{\beta})$) and solving for the minimum mean squared error estimator revealed that

$$\mathrm{Pro}(B_i \mid L(1,\mathbf{Y})) = \begin{cases} \mathrm{Pro}(B_i \mid L(1,\bar{\mathbf{Y}})) & \text{in the multivariate case,} \\ \mathrm{Pro}(B_i \mid L(1,\boldsymbol{\beta})) & \text{in the linear regression case,} \end{cases}$$

cf. (2.9) and (2.16). Difficulty arises in the present case of GLMs since 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. In fact, $L(1,\mathbf{Y})$ does not contain the individual solutions, which will be introduced soon, and we should directly select the credibility estimator within the class of the best individual solutions instead. From a different point of view, one could have also defined the multivariate and regression credibility estimators by means of (2.9) and (2.16) respectively. Their optimality even in the larger class $L(1,\mathbf{Y})$ is a nice-to-have property.

What do the best individual solutions look like? If we treat B_i as an fixed but unknown constant β_i, the natural choice will be the best estimator of a GLM, i.e. the maximum likelihood estimator (MLE) $\hat{\beta}_i$. The MLE is defined through the quantities (2.22) to (2.24) which now involve an additional cluster index i. More precisely, $\hat{\beta}_i$ solves

$$s_{in}(\hat{\beta}_i) = \frac{\partial l_{in}(\hat{\beta}_i)}{\partial \beta} = \sum_{j=1}^{n} X_j' \left(Y_{ij} - \mu_j(\hat{\beta}_i) \right) = 0. \tag{3.3}$$

Special caution is needed as the GLM parameter itself is actually a random variable. The maps l_{in} and s_{in} represent the true log-likelihood function and the true score function under the conditional measure $\mathbb{P}_\beta$ respectively but not under the unconditional measure $\mathbb{P}$. The latter requires integration involving the prior distribution of the 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 learning 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

$$L(1,\boldsymbol{\beta}) = \left\{ a + \sum_{i=1}^{N} A_i\hat{\beta}_i : a \in \mathbb{R}^p, A_i \in \mathbb{R}^{p,p} \right\}. \tag{3.4}$$

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

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

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 3.1. We consider the Poisson case with the simple $n \times 1$ design matrix $X = (1 \dots 1)'$ giving that, conditional on $B_i = \beta_i$,

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

Then,

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

and it follows that

$$\hat{\beta}_i = \log\left(\frac{1}{n}\sum_{j=1}^{n} Y_{ij}\right). \tag{3.5}$$

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 and a possible approach has been inconspicuously given in Example 3.1. Structure (3.5) seems not only to be a counterexample for $\hat{\beta}_i \notin \mathcal{L}^2$ 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 (3.5), one can easily check that

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

We say that

$$\hat{\beta}_i \mathbb{1}_{\{\sum_{j=1}^{n} Y_{ij} > 0\}} \tag{3.6}$$

is a MLE on $\{\sum_{j=1}^{n} Y_{ij} > 0\}$. It is such a defining set which plays a crucial role in credibility estimation and its proper construction is already half the battle.

3.2. The Pseudo Maximum Likelihood Estimator

To stress the role of the defining set and to clearly distinguish between the unrestricted MLE, we introduce an explicit notation for estimators of type (3.6). For some family of measurable sets $(M_{in})_{n\in\mathbb{N}} \subset \mathcal{F}$, we call

$$\tilde{\beta}_{in} := \hat{\beta}_{in} \mathbb{1}_{M_{in}}, \quad n \in \mathbb{N}, \tag{3.7}$$

the pseudo maximum likelihood estimator (PMLE). 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 also further properties will follow. For that purpose and for the remainder of the chapter, we will work 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.

$$\mathbb{P}(B_i = 0) = 0, \tag{3.8}$$

$$\mathbb{P}(B_i \in \partial\mathcal{B}) = 0. \tag{3.9}$$

(R2) 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 $\beta \in \mathcal{B}$.

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

(R5) The scaled Fisher information matrix

$$\frac{F_n(\beta)}{n} = \frac{1}{n}(X'W(\beta)X),$$

with $W(\beta) = \operatorname{diag}\{v_j(\beta) : j = 1, \ldots, n\}$, converges pointwise to a positive definite limit $F(\beta)$ for all $\beta \in \mathcal{B}$.

Remark. The above assumptions have to ensure that the conditional GLM is pointwise valid for all realizations $\beta \in \mathcal{B}$ of B_i. Thus, they naturally coincide with the assumptions (L1) to (L4) of a classical GLM. In particular, (R2) and (R4) are exactly the same as (L2) and (L4) respectively. Condition (R1) is the counterpart to (L1) that stated openness and convexity of $\mathcal{B}$. Compactness seems to be in conflict with (L1) but we can without loss of generality enlarge $\mathcal{B}$ to an open set, say $\tilde{\mathcal{B}}$, where $\tilde{\mathcal{B}}\backslash\mathcal{B}$ gets zero weight. The main purpose of (R1) is to make the B_i almost surely bounded and additional compactness is required for technical reasons. The second part about point zero is just for technical purposes and excludes only the trivial case where a regression model is redundant. Similar to (L3), assumption (R3) deals with the admissible set of covariates. In addition to the former prerequisites boundedness is now needed. Only (R5) is a totally new assumption and concerns with the 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 remainder of this section is devoted for the explicit construction of the sets (M_{in}) which will define the PMLE. The idea comes from Fahrmeir and Kaufmann (1985). For $\delta > 0$, we define a sequence of neighborhoods

$$N_n(\delta, B_i) := \{\beta \in \mathcal{B} : \sqrt{n}\|\beta - B_i\| \leq \delta\}, \quad n \in \mathbb{N}, \tag{3.10}$$

which are spheres with radius $\delta/\sqrt{n}$ and random central point B_i with respect to the vector 2-norm $\|\cdot\|$. In addition, let

$$M_{in}^{\delta} := \{l_{in}(\beta) - l_{in}(B_i) < 0, \quad \text{for all } \beta \in \partial N_n(\delta, B_i)\}. \tag{3.11}$$

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 an 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.

$$\mathbb{1}_{M_{in}^{\delta}} \|\hat{\beta}_{in} - B_i\| \leq \frac{\delta}{\sqrt{n}}, \quad \text{a.s.} \tag{3.12}$$

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

Theorem 3.2. *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.

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

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

Proof. See Section 3.5. □

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

Definition 3.3 (PMLE). Let (δ_n) be as in Theorem 3.2. 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 3.2 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.

Proposition 3.4. *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. Furthermore, asymptotic properties follow. In this connection, recall that matrix square roots are denoted by $A = A^{1/2}A^{T/2}$, where the superscript T addresses the matrix transpose. The inverse matrices are then given by $A^{-1} = A^{-T/2}A^{-1/2}$. See also Section 2.3 for an overview of notations.

Theorem 3.5. *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) $\tilde{\beta}_{in} - B_i$ *is asymptotically Normal, i.e.*

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

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

Proof. See Section 3.6. □

All properties except iii) are known from classical GLM theory, cf. Theorem 1 to 3 of Fahrmeir and Kaufmann (1985). Hence, if β denotes the true GLM parameter, the convergence towards β in the particular types will hold under $\mathbb{P}_\beta$. This theorem generalizes the 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 finally address the credibility estimation.

3.3. The credibility estimator

Recall that the the credibility estimator for B_i is generally defined as the orthogonal projection of the target variable on some proper linear subspace of $\mathcal{L}^2$ or, in other words, it is the Bayes estimator within this linear class. We have motivated to use structure (3.4) as the subspace but it has turned out that $L(1, \boldsymbol{\beta})$ is not contained in $\mathcal{L}^2$. We now redefine

$$L(1, \boldsymbol{\beta}) := \left\{ a + \sum_{k=1}^{N} A_k \tilde{\beta}_k : a \in \mathbb{R}^p, A_k \in \mathbb{R}^{p,p} \right\} \tag{3.13}$$

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 an univariate exponential family. It directly follows from Proposition 3.4 that $L(1, \boldsymbol{\beta})$ itself is a Hilbert space as 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 3.6 (GLM credibility estimator). The credibility estimator for B_i is defined as

$$\hat{B}_i = \mathrm{Pro}(B_i \mid L(1, \boldsymbol{\beta})) \tag{3.14}$$

or equivalently as

$$\hat{B}_i = \underset{\tilde{B}_i \in L(1,\boldsymbol{\beta})}{\arg\min}\, \mathbb{E}\left[\left(\tilde{B}_i - B_i\right)' \left(\tilde{B}_i - B_i\right)\right]. \tag{3.15}$$

By choosing $a = 0$ and $A_k = \delta_{ik} I_p$, with δ_{ik} being the Kronecker delta and $I_p \in \mathbb{R}^{p,p}$ 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 3.5, 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, see e.g. (Revuz and Yor, 1999, Chapter II). 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 now follows.

Theorem 3.7. *The solution of* (3.14) *and* (3.15) *is given by*

$$\hat{B}_i = \mathbb{E}[B_i] + A(\tilde{\beta}_i - \mathbb{E}[\tilde{\beta}_i]) \tag{3.16}$$

with credibility matrix

$$A = \mathrm{Cov}(B_i, \tilde{\beta}_i)\, \mathrm{Cov}(\tilde{\beta}_i)^{-1}. \tag{3.17}$$

Proof. By plugging the linear representation of $\tilde{B}_i$ according to (3.13) into the mean squared error (3.15), 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

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

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

whereat the latter equation simplifies to

$$\mathrm{Cov}(B_i, \tilde{\beta}_l) = \sum_{k=1}^N A_k\, \mathrm{Cov}(\tilde{\beta}_k, \tilde{\beta}_l), \quad l = 1, \ldots, N, \tag{3.18}$$

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, (3.18) 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

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

and putting together the pieces completes the proof. □

Notice that the covariance matrix $\operatorname{Cov}(\tilde{\beta}_i)$ is symmetric and positive semidefinite so it may happen that it is singular. However, this is the case 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. We can without loss of generality assume that $\operatorname{Cov}(\tilde{\beta}_i)$ is positive definite and thus invertible such that the credibility matrix A always exists. The resulting credibility formula (3.16) slightly differs from those for the multivariate and the regression models (2.8) and (2.15) respectively. However, an analogous structure can be established in an n-asymptotic meaning. Notice that A 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

$$(G_n - H_n) \in o(H_n). \tag{3.19}$$

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. We use the additional superscript n in $\sim$ to distinguish between the same symbol without a superscript which stands for "distributed as". As the name suggests, asymptotic equivalence is indeed an equivalence relation providing reflexivity, symmetry and transitivity. Reflexivity trivially follows. Symmetry is provided since for every $c > 0$ and large n,

$$\begin{aligned} \frac{\|H_n - G_n\|}{\|G_n\|} &= \frac{\|G_n - H_n\|}{\|H_n\|} \frac{\|G_n + (H_n - G_n)\|}{\|G_n\|} \\ &\leq \frac{c}{1+c}\left(1 + \frac{\|H_n - G_n\|}{\|G_n\|}\right) \\ \Rightarrow \quad \frac{1}{1+c}\frac{\|H_n - G_n\|}{\|G_n\|} &\leq \frac{c}{1+c}. \end{aligned}$$

$(H_n - G_n) \in o(G_n)$ then follows by multiplying both sides with $1 + c$. This property together with the triangular inequality directly imply transitivity of $\overset{n}{\sim}$. We can generalize the notion of asymptotic equivalence to random vectors and matrices by interpreting (3.19) in a probabilistic manner. To this end, we use the $o_{\mathbb{P}}$-notation introduced by Pratt (1959).

Definition 3.8. 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 3.9. *An n-asymptotic credibility formula is given by*

$$\hat{B}_{in} \overset{n}{\sim} A_n\tilde{\beta}_{in} + (I_p - A_n)\mathbb{E}[\tilde{\beta}_{in}], \tag{3.20}$$

where A_n is the same as in (3.17).

The right hand side of (3.20) simply follows by replacing $\mathbb{E}[B_i]$ in (3.16) with its approximation $\mathbb{E}[\tilde{\beta}_{in}]$ but the formal proof requires some preparation. Notice that the asymptotic formula now has the familiar structure which is easy to interpret. The credibility estimator is composed of a cluster specific term $A_n\tilde{\beta}_{in}$ and a cluster common term $(I_p - A_n)\mathbb{E}[\tilde{\beta}_{in}]$. The best individual estimator $\tilde{\beta}_{in}$ is weighted according to its credibility A_n. If it is evaluated to be highly credible, i.e. $A_n \approx I_p$, 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 cluster common part will compensate the lack of information if necessary. Thus, it can be interpreted as a learning effect which vanishes as n increases.

Lemma 3.10. *We have*

$$A_n \overset{n\to\infty}{\longrightarrow} I_p$$

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

Proof. By (3.17),

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

so that

$$A_n - I_p = \Big(\operatorname{Cov}(B_i, \tilde{\beta}_{in}) - \operatorname{Cov}(\tilde{\beta}_{in})\Big) \operatorname{Cov}(\tilde{\beta}_{in})^{-1}. \tag{3.21}$$

The law of total covariance, cf. Sheldon et al. (2002), 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, (3.21) can be written as

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

By claims iii) and iv) of Theorem 3.5, 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

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

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

Lemma 3.11. *The credibility estimator $\hat{B}_{in}$ is weakly consistent.*

Proof. The claim directly follows by the weak consistence of $\tilde{\beta}_i$, its asymptotically unbiasedness and Lemma 3.10. To be more precise,

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

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

We can now prove the asymptotic credibility formula (3.20).

Proof of Theorem 3.9. We have to show that

$$\left(A_n\tilde{\beta}_{in} + (I_p - A_n)\mathbb{E}[\tilde{\beta}_{in}]\right) \overset{n}{\sim} \left(\mathbb{E}[B_i] + A(\tilde{\beta}_{in} - \mathbb{E}[\tilde{\beta}_{in}])\right) \tag{3.22}$$

in the sense of Definition 3.8. In fact, their difference is

$$\mathbb{E}[\tilde{\beta}_{in}] - \mathbb{E}[B_i],$$

whose norm vanishes by ii) of Theorem 3.5. It suffices to show that it is also asymptotically dominated by the right hand side of (3.22) which is $\hat{B}_{in}$. Lemma 3.11 implies that $\|\hat{B}_{in}\|$ converges in probability to $\|B_i\|$. Therefore,

$$\frac{\|\mathbb{E}[\tilde{\beta}_{in}] - \mathbb{E}[B_i]\|}{\|\hat{B}_{in}\|}$$

is a product of two converging sequences. It converges 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. □

It remains to estimate the structural parameters of the model. Considering the asymptotic credibility formula (3.20) and Definition (3.17), these are specifically

$$\begin{aligned}\beta_0 &:= \mathbb{E}[\tilde{\beta}_i], \\ \tau^{-1} &:= \mathrm{Cov}(\tilde{\beta}_i)^{-1}, \\ T &:= \mathrm{Cov}(B_i, \tilde{\beta}_i)\end{aligned}$$

and their estimation is based on the N iid samples $\tilde{\beta}_1, \ldots, \tilde{\beta}_N$. The first two quantities allow empirical estimation but the third one requires some work. The next theorem contains convergence statements with two control variables n and N. Limiting behavior

of a sequence $(G_{n,N})$ as both $n \to \infty$ and $N \to \infty$ should be interpreted as their successive application, i.e. as

$$\lim_{n\to\infty} \lim_{N\to\infty} G_{n,N}.$$

Theorem 3.12. *The structural parameters can be estimated as follows.*

i) An unbiased and strongly N-consistent estimator for $\mathbb{E}[\tilde{\beta}_i]$ is given by

$$\hat{\beta}_0 = \frac{1}{N} \sum_{i=1}^{N} \tilde{\beta}_i. \tag{3.23}$$

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

$$\widehat{\tau^{-1}} = \frac{N-p-2}{N-1} (\hat{\tau})^{-1} \tag{3.24}$$

is a weakly N-consistent estimator of $\mathrm{Cov}(\tilde{\beta}_i)^{-1}$.

iii) Let

$$\hat{S} = \frac{1}{N} \sum_{i=1}^{N} F_n^{-1}(\tilde{\beta}_i). \tag{3.25}$$

Then,

$$\hat{T} = \hat{\tau} - \hat{S} \tag{3.26}$$

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

iv) The credibility matrix can be estimated by

$$\hat{A} = \hat{T}\widehat{\tau^{-1}}, \tag{3.27}$$

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

The estimator (3.24) for the inverse covariance matrix includes a factor $\frac{N-p-2}{N-1}$, which will be discussed after the proof.

Proof. *ad i).* As the $\tilde{\beta}_i$ are iid, the strong law of large numbers provides strong consistency of $\hat{\beta}_0$. Unbiasedness directly follows by construction.

ad ii). The sample covariance matrix

$$\hat{\tau} = \frac{1}{N-1} \sum_{i=1}^{N} (\tilde{\beta}_i - \hat{\beta}_0)(\tilde{\beta}_i - \hat{\beta}_0)'$$

is an unbiased and weakly consistent estimator of $\mathrm{Cov}(\tilde{\beta}_i)$ as $N \to \infty$. Therefore, by the continuous mapping theorem, cf. (Klenke, 2006, Theorem 13.25), $\widehat{\tau^{-1}}$ consistently estimates the inverse covariance matrix.

ad iii). First, we show the asymptotic equivalence of the expressions

$$\mathrm{Cov}(B_i, \tilde{\beta}_{in}) \overset{n}{\sim} \mathrm{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right). \tag{3.28}$$

Since

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

it suffices to prove that

$$\operatorname{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) \overset{n}{\sim} \operatorname{Cov}\left(B_i, \mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right).$$

This is easy to see as both terms converge to $\operatorname{Cov}(B_i)$, cf. Theorem 3.5. Thus,

$$\operatorname{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) \overset{n}{\sim} \operatorname{Cov}(B_i) \quad \text{and} \quad \operatorname{Cov}\left(B_i, \mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) \overset{n}{\sim} \operatorname{Cov}(B_i)$$

so (3.28) follows by the transitivity of $\overset{n}{\sim}$. We now estimate $\operatorname{Cov}(B, \tilde{\beta}_i)$ by means of (3.28) and we use the identity

$$\operatorname{Cov}\left(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]\right) = \operatorname{Cov}(\tilde{\beta}_{in}) - \mathbb{E}\left[\operatorname{Cov}(\tilde{\beta}_{in} \mid B_i)\right]. \tag{3.29}$$

As we have already seen, $\hat{\tau}$ is an unbiased and weakly N-consistent estimator for $\operatorname{Cov}(\tilde{\beta}_{in})$. An intuitive choice for the second summand is

$$\frac{1}{N}\sum_{i=1}^{N} \operatorname{Cov}(\tilde{\beta}_{in} \mid B_i). \tag{3.30}$$

However, these conditional covariance matrices are not known so that we look for estimators

$$\widehat{\operatorname{Cov}(\tilde{\beta}_{in} \mid B_i)}$$

that are asymptotically unbiased and consistent as n grows to infinity. By claim v) of Theorem 3.5, the inverse Fisher information matrix is the asymptotic conditional covariance matrix of $\tilde{\beta}_{in}$ and is therefore a natural choice. In fact, it follows from Lemma 3.14 and (3.9) that

$$\lim_{n\to\infty} F_n^{-1}(\tilde{\beta}_{in}) = 0, \quad \text{a.s. and in } \mathcal{L}^1$$

and from Theorem 3.5 claim iv) that

$$\lim_{n\to\infty} \operatorname{Cov}(\tilde{\beta}_{in} \mid B_i) = 0, \quad \text{a.s. and in } \mathcal{L}^1.$$

Thus,

$$\hat{S} = \frac{1}{N}\sum_{i=1}^{N} F_n^{-1}(\tilde{\beta}_{in})$$

is an n-asymptotically unbiased and strongly n-consistent estimator for (3.30), which itself is an unbiased and strongly N-consistent estimator for $\mathbb{E}[\operatorname{Cov}(\tilde{\beta}_{in} \mid B_i)]$. Altogether, $\hat{T} = \hat{\tau} - \hat{S}$ estimates (3.29) in an asymptotically unbiased and consistent way as both n and N grow to infinity. $\hat{T}$ preserves these properties for the left hand side of (3.28). More precisely,

$$\begin{aligned}
&\lim_{n\to\infty}\lim_{N\to\infty} \left\| \hat{T} - \operatorname{Cov}(B_i, \tilde{\beta}_{in}) \right\| \\
&\leq \lim_{n\to\infty}\lim_{N\to\infty} \left\| \hat{T} - \operatorname{Cov}(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]) \right\| + \lim_{n\to\infty}\lim_{N\to\infty} \left\| \operatorname{Cov}(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]) - \operatorname{Cov}(B_i, \tilde{\beta}_{in}) \right\| \\
&\leq \lim_{n\to\infty}\lim_{N\to\infty} \left\| \hat{\tau} - \operatorname{Cov}(\tilde{\beta}_{in}) \right\| + \lim_{n\to\infty}\lim_{N\to\infty} \left\| \mathbb{E}\left[\operatorname{Cov}(\tilde{\beta}_{in} \mid B_i)\right] - \hat{S} \right\| \\
&\qquad + \lim_{n\to\infty} \left\| \operatorname{Cov}(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]) - \operatorname{Cov}(B_i, \tilde{\beta}_{in}) \right\|
\end{aligned}$$

almost surely. Convergences of the first two summands have just been discussed. For the last summand, asymptotic dominance (3.28) provides

$$\left\|\mathrm{Cov}(\mathbb{E}[\tilde{\beta}_{in} \mid B_i]) - \mathrm{Cov}(B_i, \tilde{\beta}_{in})\right\| \leq c \left\|\mathrm{Cov}(B_i, \tilde{\beta}_{in})\right\|$$

for all $c > 0$ and n large enough. The right hand side vanishes by taking $c \downarrow 0$ and since

$$\sup_n \left\|\mathrm{Cov}(B_i, \tilde{\beta}_{in})\right\| < \infty$$

as $\mathrm{Cov}(B_i, \tilde{\beta}_{in})$ converges.

ad iv). The claim directly follows by applying the continuous mapping theorem on ii) and iii). □

The credibility formula can now be evaluated by replacing its structural parameters by their consistent estimators.

Corollary 3.13. *The estimator*

$$\hat{\tilde{B}}_i = \hat{A}\tilde{\beta}_i + (I_p - \hat{A})\hat{\beta}_0. \tag{3.31}$$

is a weakly consistent estimator for the exact credibility estimator (3.16) *as both* $n \to \infty$ *and* $N \to \infty$.

Proof. The claim is a direct consequence of Theorem 3.9 and Theorem 3.12. □

In practice, we recommend two slight modifications of the presented estimators. The first has been already indicated and concerns the estimation of the inverse covariance matrix τ^{-1}. Not only $\widehat{\tau^{-1}}$ but also $(\hat{\tau})^{-1}$ consistently estimates τ^{-1}. The additional factor of the former is justified by the asymptotic distribution. Since the $\tilde{\beta}_i$ are n-asymptotically Normal, cf. Theorem 3.5, $(N-1)\hat{\tau}$ is n-asymptotically Wishart distributed with $N-1$ degrees of freedom. The asymptotic distribution of $\frac{1}{N-1}(\hat{\tau})^{-1}$ is the inverse Wishart distribution with $N-1$ degrees of freedom and its expectation contains a factor $(N-p-2)^{-1}$. Thus, $\frac{N-p-2}{N-1}$ in (3.24) serves as a bias correction term. The estimator is not necessarily asymptotically unbiased due to the fact that the convergence to the inverse Wishart distribution only holds in law. Nevertheless, simulation studies reveal that $\widehat{\tau^{-1}}$ performs much better than $(\hat{\tau})^{-1}$.
The second modification applies to the estimator of $T = \mathrm{Cov}(\mathbb{E}[\tilde{\beta}_i \mid B_i])$. As $\hat{T}$ estimates a covariance matrix, it should be symmetric and positive semidefinite. By the nature of structure (3.26), $\hat{T}$ is symmetric but the latter property is not guaranteed. We therefore propose a transformation of $\hat{T}$ in the following way. 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

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

and empirical evidence shows that this estimator performs better than $\hat{T}$. Details will follow shortly.

$N\backslash n$	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 3.1. Relative improvement of the credibility estimator compared to the PMLE. The line MSE shows the simulated mean squared error of the PMLE.

One drawback of the credibility estimation is that the PMLE $\tilde{\beta}_i$ cannot be explicitly calculated since the corresponding sets M_{in}, cf. (3.11), have an abstract structure. The PMLE is however more a theoretical construct introduced to build up the theory in a clean mathematical way. Whenever the MLE $\hat{\beta}_i$ can be numerically computed, we pretend as M_{in} has actually occurred so that the PMLE and the MLE agree. This step can be theoretically justified by making the probability of agreement arbitrarily close to 1. According to Theorem 3.2, $\mathbb{P}(M_{in}) \geq 1 - \eta_n$ and we can choose a small $\eta_n > 0$ for the particular sample size n of the data in fit.

3.4. Simulation Study

We conclude the theory with an illustrative example that shows the performance of the credibility estimator. As mentioned earlier, credibility models target at situation where only little statistical information is available for the single clusters. In this section, we consider the case of a Poisson-CGLM. Let

$$B_i \sim \mathcal{N}(\begin{pmatrix}2 & 1\end{pmatrix}', I_2), \quad i = 1, \ldots, N,$$

be iid and capped whenever $\|B_i\| > 1000$. The cap is needed to comply the compactness condition (R1) even though it will never be reached in simulation. Conditional, given these B_i, Y_{ij} follow a Poisson-GLM with parameter B_i and covariates

$$X_j = \begin{pmatrix}1 & \frac{j}{n}\end{pmatrix}, \quad j = 1, \ldots, n.$$

Thus, the component B_{i1} describes the sensitivity to overall effects and B_{i2} that to linear effects. For several constellations of the portfolio and sample sizes (N, n), we run $m = 1000$ scenarios $\omega_1, \ldots, \omega_m$ and compute the performance of the credibility estimator. In detail, we calculate the aggregate realized quadratic losses of the credibility and the (pseudo) maximum likelihood estimator and measure the relative improvement by

$$\frac{\sum_{k=1}^{m} \sum_{i=1}^{N} \|\hat{\tilde{B}}_i(\omega_k) - B_i(\omega_k)\|^2}{\sum_{k=1}^{m} \sum_{i=1}^{N} \|\tilde{\beta}_i(\omega_k) - B_i(\omega_k)\|^2}. \tag{3.32}$$

In addition, modifications of the estimators are evaluated by means of this value. These deal with the bias correction term $\frac{N-p-2}{N-1}$ in (3.24) and the transformation of $\hat{T}$ into a positive semidefinite matrix. Recall that we have recommended to use both of them.

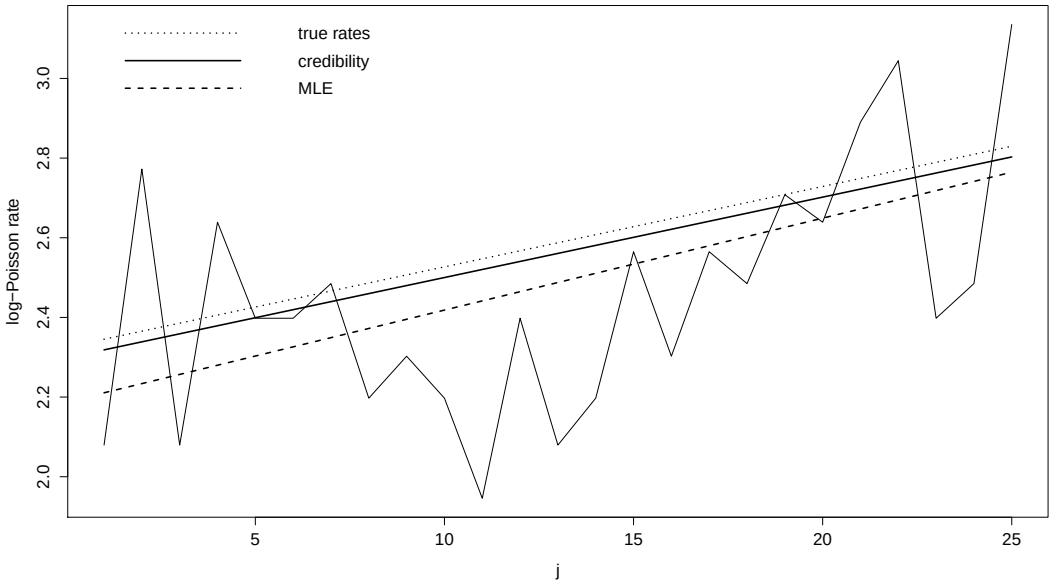

FIGURE 3.1. A particular realization and the associated estimators for $N = 30$, $n = 25$ and $\mathbb{E}[B_i] = (2.5, 1)'$.

Table 3.1 presents the simulation results. The columns labeled no.1 list the relative improvement of the credibility estimator without the two modifications. Both of them are applied in the estimation whose results are given in columns no.2. There are noticeable differences between these two estimators but differences reduce as N gets larger. We have also considered the modifications separately and the corresponding values lie somewhere between the displayed ones. 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 3.12 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. This effect is especially large if observations are atypical or contain irregularities as shown in Figure 3.1. The PMLE underestimates the Poisson rates since most observations deviate to the downside. However, the credibility estimator successfully adjusts towards the true Poisson rates. One should also notice that the credibility matrix A and its estimator $\hat{A}$ are not necessarily diagonal. The solid line and the dashed line representing the two estimators are in fact not parallel. In many scenarios, both estimators almost agree.

3.5. Proof of Theorem 3.2

The remainder of the chapter is devoted for proofs and we begin with that of Theorem 3.2 which claims asymptotic existence of the PMLE. Throughout the proof the cluster index i is omitted for notational simplicity, e.g. a particular element of $\mathbf{B} = (B_1, \dots, B_N)$ will be denoted by B instead of B_i.

Lemma 3.14. *For all $n \in \mathbb{N}$, $\frac{1}{n}F_n$ is Lipschitz continuous with Lipschitz constant $L > 0$ and the sequence $(\frac{1}{n}F_n)_{n\in\mathbb{N}}$ converges uniformly on $\mathcal{B}$, i.e.*

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

as $n \to \infty$.

Proof. Since $\frac{F_n}{n}$ converges pointwise and its domain $\mathcal{B}$ is compact, uniform convergence is guaranteed if the sequence $(\frac{F_n}{n})$ 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 $\beta_1, \beta_2 \in \mathcal{B}$, applying the mean value theorem on the variance function $v_j(\cdot) = b''(X_j\,\cdot)$ yields that

$$\begin{aligned}
\left\| \frac{1}{n}F_n(\beta_1) - \frac{1}{n}F_n(\beta_2) \right\| &\le \frac{1}{n}\sum_{j=1}^{n} \left\| X_j' X_j \left(v_j(\beta_1) - v_j(\beta_2) \right) \right\| \\
&\le \frac{1}{n}\sum_{j=1}^{n} \left\| X_j' X_j \right\| \left\| \sup_{\xi\in\overline{\beta_1\beta_2}} X_j' b^{(3)}(X_j\xi) \right\| \|\beta_1 - \beta_2\| \\
&\le L \|\beta_1 - \beta_2\| .
\end{aligned}$$

Such a bound $L > 0$ exists since $b' = g^{-1}$ is twice continuously differentiable by (R2) and since the domains of X_j and ξ are bounded. Also notice that $\mathcal{B}$ is convex and therefore contains ξ. □

The set M_n^δ. Recall the constructions (3.10)

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

for $\delta > 0$ and (3.11)

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

We have already seen that

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

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

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

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

$$l_n(\beta) - l_n(B) = (\beta - B)' s_n(B) - \frac{1}{2}(\beta - B)' F_n(\xi)(\beta - B), \tag{3.33}$$

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 (3.33) 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

$$v' s_n(B) \leq \max_{\|w\|=1} |w' s_n(B)| \leq \|s_n(B)\| \tag{3.34}$$

$$v' \frac{F_n(\xi)}{n} v \geq \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w. \tag{3.35}$$

The last inequality of (3.34) follows by the Cauchy-Schwarz inequality.

Lower bound for (3.35). 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 3.14, yields a lower bound for the expression

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

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

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

It follows from the structure of an induced matrix norm and from the Cauchy-Schwarz inequality that for all $w \in \mathbb{R}^p$ with $\|w\| = 1$,

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

and thus,

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

By assumption the matrix $F(\xi)$ is positive definite for all ξ. Furthermore, as an 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

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

Putting the pieces together. Altogether, we have

$$\begin{aligned}
&\mathbb{P}\left(\exists \beta \in \partial N_n(\delta, B) : l_n(\beta) - l_n(B) \geq 0\right) \\
&= \mathbb{P}\left(\exists \beta \in \partial N_n(\delta, B) : (\beta - B)' s_n(B) \geq \frac{1}{2}(\beta - B)' F_n(\xi)(\beta - B)\right) \\
&= \mathbb{P}\left(\exists \beta \in \partial N_n(\delta, B) : \frac{\delta}{\sqrt{n}} v' s_n(B) \geq \frac{1}{2}\delta^2 v' \frac{F_n(\xi)}{n} v\right) \\
&\leq \mathbb{P}\left(\frac{\delta}{\sqrt{n}}\|s_n(B)\| \geq \frac{1}{2}\delta^2 \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w\right) \\
&\leq \mathbb{P}\left(\|s_n(B)\| \geq \frac{1}{2}\sqrt{n}\delta d\right) \qquad (3.36)
\end{aligned}$$

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

$$\mathbb{P}\left(\|s_n(B)\| \geq \frac{1}{2}\sqrt{n}\delta d\right) \leq \frac{4}{n\delta^2 d^2}\mathbb{E}\left[\|s_n(B)\|^2\right], \qquad (3.37)$$

where the expectation linearly grows in n. More precisely, using the explicit structure of the vector 2-norm $\|\cdot\|$,

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

The norm $\|X_j\|$ is bounded according to assumption (R3). In addition, the boundedness of $\{X_j\}$ and $\mathcal{B}$ ensures that the conditional variances as continuous images $b''(X_j B)$ 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). Credibility estimation is purely based on the PMLEs, which are the best individual solutions. Drawing observations from q-variate simple exponential families only affects the construction of these estimators. In particular, conditional expectations and variances have to be consequently replaced by conditional expectation vectors and covariance matrices respectively. These changes concern the likelihood functions, score functions and the Fisher information matrices which now involve multivariate quantities. Their general structures have been already introduced in (2.22) to (2.24). Nevertheless, the proof of Theorem 3.2 remains valid. We only need to verify the last step (3.37). Specifically, we have q-variate quantities X_{jk}, Y_j and $\mu_j(B)$ so that

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

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.

3.6. Proof of Theorem 3.5

All proofs of the asymptotic properties are based on (3.12) and Theorem 3.2, 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

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

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,

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

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

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

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

$$\sup_n \|\mathbb{E}[\tilde{\beta}_n]\| \leq \sup_n \mathbb{E}[\|\tilde{\beta}_n\|] < \infty. \tag{3.38}$$

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 (3.38),

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

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

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

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

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

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

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

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

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

and

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

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

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

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, Theorem 13.18).

CHAPTER 4

Particular cases of CGLMs

Particular distributions and types of covariates often appear in regression models for insurance data. This chapter investigates CGLM for two important situations.

4.1. Poisson data

Modeling frequency variables is a common task in many applications, e.g. the number of certain events within a certain time period. Generally, the Poisson distribution is the first point of contact for these cases. The previous chapter has broadly dealt with simple exponential families and we are going to take a closer look at Poisson-CGLMs. They are characterized by the conditional pdf of the Y_{ij} given $B_i = \beta_i$, which now has the structure

$$f_{\beta_i}(y) = \prod_{j=1}^{n} \frac{1}{y_j!} \exp\left(\sum_{j=1}^{n} \theta_{ij} y_j - \exp(\theta_{ij}) \right), \quad y \in \mathbb{N}^n.$$

We will review the construction of the PMLE under the Poisson assumption and discuss its effects on the credibility estimator. In doing so, the regularity assumptions (R1) to (R5) remain unchanged. Notice that the natural link function of a Poisson distribution is the log-link, cf. Table 2.1, and it satisfies the differentiability condition (R2) on the parameter space $(0, \infty)$.

Large Deviation Techniques. First recall the construction of the events $M_{in}^{\delta_n}$, which led to Definition 3.3 of the PMLE. We have chosen

$$M_{in}^{\delta_n} = \{ l_{in}(\beta) - l_{in}(B_i) < 0, \quad \text{for all } \beta \in \partial N_n(\delta_n, B_i) \},$$

where

$$N_n(\delta_n, B_i) = \{ \beta \in \mathcal{B} : \sqrt{n} \|\beta - B_i\| \leq \delta_n \}.$$

A vanishing upper bound for the probability of the complement of $M_{in}^{\delta_n}$ has been derived by Chebyshev's inequality, see (3.37). While the inequality can be applied to a broad class of random variables, the provided upper bound is in general very rough. Cramér's theorem, see for instance Klenke (2006), is restricted to a special class of random variables but usually provides a better approximation as the upper bound converges to the exact probability. More precisely, let S_n be the sum of n iid random variables $Z_1, \ldots, Z_n$ and $x > \mathbb{E}[Z_1]$. Then, the probability of large deviation satisfies

$$\frac{1}{n} \log \mathbb{P}(S_n \geq xn) \leq -I(x)$$

$$\text{and} \qquad \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq xn) = -I(x),$$

where

$$I(x) = \sup_{t\in\mathbb{R}} \left(tx - \log \mathbb{E}[\exp(tZ_1)]\right) \tag{4.1}$$

is the Legendre transform. In contrast to Chebyshev's inequality, the upper bound decays exponentially fast in n. We will repeat the last step of the proof of Theorem 3.2 using ideas from Cramér's large deviation theory. In doing so, we will rely on the explicit structure of the conditional expectations and variances, which are

$$b'(X_jB) = b''(X_jB) = \exp(X_jB) =: \lambda_j(B).$$

The cluster index i will agin be omitted.

We have already seen in (3.36) that for sufficiently large n

$$1 - \mathbb{P}(M_n^{\delta_n}) \leq \mathbb{P}\left(\|s_n(B)\| \geq \frac{d}{2}\sqrt{n}\delta_n\right) \tag{4.2}$$

and the right hand side had an upper bound that, up to some constant factors, vanished at rate δ_n^{-2}. This rate was limited since we required $\frac{\delta_n}{\sqrt{n}}$ to converge to zero at the same time. The approximation of the upper bound can now be improved. Using the definition of the 2-norm, the inequality (4.2) can be further continued as

$$\begin{aligned}
\mathbb{P}\left(\|s_n(B)\| \geq \frac{d}{2}\sqrt{n}\delta_n\right) &= \mathbb{P}\left(\sum_{k=1}^{p}\left(\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(B))\right)^2 \geq \frac{d^2}{4}n\delta_n^2\right) \\
&\leq \sum_{k=1}^{p}\mathbb{P}\left(\left|\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(B))\right| \geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right) \\
&\leq \sum_{k=1}^{p}\mathbb{P}\left(\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(B)) \geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right) \\
&\quad + \sum_{k=1}^{p}\mathbb{P}\left(\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(B)) \leq -\frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right).
\end{aligned} \tag{4.3}$$

The second line follows since at least one of the p summands must exceed the average. All summands on the last line are probabilities of large deviations. Let $1 \leq k \leq p$. Then, writing

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(B)) \geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right) \\
&\quad = \int_{\mathcal{B}} \mathbb{P}_\beta\left(\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(\beta)) \geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right)\mathbb{P}(B \in d\beta)
\end{aligned}$$

allows us to consider the conditional probabilities instead under which the Y_j are Poisson distributed random variables.

We apply Markov's inequality with the exponential function to obtain

$$\mathbb{P}_\beta\left(\sum_{j=1}^n X_{jk}(Y_j-\lambda_j(\beta))\geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right)$$
$$\leq \exp\left(-\frac{d}{2\sqrt{p}}t_n\sqrt{n}\delta_n\right)\mathbb{E}_\beta\left[\exp\left(t_n\sum_{j=1}^n X_{jk}(Y_j-\lambda_j(\beta))\right)\right] \tag{4.4}$$

with an arbitrary $t_n>0$. The expectation on the right hand side can be further simplified as the Y_j are conditionally independent Poisson variables. Specifically,

$$\begin{aligned}\mathbb{E}_\beta\left[\exp\left(t_n\sum_{j=1}^n X_{jk}(Y_j-\lambda_j(\beta))\right)\right] &= \prod_{j=1}^n \mathbb{E}_\beta\Big[\exp(t_nX_{jk}(Y_j-\lambda_j(\beta)))\Big]\\ &= \prod_{j=1}^n \exp(-t_nX_{jk}\lambda_j(\beta))\mathbb{E}_\beta\left[\exp((t_nX_{jk})Y_j)\right]\\ &= \prod_{j=1}^n \exp(-t_nX_{jk}\lambda_j(\beta))\exp(\lambda_j(\beta)(\exp(t_nX_{jk})-1))\\ &= \exp\left(\sum_{j=1}^n \lambda_j(\beta)(\exp(t_nX_{jk})-t_nX_{jk}-1)\right).\end{aligned}$$

Hence, the probability on the left hand side of (4.4) is bounded from above by $\exp(-r_n)$, where

$$r_n := \frac{d}{2\sqrt{p}}t_n\sqrt{n}\delta_n - \sum_{j=1}^n \lambda_j(\beta)(\exp(t_nX_{jk})-t_nX_{jk}-1). \tag{4.5}$$

We maximize r_n with respect to t_n to obtain the analogous structure to the Legendre transform (4.1). However, structure (4.5) allows no analytic solution of the maximization problem. We will construct a lower bound $\underline{r_n}$ of r_n with appropriate sequences $(t_n)_n$ and $(\delta_n)_n$ such that $\underline{r_n}\to\infty$ and $\delta_n/\sqrt{n}\to 0$ to avoid this problem.

The regularity assumptions (R1) to (R3) immediately imply that there exist constants c_X and $c_\lambda>0$ with

$$|X_{jk}|<c_X \quad \text{and} \quad \lambda_j(B)\leq c_\lambda$$

for all $j\in\mathbb{N}$ and $1\leq k\leq p$. Writing the exponential function as a power series gives that

$$\begin{aligned}\sum_{j=1}^n \lambda_j(\beta)(\exp(t_nX_{jk})-t_nX_{jk}-1) &= \sum_{j=1}^n \lambda_j(\beta)\sum_{l=2}^\infty \frac{(t_nX_{jk})^l}{l!}\\ &\leq \sum_{j=1}^n \lambda_j(\beta)\sum_{l=2}^\infty \frac{(t_nc_X)^l}{l!}\\ &= \sum_{j=1}^n \lambda_j(\beta)(\exp(t_nc_X)-t_nc_X-1).\end{aligned}$$

Since

$$\exp(t_n c_X) - t_n c_X - 1 \geq 0$$

we obtain

$$r_n \geq \frac{d}{2\sqrt{p}} t_n \sqrt{n}\delta_n - c_\lambda n(\exp(t_n c_X) - t_n c_X - 1) =: \underline{r_n}.$$

Notice that this lower bound does neither depend on β nor on k. Now, choose

$$t_n = \frac{1}{c_X} \log\left(1 + \frac{1}{\sqrt{n}}\right), \quad n \in \mathbb{N}.$$

Then,

$$\underline{r_n} = \frac{d}{2\sqrt{p}c_X} \log\left(1 + \frac{1}{\sqrt{n}}\right) \sqrt{n}\delta_n - \sqrt{n}c_\lambda \left(1 - \sqrt{n} \log\left(1 + \frac{1}{\sqrt{n}}\right)\right) \tag{4.6}$$

and it suffices to find a sequence $\delta_n \in o(\sqrt{n})$ such that $\underline{r_n} \to \infty$. Let

$$\delta_n = n^\alpha, \quad 0 < \alpha < \frac{1}{2}.$$

The α-condition ensures that $\frac{\delta_n}{\sqrt{n}}$ converges to zero. We consider the two summands in (4.6) separately. By applying l'Hôpital's rule, we get

$$\begin{aligned}
\lim_{n\to\infty} \log\left(1 + \frac{1}{\sqrt{n}}\right) \sqrt{n}\delta_n &= \lim_{n\to\infty} \frac{\log\left(1 + \frac{1}{\sqrt{n}}\right)}{n^{-\alpha-\frac{1}{2}}} \\
&= \lim_{n\to\infty} \frac{(-\frac{1}{2})n^{-\frac{3}{2}}}{\left(1 + \frac{1}{\sqrt{n}}\right)(-\alpha - \frac{1}{2})n^{-\alpha-\frac{3}{2}}} \\
&= \lim_{n\to\infty} \frac{1}{2}\left(\alpha + \frac{1}{2}\right)^{-1} n^\alpha \left(1 + \frac{1}{\sqrt{n}}\right)^{-1} \\
&= \lim_{n\to\infty} \frac{1}{2}\left(\alpha + \frac{1}{2}\right)^{-1} n^\alpha \\
&= \infty.
\end{aligned}$$

Since $d/(2\sqrt{p}c_X) > 0$, the first summand grows to infinity with speed n^α. Next, we study

$$l_\infty := \lim_{n\to\infty} \sqrt{n} \left(1 - \sqrt{n} \log\left(1 + \frac{1}{\sqrt{n}}\right)\right)$$

to determine the limit of the second summand in (4.6). Since

$$\begin{aligned}
\lim_{n\to\infty} \sqrt{n} \log\left(1 + \frac{1}{\sqrt{n}}\right) &= \lim_{n\to\infty} \frac{\log\left(1 + \frac{1}{\sqrt{n}}\right)}{n^{-\frac{1}{2}}} \\
&= \lim_{n\to\infty} \frac{(-\frac{1}{2})n^{-\frac{3}{2}}}{\left(1 + \frac{1}{\sqrt{n}}\right)(-\frac{1}{2})n^{-\frac{3}{2}}} \\
&= 1,
\end{aligned}$$

we can apply l'Hôpital's rule to determine l_∞. More precisely,

$$\begin{aligned} l_\infty &= \lim_{n\to\infty} \frac{1-\sqrt{n}\log\left(1+\frac{1}{\sqrt{n}}\right)}{n^{-\frac{1}{2}}} \\ &= \lim_{n\to\infty} \frac{(-\frac{1}{2})n^{-\frac{1}{2}}\log\left(1+\frac{1}{\sqrt{n}}\right) - \sqrt{n}\left(1+\frac{1}{\sqrt{n}}\right)^{-1}(-\frac{1}{2})n^{-\frac{3}{2}}}{(-\frac{1}{2})n^{-\frac{3}{2}}} \\ &= \lim_{n\to\infty} n\log\left(1+\frac{1}{\sqrt{n}}\right) - \sqrt{n}\left(1+\frac{1}{\sqrt{n}}\right)^{-1} \\ &= \lim_{n\to\infty} \sqrt{n}\left(\sqrt{n}\log\left(1+\frac{1}{\sqrt{n}}\right) - \left(1+\frac{1}{\sqrt{n}}\right)^{-1}\right). \end{aligned} \tag{4.7}$$

Applying l'Hôpital another time gives

$$\begin{aligned} l_\infty &= \lim_{n\to\infty} \frac{\sqrt{n}\log\left(1+\frac{1}{\sqrt{n}}\right) - \left(1+\frac{1}{\sqrt{n}}\right)^{-1}}{n^{-1/2}} \\ &= \lim_{n\to\infty} \frac{\frac{1}{2}n^{-1/2}\log\left(1+\frac{1}{\sqrt{n}}\right) + \sqrt{n}\left(1+\frac{1}{\sqrt{n}}\right)^{-1}(-1/2)n^{-3/2} + \left(1+\frac{1}{\sqrt{n}}\right)^{-2}(-1/2)n^{-3/2}}{(-1/2)n^{-3/2}} \\ &= \lim_{n\to\infty} -n\log\left(1+\frac{1}{\sqrt{n}}\right) + \sqrt{n}\left(1+\frac{1}{\sqrt{n}}\right)^{-1} + \left(1+\frac{1}{\sqrt{n}}\right)^{-2} \\ &= \lim_{n\to\infty} -\sqrt{n}\left(\sqrt{n}\log\left(1+\frac{1}{\sqrt{n}}\right) - \left(1+\frac{1}{\sqrt{n}}\right)^{-1}\right) + 1 \end{aligned} \tag{4.8}$$

Comparing (4.7) with (4.8), we get

$$l_\infty = -l_\infty + 1$$

so that $l_\infty = 1/2$ follows. Altogether, $r_n \geq \underline{r_n} \to \infty$ and

$$\mathbb{P}_\beta\left(\sum_{j=1}^n X_{jk}(Y_j - \lambda_j(\beta)) \geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right) \leq \exp(-r_n) \leq \exp(-\underline{r_n}) \overset{n\to\infty}{\longrightarrow} 0.$$

Moreover, since the rate $\underline{r_n}$ does not depend on β and k, we even obtain a vanishing bound for the total probability and altogether,

$$\sum_{k=1}^p \mathbb{P}\left(\sum_{j=1}^n X_{jk}(Y_j - \lambda_j(B)) \geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right) \leq \exp\left(-(\log p)\underline{r_n}\right) \overset{n\to\infty}{\longrightarrow} 0.$$

The probability of negative deviations in (4.3) similarly follows. We have

$$\begin{aligned}
&\mathbb{P}_\beta\left(\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(\beta)) \leq -\frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right)\\
=&\mathbb{P}_\beta\left(-\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(\beta)) \geq \frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right)\\
\leq&\exp\left(-\frac{d}{2\sqrt{p}}t_n\sqrt{n}\delta_n\right)\mathbb{E}_\beta\left[\exp\left(-t_n\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(\beta))\right)\right]\\
=&\exp\left(-\left(\frac{d}{2\sqrt{p}}t_n\sqrt{n}\delta_n - \sum_{j=1}^{n}\lambda_j(\beta)(\exp(-t_nX_{jk}) + t_nX_{jk} - 1)\right)\right)\\
=&\exp(-\tilde{r}_n) \qquad (4.9)
\end{aligned}$$

with $t_n > 0$. The only difference to the former part is that the moment generating function is now evaluated at $-t_n$. As the covariates X_{jk} are bounded from below by $-c_X$, it holds that

$$\begin{aligned}
\exp(-t_nX_{jk}) + t_nX_{jk} - 1 &= \sum_{l=2}^{\infty}\frac{(-t_nX_{jk})^l}{l!}\\
&\leq \sum_{l=2}^{\infty}\frac{(t_nc_X)^l}{l!}\\
&= \exp(t_nc_X) - t_nc_X - 1.
\end{aligned}$$

Thus, $\tilde{r}_n \geq \underline{r_n}$, where $\underline{r_n}$ is defined as in (4.6). Finally, we again choose $\delta_n = n^\alpha$ with $0 < \alpha < 1/2$ to obtain

$$\sum_{k=1}^{p}\mathbb{P}\left(\sum_{j=1}^{n} X_{jk}(Y_j - \lambda_j(B)) \leq -\frac{d}{2\sqrt{p}}\sqrt{n}\delta_n\right) \leq \exp\left(-(\log p)\underline{r_n}\right) \overset{n\to\infty}{\longrightarrow} 0.$$

We summarize the result, which we have just proved, in the following theorem. It has major impact on the estimation of the structural parameters.

Theorem 4.1. *Let $\delta_n = n^\alpha$ with $0 < \alpha < 1/2$. Then, there exists a sequence $(\eta_n)_{n\in\mathbb{N}}$ such that*

$$\lim_{n\to\infty}\frac{\eta_n}{n^\alpha} = \eta_0 > 0$$

and for $M_{in} := M_{in}^{\delta_n}$,

$$1 - \mathbb{P}_\beta(M_{in}^{\delta_n}) \leq \exp(-\eta_n), \quad \textit{for all } \beta \in \mathcal{B}, \qquad (4.10)$$

$$1 - \mathbb{P}(M_{in}^{\delta_n}) \leq \exp(-\eta_n). \qquad (4.11)$$

for sufficiently large n.

Theorem 4.2. *Let $\delta_n = n^\alpha$ with $0 < \alpha < 1/6$. Then, for $i = 1, \ldots, N$,*

$$\lim_{n\to\infty} nF_n^{-1}(\tilde{\beta}_{in}) = F^{-1}(B_i), \qquad (4.12)$$

$$\lim_{n\to\infty} n\,\mathrm{Cov}(\tilde{\beta}_{in} \mid B_i) = F^{-1}(B_i), \qquad (4.13)$$

where both convergences are in $\mathcal{L}^1$-sense. The limiting function F comes from the convergence condition (R5).

The proof, which is mainly technical, will follow soon. Equations (4.12) and (4.13) claim that the inverse Fisher information matrix and the conditional covariance matrix converge to the same limit with the same speed $\frac{1}{n}$. Especially, they are asymptotically equivalent in probability, i.e.

$$F_n^{-1}(\tilde{\beta}_{in}) \overset{n}{\sim} \mathrm{Cov}(\tilde{\beta}_{in} \mid B_i).$$

Recall Theorem 3.12, where we have chosen

$$\hat{S} = \frac{1}{N}\sum_{i=1}^{N} F_n^{-1}(\tilde{\beta}_{in})$$

as the estimator of $S = \mathbb{E}[\mathrm{Cov}(\tilde{\beta}_{in} \mid B_i)]$. We have motivated this choice by the asymptotic normality of $\tilde{\beta}_{in}$, where $F_n^{-1}(\tilde{\beta}_{in})$ was the appropriate scaling matrix. However, asymptotic statements in the sense of $\mathcal{L}^1$ or $\mathbb{P}$ could not be provided besides both being $\mathcal{L}^1$ and $\mathbb{P}$-null sequences. This is strongly connected to the convergence speed of $1 - \mathbb{P}(M_{in})$ which is in general slower than $\frac{1}{n}$. Until now, all proofs concerning the PMLE were based on the fact that quantities were negligible on the complement of M_{in}. To mention a particular usage in the proof of Theorem 4.2, we want

$$\begin{aligned}\mathbb{E}[\|s_{in}(\tilde{\beta}_{in})\|] &= \mathbb{E}[\|s_{in}(\hat{\beta}_{in})\mathbb{1}_{M_{in}} + s_{in}(0)\mathbb{1}_{M_{in}^c}\|] \\ &= \mathbb{E}[\|s_{in}(0)\mathbb{1}_{M_{in}^c}\|] \\ &\leq \sqrt{\mathbb{P}(M_{in}^c)\mathbb{E}[\|s_{in}(0)\|^2]}.\end{aligned}$$

to vanish. As we will see later, $\mathbb{E}[\|s_{in}(0)\|^2] \leq cn^2$ for some $c > 0$ so that

$$\mathbb{E}[\|s_{in}(\tilde{\beta}_{in})\|] \leq \sqrt{\mathbb{P}(M_{in}^c)}\sqrt{c}n.$$

In the case of a general simple exponential family, we have $n\sqrt{\mathbb{P}(M_{in}^c)} \to \infty$ but an exponentially fast decay of $\mathbb{P}(M_{in}^c)$ solves this issue.

Proof of Theorem 4.2. The proof holds for all clusters $i = 1, \ldots, N$ and the corresponding index i is omitted. First recall Lemma 3.14 which has provided uniform convergence of the Fisher information matrix, i.e.

$$\lim_{n\to\infty} \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n} F_n(\beta) - F(\beta) \right\| = 0. \tag{4.14}$$

This property can be easily extended to the uniform convergence of the inverse functions.

Lemma 4.3. *It holds that*

$$\lim_{n\to\infty} \sup_{\beta\in\mathcal{B}} \left\| nF_n^{-1}(\beta) - F^{-1}(\beta) \right\| = 0. \tag{4.15}$$

Furthermore, there exist a $c_F > 0$ and an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\sup_{\beta\in\mathcal{B}} \|F_n^{-1}(\beta)\| \leq \frac{1}{n} c_F. \tag{4.16}$$

PROOF. Let $\epsilon > 0$. Since matrix inversion is a continuous operation and $\mathcal{B}$ is compact, there exists a $\delta > 0$ such that for all $\beta \in \mathcal{B}$,

$$\left\| \frac{1}{n} F_n(\beta) - F(\beta) \right\| < \delta \Rightarrow \left\| n F_n^{-1}(\beta) - F^{-1}(\beta) \right\| < \epsilon.$$

By (4.14), the δ-condition is always satisfied if n is chosen sufficiently large so that (4.15) follows.

For the second claim, let $\beta \in \mathcal{B}$. We have

$$\|F_n^{-1}(\beta)\| = \lambda_{\max}(F_n^{-1}(\beta)) = \lambda_{\min}^{-1}(F_n(\beta)),$$

where $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and the smallest eigenvalue, respectively. We have already seen in the proof of Theorem 3.2 that there exist constants $d > 0$ and $n_0 \in \mathbb{N}$ such that

$$\lambda_{\min}\left(\frac{F_n(\beta)}{n} \right) \geq d > 0, \quad n \geq n_0.$$

These constants do not depend on β. Thus, by putting $c_F = d^{-1}$,

$$\lambda_{\min}^{-1}(F_n(\beta)) = \left(n \lambda_{\min}\left(\frac{F_n(\beta)}{n} \right) \right)^{-1} \leq \frac{1}{n} d^{-1} = \frac{1}{n} c_F$$

for all $n \geq n_0$. □

Remark. Since the Cholesky square root $F_n^{-T/2}(\beta)$ of $F_n^{-1}(\beta)$, which is defined through the identity

$$F_n^{-1}(\beta) = F_n^{-T/2}(\beta) F_n^{-1/2}(\beta),$$

satisfies

$$\|F_n^{-1}(\beta)\| = \|F_n^{-T/2}(\beta)\|^2 = \|F_n^{-1/2}(\beta)\|^2,$$

the above lemma can be extended to the norm of the square root, i.e.

$$\sup_{\beta \in \mathcal{B}} \left\| F_n^{-T/2}(\beta) \right\| \leq \frac{1}{\sqrt{n}} \sqrt{c_F} \tag{4.17}$$

for sufficiently large n.

Lemma 4.3 shows that the inverse Fisher information matrix behaves like n^{-1} on $\mathcal{B}$. Moreover, F_n^{-1} is almost constant on sufficiently small neighborhoods as the next lemma reveals. For that purpose, we define

$$V_n(\beta, B) := F_n^{-1/2}(B) F_n(\beta) F_n^{-T/2}(B).$$

Lemma 4.4. *There exist a constant $c_V > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,*

$$\sup_{\beta \in N_n(\delta_n, B)} \|V_n(\beta, B) - I_p\| \leq c_V \frac{\delta_n}{\sqrt{n}}, \quad \text{a.s.}$$

PROOF. Recall that F_n/n is Lipschitz continuous on $\mathcal{B}$ with some Lipschitz constant $L > 0$ that does not depend on n. Moreover, let c_F and n_0 as in Lemma 4.3 and choose

$c_V = c_F L$. Then, we have for all $\beta \in N_n(\delta, B)$ and $n \geq n_0$,

$$\begin{aligned}
\|V_n(\beta, B) - I_p\| &= \left\| F_n^{-1/2}(B) F_n(\beta) F_n^{-T/2}(B) - I_p \right\| \\
&= \left\| F_n^{-1/2}(B) \left(F_n(\beta) - F_n(B) \right) F_n^{-T/2}(B) \right\| \\
&\leq n \left\| F_n^{-1/2}(B) \right\| \left\| \frac{1}{n} \left(F_n(\beta) - F_n(B) \right) \right\| \left\| F_n^{-T/2}(B) \right\| \\
&\leq n \frac{c_F}{n} L \|\beta - B\| \\
&\leq c_V \frac{\delta_n}{\sqrt{n}}
\end{aligned}$$

which proves the claim. □

Since $\delta_n / \sqrt{n} \to 0$ for $\delta_n = n^\alpha$, it immediately follows from Lemma 4.4 that V_n converges uniformly to I_p on $N_n(\delta_n, B)$. Moreover, the convergence speed does not depend on B. Similar to the uniform convergence of nF_n^{-1}, this convergence can be extended to that of V_n^{-1} towards I_p. Again, its convergence speed does not depend on B. We can now proof (4.12), i.e.

$$\lim_{n \to \infty} n F_n^{-1}(\tilde{\beta}_n) \overset{\mathcal{L}^1}{=} F^{-1}(B).$$

Let $\epsilon > 0$ and n sufficiently large. By Lemma 3.14, Lemma 4.3 and Lemma 4.4, we get

$$\begin{aligned}
\left\| n F_n^{-1}(\tilde{\beta}_n) - F^{-1}(B) \right\| &\leq n \left\| F_n^{-1}(\tilde{\beta}_n) - F_n^{-1}(B) \right\| + \left\| n F_n^{-1}(B) - F^{-1}(B) \right\| \\
&= n \left\| F_n^{-1}(\tilde{\beta}_n) - F_n^{-1}(B) \right\| \left(\mathbb{1}_{M_n} + \mathbb{1}_{M_n^c} \right) + \left\| n F_n^{-1}(B) - F^{-1}(B) \right\| \\
&\leq n \left\| F_n^{-T/2}(B) \left(V_n^{-1}(\hat{\beta}_n, B) - I_p \right) F_n^{-1/2}(B) \right\| \mathbb{1}_{M_n} \\
&\quad + n \left\| F_n^{-1}(0) - F_n^{-1}(B) \right\| \mathbb{1}_{M_n^c} + \frac{\epsilon}{2} \\
&\leq n \left\| F_n^{-1}(B) \right\| \left\| V_n^{-1}(\hat{\beta}_n, B) - I_p \right\| \mathbb{1}_{M_n} \\
&\quad + n \left\| F_n^{-1}(0) - F_n^{-1}(B) \right\| \mathbb{1}_{M_n^c} + \frac{\epsilon}{2} \\
&\leq n \frac{c_F}{n} \frac{\epsilon}{2 c_F} \mathbb{1}_{M_n} + 2 c_F \mathbb{1}_{M_n^c} + \frac{\epsilon}{2}
\end{aligned}$$

almost surely. Thus,

$$\begin{aligned}
\mathbb{E}\left[\left\| n F_n^{-1}(\tilde{\beta}_n) - F^{-1}(B) \right\| \right] &\leq \frac{\epsilon}{2} \mathbb{P}(M_n) + 2 c_F \mathbb{P}(M_n^c) + \frac{\epsilon}{2} \\
&\overset{n \to \infty}{\longrightarrow} \epsilon.
\end{aligned}$$

Since $\epsilon > 0$ is arbitrary, the claim follows by taking $\epsilon \downarrow 0$.

One similarly proves the second claim (4.13), which was

$$\lim_{n \to \infty} n \operatorname{Cov}(\tilde{\beta}_n \mid B) \overset{\mathcal{L}^1}{=} F^{-1}(B).$$

In fact, for $\epsilon > 0$ and n large enough,

$$\begin{aligned}
&\left\| n \operatorname{Cov}(\tilde{\beta}_n \mid B) - F^{-1}(B) \right\| \\
\leq\ &\left\| n F_n^{-T/2}(B) \operatorname{Cov}(F_n^{T/2}(B)\tilde{\beta}_n \mid B) F_n^{-1/2}(B) - F^{-1}(B) \right\| \\
\leq\ &\left\| \sqrt{n} F_n^{-T/2}(B) \left(\operatorname{Cov}(F_n^{T/2}(B)\tilde{\beta}_n \mid B) - I_p \right) \sqrt{n} F_n^{-1/2}(B) \right\| \\
&\qquad + \left\| n F_n^{-1}(B) - F^{-1}(B) \right\| \\
\leq\ &n \left\| F_n^{-1}(B) \right\| \left\| \operatorname{Cov}(F_n^{T/2}(B)\tilde{\beta}_n \mid B) - I_p \right\| + \left\| n F_n^{-1}(B) - F^{-1}(B) \right\| \\
\leq\ &c_F \left\| \operatorname{Cov}(F_n^{T/2}(B)(\tilde{\beta}_n - B) \mid B) - I_p \right\| + \frac{\epsilon}{2}
\end{aligned}$$

almost surely. It remains to show that the norm on the last line vanishes on average. We have,

$$\begin{aligned}
&\left\| \operatorname{Cov}(F_n^{T/2}(B)(\tilde{\beta}_n - B) \mid B) - I_p \right\| \\
\leq &\left\| \operatorname{Cov}(F_n^{T/2}(B)(\tilde{\beta}_n - B) \mid B) - \operatorname{Cov}(V_n(\xi, B) F_n^{T/2}(B)(\tilde{\beta}_n - B) \mid B) \right\| \\
&\quad + \left\| \operatorname{Cov}(V_n(\xi, B) F_n^{T/2}(B)(\tilde{\beta}_n - B) \mid B) - I_p \right\|, \qquad (4.18)
\end{aligned}$$

where ξ lies on the line segment connecting $\tilde{\beta}_n$ and B. The ξ results from the Taylor expansion of s_n, that is

$$s_n(B) - s_n(\tilde{\beta}) = F_n(\xi)(\tilde{\beta}_n - B).$$

The remaining steps can be roughly summarized as follows. Since $V_n(\xi, B)$ converges to I_p, see Lemma 4.4, the first summand on the right hand side of (4.18) vanishes. The second summand vanishes as $F_n^{-1}(B)$ is, conditional on B, the asymptotic covariance matrix of $\tilde{\beta}_n$.

More specifically, we have

$$\begin{aligned}
V_n(\xi, B) F_n^{T/2}(B)(\tilde{\beta}_n - B) &= V_n(\xi, B) F_n^{T/2}(B) F_n^{-1}(\xi)(s_n(B) - s_n(\tilde{\beta}_n)) \\
&= F_n^{-1/2}(B) F_n(\xi) F_n^{-T/2}(B) F_n^{T/2}(B) F_n^{-1}(\xi)(s_n(B) - s_n(\tilde{\beta}_n)) \\
&= F_n^{-1/2}(B)(s_n(B) - s_n(\tilde{\beta}_n)).
\end{aligned}$$

We use the almost sure identities

$$\begin{aligned}
\operatorname{Cov}(s_n(B) \mid B) &= F_n(B), \\
\mathbb{E}[s_n(B) \mid B] &= 0, \\
s_n(\hat{\beta}_n) &= 0
\end{aligned}$$

to obtain

$$
\begin{aligned}
&\mathrm{Cov}(V_n(\xi,B)F_n^{T/2}(B)(\tilde\beta_n - B) \mid B) - I_p \\
&= \mathrm{Cov}(F_n^{-1/2}(B)(s_n(B) - s_n(\tilde\beta_n)) \mid B) - I_p \\
&= F_n^{-1/2}(B)\Big(\mathbb{E}[(s_n(B) - s_n(\tilde\beta_n))(s_n(B) - s_n(\tilde\beta_n))' \mid B] \\
&\qquad -\mathbb{E}[s_n(B) - s_n(\tilde\beta_n) \mid B]\mathbb{E}[s_n(B) - s_n(\tilde\beta_n) \mid B]' - F_n(B)\Big) F_n^{-T/2}(B) \\
&= F_n^{-1/2}(B)\Big(\mathbb{E}[s_n(\tilde\beta_n)s_n(\tilde\beta_n)' - s_n(B)s_n(\tilde\beta_n)' - s_n(\tilde\beta_n)s_n(B)' \mid B] \\
&\qquad -\mathbb{E}[s_n(\tilde\beta_n) \mid B]\mathbb{E}[s_n(\tilde\beta_n) \mid B]'\Big) F_n^{-T/2}(B) \\
&= F_n^{-1/2}(B)\Big(\mathbb{E}[(s_n(\hat\beta_n)s_n(\hat\beta_n)' - s_n(B)s_n(\hat\beta_n)' - s_n(\hat\beta_n)s_n(B)')\mathbb{1}_{M_n} \mid B] \\
&\qquad -\mathbb{E}[s_n(\hat\beta_n)\mathbb{1}_{M_n} \mid B]\mathbb{E}[s_n(\hat\beta_n)\mathbb{1}_{M_n} \mid B]'\Big) F_n^{-T/2}(B) \\
&\quad + F_n^{-1/2}(B)\left(\mathbb{E}[(s_n(0)s_n(0)' - s_n(B)s_n(0)' - s_n(0)s_n(B)')\mathbb{1}_{M_n^c} \mid B]\right. \\
&\qquad \left.-\mathbb{E}[s_n(0)\mathbb{1}_{M_n^c} \mid B]\mathbb{E}[s_n(0)\mathbb{1}_{M_n^c} \mid B]'\right) F_n^{-T/2}(B) \\
&= F_n^{-1/2}(B)\left(\mathbb{E}[(s_n(0)s_n(0)' - s_n(B)s_n(0)' - s_n(0)s_n(B)')\mathbb{1}_{M_n^c} \mid B]\right. \\
&\qquad \left.-\mathbb{E}[s_n(0)\mathbb{1}_{M_n^c} \mid B]\mathbb{E}[s_n(0)\mathbb{1}_{M_n^c} \mid B]'\right) F_n^{-T/2}(B)
\end{aligned}
$$

Therefore, the second summand on the right hand side of (4.18) satisfies

$$
\begin{aligned}
&\Big\|\mathrm{Cov}(V_n(\xi,B)F_n^{T/2}(B)(\tilde\beta_n - B) \mid B) - I_p\Big\| \\
&\le \big\|F_n^{-1}(B)\big\|\left(\mathbb{E}[(\|s_n(0)s_n(0)'\| + 2\|s_n(B)s_n(0)'\|)\mathbb{1}_{M_n^c} \mid B] + \mathbb{E}[\|s_n(0)\|\mathbb{1}_{M_n^c} \mid B]^2\right) \\
&\le \big\|F_n^{-1}(B)\big\|\left(\sqrt{\mathbb{E}[\mathbb{1}_{M_n^c} \mid B]}\left(\sqrt{\mathbb{E}[\|s_n(0)s_n(0)'\|^2 \mid B]} + 2\sqrt{\mathbb{E}[\|s_n(B)s_n(0)'\|^2 \mid B]}\right)\right. \\
&\qquad \left.+ \mathbb{E}[\mathbb{1}_{M_n^c} \mid B]\mathbb{E}[\|s_n(0)\|^2 \mid B]\right). \qquad (4.19)
\end{aligned}
$$

As stated in Theorem 4.1, the conditional probability of M_n^c is almost surely bounded by an exponentially decaying term $\exp(-\eta_n)$. Moreover, one easily shows that all the other conditional expectations in (4.19) are bounded by terms of polynomial order. For instance, the Cauchy-Schwarz inequality gives that

$$
\begin{aligned}
\mathbb{E}\left[\|s_n(0)\|^2 \mid B\right] &= \mathbb{E}\left[\sum_{k=1}^{p}\left(\sum_{j=1}^{n} X_{jk}(Y_j - \mu_j(0))\right)^2 \mid B\right] \\
&\le \sum_{k=1}^{p}\mathbb{E}\left[\left(\sum_{j=1}^{n} X_{jk}^2\right)\left(\sum_{j=1}^{n}(Y_j - \mu_j(0))^2\right) \mid B\right] \\
&\le pc_X^2 n\sum_{j=1}^{n}\mathbb{E}[(Y_j - \mu_j(0))^2 \mid B] \\
&= pc_X^2 n\sum_{j=1}^{n}(v_j(B) + \mu_j(B)^2 - 2\mu_j(B)\mu_j(0) + \mu_j(0)^2) \\
&\le \mathrm{const}\cdot n^2.
\end{aligned}
$$

Such a constant exists since v and μ are bounded by the compactness of $\mathcal{B}$ and the admissible covariate space. Other upper bounds similarly follow.

To complete the proof, it remains to study the first summand on the right hand side of (4.18). We have,

$$
\begin{aligned}
&\left\|\mathrm{Cov}(F_n^{T/2}(B)(\tilde\beta_n - B) \mid B) - \mathrm{Cov}(V_n(\xi,B)F_n^{T/2}(B)(\tilde\beta_n - B) \mid B)\right\| \\
\leq\ &\Big\|\mathbb{E}\left[F_n^{T/2}(B)(\tilde\beta_n - B)(\tilde\beta_n - B)'F_n^{1/2}(B) \mid B\right] \\
&\quad -\mathbb{E}\left[V_n(\xi,B)F_n^{T/2}(B)(\tilde\beta_n - B)(\tilde\beta_n - B)'F_n^{1/2}(B)V_n'(\xi,B) \mid B\right]\Big\| \\
&\quad + \left.\begin{aligned}&\Big\|-\mathbb{E}\left[F_n^{T/2}(B)(\tilde\beta_n - B) \mid B\right]\mathbb{E}\left[(\tilde\beta_n - B)'F_n^{1/2}(B) \mid B\right] \\ &\quad +\mathbb{E}\left[V_n(\xi,B)F_n^{T/2}(B)(\tilde\beta_n - B) \mid B\right]\mathbb{E}\left[(\tilde\beta_n - B)'F_n^{1/2}(B)V_n'(\xi,B) \mid B\right]\Big\|\end{aligned}\right\} =: \mathrm{V} \\
\leq\ &\Big\|\mathbb{E}\left[(I_p - V_n(\xi,B))F_n^{T/2}(B)(\tilde\beta_n - B)(\tilde\beta_n - B)'F_n^{1/2}(B) \mid B\right]\Big\| \\
&\quad + \Big\|\mathbb{E}\left[V_n(\xi,B)F_n^{T/2}(B)(\tilde\beta_n - B)(\tilde\beta_n - B)'F_n^{1/2}(B)(I_p - V_n(\xi,B))' \mid B\right]\Big\| + \mathrm{V} \\
\leq\ &\mathbb{E}\left[\|I - V_n(\xi,B))\|\,\|F_n(B)\|\,\left\|\tilde\beta_n - B\right\|^2 \mid B\right] \\
&\quad + \mathbb{E}\left[\|V_n(\xi,B)\|\,\|F_n(B)\|\,\left\|\tilde\beta_n - B\right\|^2\|I_p - V_n(\xi,B)\| \mid B\right] + \mathrm{V} \\
\leq\ &\mathbb{E}\left[\|I_p - V_n(\xi,B))\|\,\|F_n(B)\|\,\left\|\hat\beta_n - B\right\|^2 \mathbb{1}_{M_n} \mid B\right] \\
&\quad + \mathbb{E}\left[\|I_p - V_n(\xi,B))\|\,\|F_n(B)\|\,\|B\|^2\,\mathbb{1}_{M_n^c} \mid B\right] \\
&\quad + \mathbb{E}\left[\|V_n(\xi,B)\|\,\|F_n(B)\|\,\left\|\hat\beta_n - B\right\|^2\|I_p - V_n(\xi,B)\|\,\mathbb{1}_{M_n} \mid B\right] \\
&\quad + \mathbb{E}\left[\|V_n(\xi,B)\|\,\|F_n(B)\|\,\|B\|^2\left\|I_p - V_n(\xi,B)\mathbb{1}_{M_n^c}\right\| \mid B\right] + \mathrm{V} \\
=:\ &\mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V}.
\end{aligned}
$$

If M_n occurs, both $\hat\beta_n$ and ξ lie in $N_n(\delta_n, B)$. Thus, by Lemma 4.4,

$$\|I_p - V_n(\xi,B))\| < c_V\frac{\delta_n}{\sqrt{n}}, \quad \text{a.s.}$$

In addition, by the construction of $N_n(\delta_n, B)$,

$$\|\hat\beta_n - B\| \leq \frac{\delta_n}{\sqrt{n}}, \quad \text{a.s.}$$

and there exists a constant $c_F > 0$ with

$$\|F_n(B)\| \leq nc_F$$

for sufficiently large n. Hence, with $\delta_n = n^\alpha$ and some constant $c > 0$,

$$\begin{aligned}\mathrm{I} &= \mathbb{E}\left[\|I_p - V_n(\xi, B))\| \, \|F_n(B)\| \left\|\hat{\beta}_n - B\right\|^2 \mathbb{1}_{M_n} \mid B\right] \\ &\leq c\frac{\delta_n}{\sqrt{n}} n \frac{\delta_n^2}{n} \mathbb{P}(M_n \mid B) \\ &\leq c n^{3\alpha - \frac{1}{2}}\end{aligned}$$

Since $\alpha < \frac{1}{6}$, the right hand side vanishes as $n \to \infty$ and so does $\mathbb{E}[\mathrm{I}]$. By the same arguments, $\mathbb{E}[\mathrm{III}]$ converges to zero. On the other hand, if M_n^c occurs, the components $\|I_p - V_n(\xi, B)\|$, $\|F_n(B)\|$ and $\|B\|$ are either bounded or grow at most at polynomial speed. Since the decay of $\mathbb{P}(M_n^c)$ is of exponential order, the summands $\mathbb{E}[\mathrm{II}]$ and $\mathbb{E}[\mathrm{IV}]$ converge to zero. Assuming that $\alpha < 1/4$, the same procedure works with the first moments denoted by $\mathbb{E}[\mathrm{V}]$, too. Altogether, we have shown that

$$\lim_{n\to\infty} \left\|\mathbb{E}[n \operatorname{Cov}(\tilde{\beta}_n \mid B) - F^{-1}(B)]\right\| = 0$$

which completes the proof.

4.2. Grouped data

The particular case of grouped data concerns the structure of the covariates rather than the distribution of the data. Let

$$\mathcal{A} := \{X_j \in \mathbb{R}^{1,p} : j \in \mathbb{N}\}$$

bet the set of admissible covariates. The data is called *grouped* if $\mathcal{A}$ is finite with $g = |\mathcal{A}|$ being the number of groups. The interpretation is that each observation can be classified into one of the groups $l = 1, \ldots, g$, where all elements of group l share the same covariate $X_{(l)} \in \mathcal{A}$. Hence,

$$\mathcal{A} = \{X_{(l)} \in \mathbb{R}^{1,p} : l = 1, \ldots, g\}.$$

Of immense importance in application is categorical data, where p is the number of categories. Then, for all $j \in \mathbb{N}$ and $k = 1, \ldots, p$,

$$X_{jk} = \begin{cases} 1 & \text{if } j \text{ belongs to category } k, \\ 0 & \text{else.} \end{cases}$$

By construction, there are at most 2^p groups. Whenever g is finite, we can find a constant $c_X > 0$ such that

$$\|X_{(l)}\| \leq c_X \quad \text{and} \quad |X_{(l)k}| \leq c_X \tag{4.20}$$

for all $l = 1, \ldots, g$ and components $k = 1, \ldots, p$. Furthermore, for every $n \in \mathbb{N}$, we let

$$n_l := |\{j \leq n : X_j = X_{(l)}\}|, \quad l = 1, \ldots, g, \tag{4.21}$$

be the number of the first n observations that belong to the l-th group.

The construction of the PMLE was based on the regularity assumptions (R1) to (R5) so far, see also Section 3.2. The grouped structure provides finiteness of $\mathcal{A}$ even though (R3) only demands its boundedness. We are able to relax some of the other conditions in return and in particular, we can weaken (R1) by dropping the compactness of $\mathcal{B}$. It may be replaced by some mild integrability conditions for B_i. The regularity assumptions for grouped data look as follows.

(G1a) The support $\mathcal{B}$ of the B_i is contained in an open and convex set and

$$\mathbb{P}(B_i = 0) = 0, \tag{4.22}$$

$$\mathbb{P}(B_i \in \partial\mathcal{B}) = 0. \tag{4.23}$$

(G1b) Let v be the variance function associated to the underlying exponential family and let $v_{(l)}(B_i) = v(X_{(l)}B_i)$. Then, we have

$$\mathbb{E}[v_{(l)}(B_i)] < \infty, \quad l = 1, \dots, g, \tag{4.24}$$

$$\text{and} \qquad \mathbb{E}[v_{(l)}(B_i)^{-2}] < \infty, \quad l = 1, \dots, g_e. \tag{4.25}$$

(G1c) B_i is square integrable, i.e.

$$\mathbb{E}[\|B_i\|^2] < \infty.$$

(G2) The link function g is twice continuously differentiable with non-singular Jacobian.

(G3) The number of admissible covariates is finite and all elements satisfy $X_{(l)}\beta \in \Theta^0$ for all $\beta \in \mathcal{B}$.

(G4) There exist $g_e \geq p$ *essential* groups, i.e. they satisfy

$$\frac{n_l}{n} \overset{n\to\infty}{\longrightarrow} \alpha_l > 0.$$

Moreover, the corresponding design matrix $X_e \in \mathbb{R}^{g_e,p}$ that consists of the essential groups' covariates has full rank p.

(G5) The scaled Fisher information matrix $F_n(\beta)/n$ converges pointwise to a positive definite limit $F(\beta)$ for all $\beta \in \mathcal{B}$.

The moment conditions (G1b) and (G1c) are easy to evaluate and allow a wide variety of distributions for B_i. For instance in the Poisson case, v is the exponential function and both (4.24) and (4.25) are equivalent to the existence of the moment generating function of B_i at the values $X_{(l)}$ and $-2X_{(l)}$ respectively. Hence, common distributions as the multivariate Gaussian is admissible for B_i. One may also notice that compactness of $\mathcal{B}$ is sufficient for (G1b) and (G1c). Conditions (G2) and (R2) as well as the convergence conditions (G5) and (R5) are both identical. As we have already mentioned, assumption (G3) defines the grouped structure and replaces (R3). The second part of condition (G4) implies the rank condition (R4). Moreover, (G4) seems to be strongly related to (G5) since

$$\frac{1}{n}F_n(\beta) = \sum_{l=1}^{g} \frac{n_l}{n} v_{(l)}(\beta) X_{(l)}' X_{(l)}.$$

However, due to the presence of non-essential groups, (G4) is neither sufficient nor necessary. If $g_e = g$, i.e. if all groups are essential, (G4) will be sufficient and we will obtain an explicit structure for the limiting function F that is

$$F(\beta) = \sum_{l=1}^{g} \alpha_l v_{(l)}(\beta) X_{(l)}' X_{(l)}.$$

Given these modified regularity assumptions, we reinvestigate the construction of the PMLE on the basis of Theorem 3.2.

Theorem 4.5. *Assume that assumptions (G1) to (G5) and*

$$\mathbb{E}\left[\sup_{\xi\in N_n(\delta_n,B_i)} v_{(l)}(\xi)^{-2}\right] < \infty, \quad l = 1,\dots,g, \tag{4.26}$$

hold. Then, we can find a sequence $(\delta_n)_{n\in\mathbb{N}}$ *such that*

$$\mathbb{P}(M_{in}^{\delta_n}) \overset{n\to\infty}{\longrightarrow} 1$$
$$\text{and} \qquad N_n(\delta_n, B_i) \overset{n\to\infty}{\longrightarrow} \{B_i\}, \quad a.s.,$$

where the involved quantities have been defined in (3.10) *and* (3.11).

Proof. The proof works similarly to that of Theorem 3.2. An intermediate result was (3.36) which stated that

$$1 - \mathbb{P}(M_{in}^{\delta_n}) \leq \mathbb{P}\left(\|s_{in}(B_i)\| \geq \frac{1}{2}\sqrt{n}\delta_n d\right).$$

Compactness of $\mathcal{B}$ ensured that there exists a constant $d > 0$ with

$$\min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w \geq d$$

for sufficiently large n and all $\xi \in N_n(\delta_n, B_i)$ coming from the Taylor expansion of the log-likelihood function l_{in}. Instead, we now use

$$\min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w \geq \inf_{\xi\in N_n(\delta_n,B_i)} \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w > 0$$

so that (3.36) changes to

$$\begin{aligned}
1 - \mathbb{P}(M_{in}^{\delta_n}) &\leq \mathbb{P}\left(\|s_{in}(B_i)\| \geq \frac{1}{2}\sqrt{n}\delta_n \left(\inf_{\xi\in N_n(\delta_n,B_i)} \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w\right)\right) \\
&\leq \mathbb{P}\left(\|s_{in}(B_i)\| \left(\inf_{\xi\in N_n(\delta_n,B_i)} \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w\right)^{-1} \geq \frac{1}{2}\sqrt{n}\delta_n\right) \\
&\leq \frac{2}{\sqrt{n}\delta_n}\mathbb{E}\left[\|s_{in}(B_i)\| \left(\inf_{\xi\in N_n(\delta_n,B_i)} \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w\right)^{-1}\right]
\end{aligned}$$

for all $n \in \mathbb{N}$. The last step follows by Markov's inequality. Applying Cauchy-Schwarz's inequality on the last term yields

$$1 - \mathbb{P}(M_{in}^{\delta_n}) \leq \frac{2}{\sqrt{n}\delta_n}\left(\mathbb{E}\left[\|s_{in}(B_i)\|^2\right]\mathbb{E}\left[\left(\inf_{\xi\in N_n(\delta_n,B_i)} \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w\right)^{-2}\right]\right)^{\frac{1}{2}} \tag{4.27}$$

and we consider both expectations separately. By (4.20), we have

$$\begin{aligned}\mathbb{E}\left[\|s_{in}(B_i)\|^2\right] &= \sum_{k=1}^{p}\sum_{j=1}^{n} X_{jk}^2 \mathbb{E}[\mathrm{Var}(Y_{ij} \mid B_i)] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{g} n_l X_{(l)k}^2 \mathbb{E}[v_{(l)}(B_i)] \\ &\leq p c_X^2 \sum_{l=1}^{g} n_l \mathbb{E}[v_{(l)}(B_i)] \\ &\leq p c_X^2 n \sum_{l=1}^{g} \mathbb{E}[v_{(l)}(B_i)].\end{aligned}$$

By assumption (G1b), $\mathbb{E}[v_{(l)}(B_i)]$ exists for all groups $l = 1, \ldots, g$. Finally, the $\sqrt{n}$ on the right hand side of (4.27) cancels out. It remains to analyze the second expectation in (4.27). Let $\xi \in N_n(\delta_n, B_i)$ and $w \in \mathbb{R}^p$ with $\|w\| = 1$. Let without loss of generality the groups $l = 1, \ldots, g_e$ be essential. By (G4), we can choose $0 < \epsilon < \min_l \alpha_l$ and $\alpha > 0$ such that for sufficiently large n,

$$\frac{n_l}{n} \geq \alpha_l - \epsilon \geq \alpha > 0, \quad l = 1, \ldots, g_e.$$

Therefore,

$$\begin{aligned}w' \frac{F_n(\xi)}{n} w &= w' \left(\sum_{j=1}^{n} \frac{1}{n} v_j(\xi) X_j' X_j \right) w \\ &= \sum_{l=1}^{g} \frac{n_l}{n} v_{(l)}(\xi) w' X_{(l)}' X_{(l)} w \\ &\geq \sum_{l=1}^{g_e} \frac{n_l}{n} v_{(l)}(\xi) \left(\sum_{k=1}^{p} w_k X_{(l)k} \right)^2 \\ &\geq \alpha \sum_{l=1}^{g_e} v_{(l)}(\xi) \left(\sum_{k=1}^{p} w_k X_{(l)k} \right)^2 \\ &\geq \alpha \left(\min_{l=1,\ldots,g_e} v_{(l)}(\xi) \right) \sum_{l=1}^{g_e} \left(\sum_{k=1}^{p} w_k X_{(l)k} \right)^2.\end{aligned}$$

Thus,

$$\min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w \geq \alpha d \min_{l=1,\ldots,g_e} v_{(l)}(\xi)$$

with

$$d := \min_{\|w\|=1} \sum_{l=1}^{g_e} \left(\sum_{k=1}^{p} w_k X_{(l)k} \right)^2.$$

In order to verify that $d > 0$, we consider the matrix X_e which has been defined in condition (G4). The matrix $X_e' X_e \in \mathbb{R}^{p,p}$ is symmetric and positive semidefinite and has

an orthogonal diagonalization QDQ'. Hence,

$$\begin{aligned} d &= \min_{\|w\|=1} w' X_e' X_e w = \min_{\|w\|=1} (Q'w)' D (Q'w) \\ &= \min_{\|w\|=1} w' D w = \lambda_{\min}(X_e' X_e), \end{aligned}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue. By assumption (G4), X_e and $X_e' X_e$ have full rank p so that $d > 0$. It follows that

$$\begin{aligned} \mathbb{E}\left[\left(\inf_{\xi \in N_n(\delta_n, B_i)} \min_{\|w\|=1} w' \frac{F_n(\xi)}{n} w \right)^{-2} \right] &\leq (\alpha d)^{-2} \mathbb{E}\left[\left(\inf_{\xi \in N_n(\delta_n, B_i)} \min_{l=1,\ldots,g_e} v_{(l)}(\xi) \right)^{-2} \right] \\ &= (\alpha d)^{-2} \mathbb{E}\left[\sup_{\xi \in N_n(\delta_n, B_i)} \max_{l=1,\ldots,g_e} v_{(l)}(\xi)^{-2} \right]. \end{aligned}$$

The expectation on the right hand side exists since (4.26) implies

$$\mathbb{E}\left[\sup_{\xi \in N_n(\delta_n, B_i)} \max_{l=1,\ldots,g_e} v_{(l)}(\xi)^{-2} \right] < \sum_{l=1}^{g} \mathbb{E}\left[\sup_{\xi \in N_n(\delta_n, B_i)} v_{(l)}(\xi)^{-2} \right] < \infty.$$

Thus, (4.27) reads

$$1 - \mathbb{P}(M_{in}^{\delta_n}) \leq \frac{1}{\delta_n} \text{const}$$

and choosing (δ_n) such that

$$\delta_n \to \infty \quad \text{and} \quad \delta_n \in o(\sqrt{n})$$

proves Theorem 4.5. □

For the remaining part, we will derive sufficient conditions for (4.26), which are easy to verify. Notice that the expression $\sup_{\xi \in N_n(\delta_n, B_i)} v_{(l)}(\xi)^{-2}$ in (4.26) is indeed a measurable function. Since v is continuous by assumption (G2), we can write

$$\begin{aligned} \sup_{\xi \in N_n(\delta_n, B_i)} v(X_{(l)}\xi)^{-2} &= \sup_{\xi \in N_n(\delta_n, 0)} v(X_{(l)}(B_i + \xi))^{-2} \\ &= \sup_{\xi \in N_n(\delta_n, 0) \cap \mathbb{Q}^p} v(X_{(l)}(B_i + \xi))^{-2} \\ &= \sup_{\xi \in N_n(\delta_n, 0) \cap \mathbb{Q}^p} f_\xi(B_i), \end{aligned}$$

where each $f_\xi(B_i)$ is measurable. As a pointwise supremum of measurable functions over a countable set, the right hand side is again measurable.

Lemma 4.6. *Let the conditional distribution of (Y_{ij}) belong to either of the distributions Normal, Poisson, Gamma, Binomial or Negative Binomial, i.e. a member of the so-called natural exponential family with quadratic variance function. Then,* (4.26) *holds.*

Proof. The proof is not based on general properties of the variance functions but on their explicit structures. Therefore, we examine each distribution in the above list separately. The cluster index i of B_i will be dropped throughout the proof for notational ease.

- *Normal.* Clear, since the variance function

$$v_{(l)}(\xi) = \sigma^2$$

is constant and does not depend on ξ.

◦ *Poisson.* In the Poisson case,

$$v_{(l)}(\xi)^{-2} = \exp(-2X_{(l)}\xi).$$

Thus,

$$\begin{aligned}
\sup_{\xi\in N_n(\delta_n,B)} \exp(-2X_{(l)}\xi) &= \sup_{\xi\in N_n(\delta_n,B)} \exp(-2X_{(l)}(B+(\xi-B))) \\
&= \exp(-2X_{(l)}B) \sup_{\xi\in N_n(\delta_n,B)} \exp(-2X_{(l)}(\xi-B)) \\
&\leq \exp(-2X_{(l)}B) \sup_{\xi\in N_n(\delta_n,B)} \exp(|-2X_{(l)}(\xi-B)|) \\
&\leq \exp(-2X_{(l)}B) \sup_{\xi\in N_n(\delta_n,B)} \exp(2\|X_{(l)}\|\|\xi-B\|) \\
&= \exp(-2X_{(l)}B)\exp(2\|X_{(l)}\|\frac{\delta_n}{\sqrt{n}}).
\end{aligned}$$

Since $\exp(-2X_{(l)}B)$ is integrable by assumption (4.25), the claim follows.

◦ *Binomial.* In the Binomial case, we have

$$v_{(l)}(\xi) = m\frac{\exp(X_{(l)}\xi)}{(1+\exp(X_{(l)}\xi))^2}$$

and

$$v_{(l)}(\xi)^{-2} = m^{-2}\frac{(1+\exp(X_{(l)}\xi))^4}{\exp(2X_{(l)}\xi)},$$

where $m \in \mathbb{N}$ is a known dispersion parameter indicating the number of successive Bernoulli trials. Again, we write $X_{(l)}\xi = X_{(l)}B + X_{(l)}(\xi - B)$ so that for all $\xi \in N_n(\delta_n, B)$,

$$v_{(l)}(\xi)^{-2} = m^{-2}\frac{\big(1+\exp(X_{(l)}B)\exp(X_{(l)}(\xi-B))\big)^4}{\exp(2X_{(l)}B)\exp(2X_{(l)}(\xi-B))}.$$

The factor in the denominator satisfies

$$\exp(X_{(l)}(\xi-B)) \geq \exp(-2|X_{(l)}(\xi-B)|) \geq \exp(-2\|X_{(l)}\|\frac{\delta_n}{\sqrt{n}}).$$

For the numerator, we have

$$\begin{aligned}
&(1+\exp(X_{(l)}B)\exp(X_{(l)}(\xi-B)))^4 \\
\leq\ &\big(\exp(\|X_{(l)}\|\|\xi-B\|) + \exp(X_{(l)}B)\exp(\|X_{(l)}\|\|\xi-B\|)\big)^4 \\
\leq\ &\exp\left(4\|X_{(l)}\|\frac{\delta_n}{\sqrt{n}}\right)\big(1+\exp(X_{(l)}B)\big)^4.
\end{aligned}$$

Altogether,

$$\begin{aligned}
\mathbb{E}\left[\sup_{\xi\in N_n(\delta_n,B)} v^{-2}(X_{(l)}\xi)\right] &\leq \exp\left(2\|X_{(l)}\|\frac{\delta_n}{\sqrt{n}}\right)\mathbb{E}\left[m^{-2}\frac{(1+\exp(X_{(l)}B))^4}{\exp(2X_{(l)}B)}\right] \\
&= \exp\left(2\|X_{(l)}\|\frac{\delta_n}{\sqrt{n}}\right)\mathbb{E}[v_{(l)}(B)^{-2}]
\end{aligned}$$

and the expectation on the right hand side exists by (4.25).

∘ *Gamma.* Let $k > 0$ be a known shape parameter. Then,

$$\begin{aligned}\sup_{\xi \in N_n(\delta_n, B)} v_{(l)}(\xi)^{-2} &= \sup_{\xi \in N_n(\delta_n, B)} k^{-2}(X_{(l)}\xi)^4 \\ &\leq k^{-2} \sup_{\xi \in N_n(\delta_n, B)} \left(|X_{(l)}B| + |X_{(l)}(\xi - B)|\right)^4 \\ &\leq k^{-2} \left(|X_{(l)}B| + \|X_{(l)}\| \frac{\delta}{\sqrt{n}}\right)^4\end{aligned}$$

Since the fourth moment of $X_{(l)}B$ exists, all its lower moments exist, too.

∘ *Negative Binomial.* The variance function has the structure

$$v_{(l)}(\xi) = \frac{1}{r} \exp(2X_{(l)}\xi) + \exp(X_{(l)}\xi),$$

where $r \in \mathbb{N}$ is a known number of failures. The proof works similar to the previous cases which is why we skip it. □

The restriction to specific distributions in Lemma 4.6 is unavoidable as the lemma turns out to be generally wrong.

Example 4.7. Let the conditional distribution of Y_{ij} belong to an exponential family with

$$b(\theta) = \int_{-\infty}^{\theta} \int_{-\infty}^{t} \exp(-s^3/2) ds dt.$$

Since $v(\theta) = b''(\theta)$, we get the variance function

$$v(\theta) = \exp\left(-\frac{\theta^3}{2}\right).$$

Thus,

$$v(\theta)^{-2} = \exp(\theta^3)$$

which is monotonously increasing in θ. For simplicity, we further choose $p = 1$ and a covariate $X_{(l)} = 1$ so that $X_{(l)}B = B$. Now, let B be distributed according to the probability density function

$$f_B(b) = \begin{cases} c \exp(-b^3) b^{-2}, & b \geq 1 \\ 0, & \text{else} \end{cases}$$

where $c > 0$ is the normalizing constant. Then

$$\mathbb{E}[v(B)^{-2}] = c \int_1^{\infty} \exp(b^3) \exp(-b^3) b^{-2} db = c \int_1^{\infty} b^{-2} < \infty.$$

On the other hand, the supremum

$$\sup_{\xi \in N_n(\delta_n, B)} v(X_{(l)}\xi)^{-2}$$

is attained at the right boundary of the interval $N_n(\delta_n, B)$. Thus,

$$\sup_{\xi \in N_n(\delta_n, B)} v(X_{(l)}\xi)^{-2} = v(\xi_0)^{-2}$$

where $\xi_0 = B + \delta/\sqrt{n}$. Therefore,

$$v(\xi_0)^{-2} = \exp\left(\left(B + \frac{\delta}{\sqrt{n}}\right)^3\right).$$

But

$$\begin{aligned}\mathbb{E}[v(\xi_0)^{-2}] &= c\int_1^\infty \exp\left(\left(b+\frac{\delta}{\sqrt{n}}\right)^3\right)\exp(-b^3)b^{-2}db \\ &= c\int_1^\infty \exp\left(3b^2\frac{\delta}{\sqrt{n}}+3b\frac{\delta^2}{n}+\frac{\delta^3}{n^{3/2}}\right)b^{-2}db = \infty\end{aligned}$$

as

$$\lim_{b\to\infty}\exp\left(3b^2\frac{\delta}{\sqrt{n}}+3b\frac{\delta^2}{n}+\frac{\delta^3}{n^{3/2}}\right)b^{-2} = \infty.$$

Hence,

$$\mathbb{E}\left[\sup_{\xi\in N_n(\delta_n,B)} v(\xi)^{-2}\right] = \infty.$$

CHAPTER 5

Incorporating weights

When applying CGLM theory to real world problems one will immediately realize that portfolios are rarely homogeneous. Clusters, which may be the different regions, time periods or products, contain data of the same subject but slightly differ in some individual characteristics. They may now have individual volume parameters that are known in advance. We will discuss three major cases where the portfolio consists of N independent but not identically distributed clusters and a general framework then follows. The content of this chapter is mainly taken from Christiansen and Schinzinger (2015) but we supply a more detailed analysis in the present thesis. Furthermore, Christiansen and Schinzinger (2015) restricts to central results and coincidences will be appropriately mentioned.

5.1. Cluster specific effects

Known nuisance parameter. Recall structure (2.17) of a simple exponential family. One can extend the class of distributions by allowing for an additional nuisance or also called dispersion parameter $\lambda > 0$. Then, a random variable Y will belong to a simple exponential family if its discrete or continuous density function is given by

$$f_\theta(y) = c(y, \lambda) \exp\left(\frac{\theta y - b(\theta)}{\lambda}\right), \quad y \in \mathbb{R}. \tag{5.1}$$

Common distributions which include a non-trivial nuisance parameter $\lambda \neq 1$ are the following.

Example 5.1 (Binomial distribution). Let $Y \sim \mathrm{Bin}(m, p)$ be Binomial distributed with number of trials $m \in \mathbb{N}$ and success probability $p \in [0, 1]$. Then, not Y but the relative Binomial distribution Y/m belongs to a simple exponential family as for $k \in \mathbb{N}$ with $k \leq m$,

$$\begin{aligned}
\mathbb{P}_\theta\left(\frac{Y}{m} = \frac{k}{m}\right) &= \binom{m}{k} p^k (1-p)^{m-k} \\
&= \binom{m}{k} \exp\left(k \log p + (m-k)\log(1-p)\right) \\
&= \binom{m}{k} \exp\left(m\left(\log\left(\frac{p}{1-p}\right)\frac{k}{m} - (-\log(1-p))\right)\right).
\end{aligned}$$

Thus, the nuisance parameter equals $\frac{1}{m}$.

Example 5.2 (Gamma distribution). The Gamma distribution with mean parameter $\mu > 0$ and shape parameter $\nu > 0$ has the probability density function

$$\begin{aligned} f_\theta(y) &= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} \exp\left(-\frac{\nu}{\mu} y\right) \\ &= \frac{1}{\Gamma(\nu)} \nu^\nu y^{\nu-1} \exp\left(\nu \left(y(-\mu^{-1}) - \log \mu\right)\right) \end{aligned}$$

with $y \geq 0$. Hence, it belongs to a simple exponential family with $\theta = -\mu^{-1}$ and $\lambda = \nu^{-1}$.

In context of a CGLM, allowing cluster specific nuisance variables $\lambda_i = (\lambda_{ij})_{j=1}^n$, $i = 1, \ldots, N$, leads to a portfolio with non-identically distributed clusters. A particular example is the Binomial-CGLM, where binary responses are taken from different numbers of repetitions $m_i = (m_{ij})_{j=1}^n$. The difference may be caused by experiment j being observed in several environments in terms of time, place or group i. In the general case of simple exponential families, observations $(Y_{ij})_{j=1}^n$ are, conditional on $B_i = \beta$, independent and distributed according to (5.1). For further progression, it is more convenient to define the corresponding weights as the reciprocals $w_{ij} = \frac{1}{\lambda_{ij}}$. Likelihood function, score function and Fisher information matrix become all cluster specific, specifically

$$F_{in}(\beta) = \sum_{j=1}^n \frac{1}{\lambda_{ij}} b''(X_j\beta) X_j' X_j = \sum_{j=1}^n w_{ij} b''(X_j\beta) X_j' X_j.$$

In the univariate case $p = 1$, F_{in} is increasing in each weight w_{ij} so that the more weight is allocated to cluster i, i.e. less nuisance, the more information it contains. Credibility estimators should take these information differences into account.

Different sample sizes. In the second case, the observation vectors $Y_i = (Y_{ij})_{j=1}^{n_i}$ of the N clusters are assumed to have different lengths $n_i \in \mathbb{N}$. Then, summations in $l_{in}(\beta)$, $s_{in}(\beta)$ and $F_{in}(\beta)$ involve n_i summands. Credibility estimators for clusters with large sample sizes should allocate more weight to their individual estimates $\hat{\beta}_i$ – or to be more precise $\tilde{\beta}_i$ – than for small clusters. Different sample sizes can be incorporated in the framework of additional weighting terms by choosing them as

$$w_{ij} := \begin{cases} 1 & \text{if } j \leq n_i, \\ 0 & \text{else.} \end{cases} \tag{5.2}$$

This definition allows us to write

$$F_{in}(\beta) = \sum_{j=1}^{n_i} b''(X_j\beta) X_j' X_j = \sum_{j=1}^n w_{ij} b''(X_j\beta) X_j' X_j.$$

where $n = \max_i n_i$. This structure can be interpreted in the same way as for the nuisance parameters. Obviously, information increases with the sample size or equivalently with the amount of allocated weights.

Known offset parameter. Observations in Poisson-GLMs are claim or event counts that follow a Poisson distribution, where the log-parameters are linked to a linear predictor. In practice, data like insurance claims are observed in time periods of different lengths or the number of insured differs from period to period. These differences have to be taken into account and a natural way to do so is to involve exposure terms. More precisely, random variables Y_j, $j = 1, \ldots, n$, are said to be Poisson distributed with

Poisson parameter $\mu_j > 0$ and known exposure $E_j > 0$ if $Y_j \sim \text{Poi}(E_j\mu_j)$. Then, the linear predictors are given by

$$\log \mathbb{E}[Y_j] = \log E_j + \log \mu_j = \log E_j + X_j\beta,$$

so the terms $\log E_j$ can be treated as known offset variables. In the general case, the linear predictors may include known additive terms $\xi_j \in \mathbb{R}$, i.e.

$$\begin{aligned} g(\mathbb{E}[Y_j]) &= \xi_j + X_j\beta \\ &= \begin{pmatrix} \xi_j & X_j \end{pmatrix} \begin{pmatrix} 1 \\ \beta \end{pmatrix} \end{aligned}$$

for $j = 1, \ldots, n$. In view of the second identity, offsets can be interpreted as a component of the covariate vector whose effect on the linear predictor are known in advance. When considering a portfolio of size N, exposure terms of Poisson models or offset terms in general typically vary for each cluster. This results in an inhomogeneous portfolio in the sense that clusters are independent but not identically distributed. Special caution is required in credibility estimation since the PMLE of large-offset clusters should be treated differently than those of small-offset clusters.

5.2. The model

We summarize the three introduced cases under revised model assumptions that allow additional cluster specific parameters. The assumptions look as follows.

(W1) Conditional on $B_i = \beta_i$, the Y_{ij}, $j = 1, \ldots, n$, 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

$$f_{\beta_i}(y) = \prod_{j=1}^n c(y_j, w_{ij}) \exp\left(\sum_{j=1}^n w_{ij} \left(\theta_{ij} y_j - b(\theta_{ij})\right) \right), \quad y \in \mathbb{R}^n. \tag{5.3}$$

(W2) The natural parameters are linked to a linear predictor by the identity

$$\theta_{ij} = g(\mathbb{E}[Y_{ij} \mid B_i]) = \xi_{ij} + X_{ij}B_i, \quad \text{a.s.}, \tag{5.4}$$

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

(W3) The risk profiles $B_1, \ldots, B_N$ are iid. The pairs $(B_1, Y_1), \ldots, (B_N, Y_N)$ are independent but not necessarily identically distributed.

The model assumptions (W1) to (W3) incorporate the introduced nuisance parameters $w_{ij} = 1/\lambda_{ij}$, different sample sizes through (5.2) and offset terms in the linear predictor. Furthermore, covariate vectors X_{ij} may depend on the cluster, too. This is due to the fact that observation j is not necessarily available for every cluster of the portfolio.

The credibility model defined through the assumptions (A1) to (A3) is a particular case of the current model that incorporates weights. In fact, it can be derived by choosing all $w_{ij} = 1$, $\xi_{ij} = 0$ and $X_{ij} = X_j$. We have established credibility theory for the simpler model which now provides a fertile ground for working on CGLM based on (W1) to (W3). The first steps are completely the same. We define the maximum likelihood

estimators $\hat{\beta}_i$ as the maximizer of the function

$$l_{in}(\beta) = \sum_{j=1}^{n} w_{ij} \left(\theta_{ij} Y_{ij} - b(\theta_{ij})\right), \tag{5.5}$$

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

$$s_{in}(\beta) = \frac{\partial l_{in}(\beta)}{\partial \beta} = \sum_{j=1}^{n} w_{ij} X'_{ij} \left(Y_{ij} - \mu_{ij}(\beta)\right), \tag{5.6}$$

$$F_{in}(\beta) = \frac{\partial^2 l_{in}(\beta)}{\partial \beta \partial \beta'} = \sum_{j=1}^{n} w_{ij} v_{ij}(\beta) X'_{ij} X_{ij}. \tag{5.7}$$

The MLE solves $s_{in}(\hat{\beta}_i) = 0$. The conditional mean and variance functions are now given by

$$\mu_{ij}(B_i) = \mathbb{E}[Y_{ij} \mid B_i] = b'(\xi_{ij} + X_{ij} B_i), \tag{5.8}$$

$$w_{ij} v_{ij}(B_i) = \mathrm{Var}(Y_{ij} \mid B_i) = w_{ij} b''(\xi_{ij} + X_{ij} B_i). \tag{5.9}$$

If Y_{ij} is not observable, i.e. if $w_{ij} = 0$, its variance will be set to zero.

For credibility estimation, we instead use the PMLE $\tilde{\beta}_i$ which are the square integrable modifications of the $\hat{\beta}_i$. Their construction according to Theorem 3.2 requires a further condition in addition to the regularity conditions (R1) to (R5) as specified in Section 3.2.

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

This condition ensures that the mean and variance functions, cf. (5.8) and (5.9), remain bounded. It directly follow from (R6) and the linear growth condition (R5) that the effective sample sizes

$$n_i = \sum_{j=1}^{n} \mathbb{1}_{[w_{ij} > 0]}$$

grow linearly with n.

Lemma 5.3. *The effective sample size n_i is asymptotically bounded from both above and below by n, i.e.*

$$0 < \liminf_{n \to \infty} \frac{n_i}{n} \leq \limsup_{n \to \infty} \frac{n_i}{n} \leq 1. \tag{5.10}$$

PROOF. Since $\frac{1}{n} F_{in}(\beta)$ converges to a positive definite limit $F_i(\beta)$ and as this applies also to their norms, it follows that

$$\lim_{n \to \infty} \frac{n}{\|F_{in}(\beta)\|} = \frac{1}{\|F_i(\beta)\|}$$

for all $\beta \in \mathcal{B}$. Hence, n is asymptotically bounded by $\|F_{in}(\beta)\|$. Furthermore, it holds that

$$\begin{aligned}\|F_{in}(\beta)\| &= \left\| \sum_{j=1}^{n} w_{ij} v_{ij}(\beta) X_{ij}' X_{ij} \right\| \\ &\leq \sum_{j=1}^{n} \mathbb{1}_{[w_{ij}>0]} \left\| w_{ij} v_{ij}(\beta) X_{ij}' X_{ij} \right\| \\ &\leq V \sum_{j=1}^{n} \mathbb{1}_{[w_{ij}>0]} \\ &= V n_i\end{aligned}$$

with some $V > 0$ since all summands are bounded. The upper bound Vn_i is also an asymptotic upper bound for n and thus,

$$n \leq C n_i$$

for some $C > 0$ and sufficiently large n. The claim finally follows as

$$\liminf_{n\to\infty} \frac{n_i}{n} \geq \frac{1}{C} > 0.$$

The other inequalities in (5.10) hold by definition. □

The lemma especially implies that $n_i \to \infty$ for all i as $n \to \infty$. Furthermore, information converges uniformly on $\mathcal{B}$ as shown in Christiansen and Schinzinger (2015).

Lemma 5.4. *For all $n \in \mathbb{N}$, $\frac{1}{n} F_{in}$ is Lipschitz continuous with Lipschitz constant $L > 0$ and the sequence $(\frac{1}{n} F_{in})_{n\in\mathbb{N}}$ converges uniformly on $\mathcal{B}$, i.e.*

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

Proof. For all $\beta_1, \beta_2 \in \mathcal{B}$, the mean value theorem applied on the variance function $v_{ij}(\cdot) = b''(\xi_{ij} + X_{ij}\cdot)$ yields that

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

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. By convexity, $\gamma \in \mathcal{B}$. □

The construction of $\tilde{\beta}_i$ relies on (3.10) and (3.12), which are for $\delta > 0$ and $n \in \mathbb{N}$,

$$\begin{aligned}N_n(\delta, B_i) &= \{\beta \in \mathcal{B} : \sqrt{n}\|\beta - B_i\| \leq \delta\}, \\ M_{in}^{\delta} &= \{l_{in}(\beta) - l_{in}(B_i) < 0, \quad \text{for all } \beta \in \partial N_n(\delta, B_i)\}.\end{aligned}$$

Theorem 5.5. *Consider the model defined through (W1) to (W3) and the conditions (R1) to (R6). Then, there exists a sequence (δ_n) that satisfies*

$$\mathbb{P}(M_{in}) \overset{n\to\infty}{\longrightarrow} 1$$

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

$$N_n(\delta_n, B_i) \overset{n\to\infty}{\longrightarrow} \{B_i\} \quad a.s.$$

PROOF. The proof is almost the same as for Theorem 3.2 but mean and variance functions involve weights and offset terms, cf. proof of Lemma 5.4. They do not violate any step in the proof since condition (R6) guarantees that all bounds still exist. Christiansen and Schinzinger (2015) also presents an adjusted version of the proof. □

Theorem 5.5 defines the PMLE by means of Definition 3.3, i.e.

$$\tilde{\beta}_{in} = \hat{\beta}_{in} \mathbb{1}_{M_{in}}.$$

By construction of the sets M_{in} and since $\|B_i\| \leq c_B$ almost surely, we have

$$\|\tilde{\beta}_{in} - B_i\| \leq \frac{\delta_{in}}{\sqrt{n}} \mathbb{1}_{M_{in}} + c_B \mathbb{1}_{M_{in}^c} \quad \text{a.s.} \tag{5.11}$$

and this relation delivers the asymptotic properties of the PMLE, cf. Theorem 3.5.

Corollary 5.6. *Theorem 3.5 holds under model assumptions (W1) to (W3) and regularity conditions (R1) to (R6).*

PROOF. The claim follows by Theorem 5.5 and (5.11). The proof is exactly the same as for Theorem 3.5 and can be also found in Christiansen and Schinzinger (2015). □

Finally, the credibility estimator $\hat{B}_i$ is given in the same way as Definition 3.6. It is the orthogonal projection of B_i on $L(1, \boldsymbol{\beta})$, where

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

or equivalently, it is the minimum MSE estimator within this class. The first big change brought by the additional cluster specific quantities concerns the credibility formula, i.e. the explicit structure of $\hat{B}_i$.

Theorem 5.7. *The credibility estimator is given by*

$$\hat{B}_{in} = \mathbb{E}[B_i] + A_{in}(\tilde{\beta}_{in} - \mathbb{E}[\tilde{\beta}_{in}]) \tag{5.12}$$

with credibility matrix

$$A_{in} = \mathrm{Cov}(B_i, \tilde{\beta}_{in}) \, \mathrm{Cov}(\tilde{\beta}_{in})^{-1}. \tag{5.13}$$

Moreover, an asymptotic credibility formula is given by

$$\hat{B}_{in} \overset{n}{\sim} A_{in}\tilde{\beta}_{in} + (I_p - A_{in})\mathbb{E}[\tilde{\beta}_{in}]. \tag{5.14}$$

PROOF. The proofs are identical to these of Theorems 3.7 and 3.9. Alternatively, see Christiansen and Schinzinger (2015). □

Since the $\tilde{\beta}_i$ are not identically distributed, we obtain cluster specific credibility matrices A_{in} whose structure (5.13) equals (3.17). The result is in accordance to the considerations we have made at the beginning of the chapter. Clusters contain different amounts of information and therefore the PMLE should vary in credibility. For the particular case of different sample sizes, large sample clusters allocate more weight to their individual estimates $\tilde{\beta}_i$ since Lemma 3.10 provides that $A_{in} \to I_p$ as $n \to \infty$.

5.3. Estimation of the parameters

The structural parameters to be estimated in (5.13) and (5.14) are $\mathbb{E}[\tilde{\beta}_{in}]$, $\mathrm{Cov}(\tilde{\beta}_{in})^{-1}$ and $\mathrm{Cov}(B_i, \tilde{\beta}_{in})$. Difficulties arise since we only have one iid sample for each $\tilde{\beta}_{in}$. In the framework (A1) to (A3) of iid clusters, all N random vectors $\tilde{\beta}_{in}$ were iid. Nevertheless, the PMLE are comparable in the sense that each of them are weakly consistent estimators of the corresponding B_i. These target variables B_i are iid. Thus, cluster specific effects are less pronounced as $n \to \infty$ and this property will be frequently used in estimation. For that purpose, we use some important variables which are

$$\begin{aligned} T &:= \mathrm{Cov}(B_i), \\ \text{and} \qquad S_{in} &:= \mathbb{E}\left[\mathrm{Cov}(\tilde{\beta}_{in} \mid B_i)\right]. \end{aligned}$$

By Theorem 3.5, they satisfy the asymptotic decomposition

$$\mathrm{Cov}(\tilde{\beta}_{in}) \overset{n}{\sim} T + S_{in}, \tag{5.15}$$

which consists of a cluster common and a cluster specific term. The asymptotic equivalence is easy to see since both expressions converge to the same limit $\mathrm{Cov}(B_i)$. We first motivate the estimators on the basis of the estimation procedure in the iid case, cf. Theorem 3.12, and then discuss the necessary changes for the new setting. Readers who are mainly interested in the results may jump over to Theorem 5.9.

ad $\mathbb{E}[\tilde{\beta}_{in}]$. In the iid setting, the sample mean

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

was a strongly N-consistent and unbiased estimator for $\mathbb{E}[\tilde{\beta}_{in}]$. In the present setting, 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 on the last line almost surely vanishes by the strong law of large numbers and the fact that the B_i are iid. For the first sum, we use the dominated convergence

theorem. By (5.11), 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 because $\tilde{\beta}_{in}$ is weakly n-consistent, cf. Theorem 3.5. Hence, $\hat{\beta}_0$ is a weakly consistent estimator of $\mathbb{E}[\tilde{\beta}_{in}]$ as $n\to\infty$ and $N\to\infty$ and it is also n-asymptotically unbiased. Nonetheless, we propose a weighted sample mean of the type

$$\hat{\beta}_0=\left(\sum_{i=1}^{N}C_i\right)^{-1}\sum_{i=1}^{N}C_i\tilde{\beta}_{in} \tag{5.16}$$

with $C_i\in\mathbb{R}^{p,p}$. The sample mean can be written in form of structure (5.16) by selecting constant weights $C_i\equiv C$. The estimator obtained by choosing the C_i as the inverse covariance matrices of the $\tilde{\beta}_{in}$ has minimum MSE among all estimators of type (5.16). Thus, we choose

$$\hat{\beta}_0:=\left(\sum_{i=1}^{N}\mathrm{Cov}(\tilde{\beta}_{in})^{-1}\right)^{-1}\sum_{i=1}^{N}\mathrm{Cov}(\tilde{\beta}_{in})^{-1}\tilde{\beta}_{in}. \tag{5.17}$$

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.

$$\mathbb{E}[\tilde{\beta}_{in}]\overset{n}{\sim}\mathbb{E}[B_i]=:\beta_0,\quad i=1,\dots,n. \tag{5.18}$$

For calculation of (5.17), the inverse covariance matrices in (5.17) have to be replaced by their estimators which follow soon.

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

$$\hat{\tau}=\frac{1}{N-1}\sum_{i=1}^{N}\left(\tilde{\beta}_{in}-\frac{1}{N}\sum_{l=1}^{N}\tilde{\beta}_{ln}\right)\left(\tilde{\beta}_{in}-\frac{1}{N}\sum_{j=l}^{N}\tilde{\beta}_{ln}\right),$$

which appeared in the estimators (3.24) and (3.26). Using the independence of the $\tilde{\beta}_{in}$, we get

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

Further, by (5.18) and the convergence of the second moments,

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

This simplification will not be possible if we use the weighted sample mean (5.17) 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 arithmetic mean. Hence, for some specific cluster i, $\hat{\tau}$ over- or underestimates $\mathrm{Cov}(\tilde{\beta}_{in})$ depending on the constellation of the portfolio. In order to remove the systematic error, we use decomposition (5.15). We then obtain

$$\mathbb{E}[\hat{\tau}] \overset{n}{\sim} \mathrm{Cov}(B_i)+\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathrm{Cov}(\tilde{\beta}_{in}\mid B_i)] = T+\frac{1}{N}\sum_{i=1}^{N}S_{in}. \tag{5.19}$$

As in the iid case, we estimate $\mathrm{Cov}(B_i,\tilde{\beta}_{in})$ by means of its asymptotic equivalent T. By (5.19), we set

$$\hat{T} := \hat{\tau}-\frac{1}{N}\sum_{i=1}^{N}\hat{S}_{in} \tag{5.20}$$

and by (5.15)

$$\widehat{\tau_i^{-1}} := \widehat{\mathrm{Cov}(\tilde{\beta}_{in})}^{-1} := \left(\hat{T}+\hat{S}_{in}\right)^{-1}. \tag{5.21}$$

It remains to find estimators for the S_{in}.

ad S_i. For the estimation of

$$S_{in} = \mathbb{E}[\mathrm{Cov}(\tilde{\beta}_{in} \mid B_i)],$$

we use that the inverse Fisher information matrix is the asymptotic covariance matrix of $\tilde{\beta}_{in}$ given B_i, cf. Theorem 3.5. Recall that we had

$$\hat{S} = \frac{1}{N}\sum_{l=1}^{N} F_n^{-1}(\tilde{\beta}_{ln})$$

in the case of identically distributed clusters. Under the present setting, F_n also depends on the cluster i through the weight and offset terms so we propose

$$\hat{S}_i := \frac{1}{N}\sum_{l=1}^{N} F_{in}^{-1}(\tilde{\beta}_{ln}). \tag{5.22}$$

The idea behind this choice is that $F_{in}^{-1}(B_i)$ is, conditional on B_i, also the asymptotic covariance matrix of $\tilde{\beta}_{in}$. 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 (5.22) replaces the B_l by $\tilde{\beta}_{ln}$. The following lemma, which is also proved in Christiansen and Schinzinger (2015), justifies this procedure.

Lemma 5.8. *For all i and $l = 1, \ldots, N$, $F_{in}(B_l)$ and $F_{in}(\tilde{\beta}_{ln})$ as well as $F_{in}^{-1}(B_l)$ and $F_{in}^{-1}(\tilde{\beta}_{ln})$ are asymptotically equivalent in probability.*

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

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

By Lemma 5.4, F_{in} is Lipschitz continuous on $\mathcal{B}$ with a Lipschitz constant $L > 0$ that neither depends on n nor i. 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 (3.9), 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

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

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

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

converges in probability to zero so that asymptotic equivalence follows. The proof for the inverse sequences works similarly by using their uniform convergence on $\mathcal{B}$, see Lemma 4.3. □

We summarize the results in the following theorem, cf. Christiansen and Schinzinger (2015).

Theorem 5.9. *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

$$\left(\sum_{i=1}^{N}\mathrm{Cov}(\tilde{\beta}_{in})^{-1}\right)^{-1}\sum_{i=1}^{N}\mathrm{Cov}(\tilde{\beta}_{in})^{-1}\tilde{\beta}_{in}.$$

ii) The random matrix

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

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,

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

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

iv) A weakly consistent estimator of $\mathrm{Cov}(\tilde{\beta}_{in})^{-1}$ 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 (5.13) 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 $N \to \infty$.

PROOF. See also the proof of Theorem 3.12. □

Notice that the estimators proposed by Theorem 3.12 are particular cases of the newly introduced ones and in fact, if the clusters are iid, the corresponding estimators will agree. For instance, the weighted sample mean $\hat{\beta}_0$ will reduce to an ordinary sample mean if the weights are identical. One exception are the estimators of the inverse covariance matrices. Compared to (3.24), estimator (5.21) does not include the bias correction term $\frac{N-p-2}{N-1}$, which was motivated by the inverse Wishart distribution. By the nature of structure (5.21), the desired asymptotic distribution is no longer provided. Finally, the credibility formula can be evaluated by replacing its structural parameters by their estimators, cf. Christiansen and Schinzinger (2015).

Corollary 5.10. *The estimator*

$$\hat{\tilde{B}}_i = \hat{A}_i \tilde{\beta}_{in} + (I_p - \hat{A}_i)\hat{\beta}_0. \tag{5.23}$$

is a weakly consistent estimator for the exact credibility estimator (5.12) *as both* $n \to \infty$ *and* $N \to \infty$.

PROOF. The claim is a direct consequence of Theorem 5.9 and (5.14). □

For usage of the credibility formula in application, we again recommend slight modifications. The inverse covariance matrices in (5.17) must be replaced by their estimators, meaning that

$$\hat{\beta}_0 = \left(\sum_{i=1}^{N} \widehat{\tau_i^{-1}} \right)^{-1} \sum_{i=1}^{N} \widehat{\tau_i^{-1}} \tilde{\beta}_{in}, \tag{5.24}$$

This structure can be differently interpreted by means of the credibility matrices A_i. Since $\hat{A}_i = \hat{T}\widehat{\tau^{-1}}$,

$$\hat{\beta}_0 = \left(\sum_{i=1}^{N} \widehat{\tau^{-1}} \right)^{-1} \hat{T}^{-1} \sum_{i=1}^{N} \hat{T}\widehat{\tau^{-1}} \tilde{\beta}_{in} = \left(\sum_{i=1}^{N} \hat{A}_i \right)^{-1} \sum_{i=1}^{N} \hat{A}_i \tilde{\beta}_{in}$$

follows. The right hand side is the credibility weighted sample mean. Furthermore, since $\hat{T}$ estimates $\mathrm{Cov}(B_i)$, it should be made positive semidefinite. The required steps have been already introduced in Section 3.3.

5.4. Simulation study

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

$$\mathrm{rMSE}_i = \frac{\mathbb{E}[\|\hat{\tilde{B}}_i - B_i\|^2]}{\mathbb{E}[\|\tilde{\beta}_i - B_i\|^2]} \tag{5.25}$$

is clearly smaller than 1. By definition of the credibility estimator, $\mathrm{rMSE}_i \leq 1$ always holds. Also recall that we used the simulated rMSE (3.32) for quantifying the credibility estimator in Section 3.4.

In the multidimensional credibility model, the rMSE_i can be expressed as

$$\text{rMSE}_i = \frac{\text{tr}(A_i S_i)}{\text{tr}(S_i)}, \tag{5.26}$$

see Bühlmann and Gisler (2005), Theorem 7.5. Indeed, (5.26) holds true in the present case of CGLMs provided that $\mathbb{E}[\tilde{\beta}_{in}] = \mathbb{E}[B_i]$. However, only the n-limit of the expectations agree. We conjecture that the asymptotic relations

$$\mathbb{E}[\|\hat{B}_{in} - B_i\|^2] \overset{n}{\sim} \text{tr}(A_{in} S_{in}),$$
$$\mathbb{E}[\|\tilde{\beta}_{in} - B_i\|^2] \overset{n}{\sim} \text{tr}(S_{in})$$

hold but a formal proof is missing. Nevertheless, we include the right hand side of (5.26) in our empirical illustrations to study its behavior.

We evaluate the performance of the credibility estimators by means of the relative gain in efficiencies

$$\frac{\sum_{k=1}^m \|\hat{\tilde{B}}_i(\omega_k) - B_i(\omega_k)\|^2}{\sum_{k=1}^m \|\tilde{\beta}_i(\omega_k) - B_i(\omega_k)\|^2}, \quad i = 1, \dots N. \tag{5.27}$$

obtained as the ratio of the simulated mean squared errors (simulated rMSE). The computation is based on $m = 10000$ scenarios denoted by $\omega_1, \dots, \omega_m$. Compared to (3.32), quantity (5.27) is evaluated for all clusters individually. As we have already motivated at the beginning of the chapter, typical forms of cluster specific effects are nuisance parameters in a Binomial-CGLM and exposure terms in a Poisson-CGLM. These cases will be examined more closely.

Binomial case. We consider a portfolio of $N = 30$ independent clusters each with sample size $n = 25$. The distribution of the risk profiles and the structure of the covariate vectors agree with those of Section 3.4. Specifically, the B_i are independently drawn from a Normal distribution with mean vector $\begin{pmatrix} 2 & 1 \end{pmatrix}'$ and covariance matrix I_2. The covariate vectors are given by

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

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 (5.3) are equal to the number of trials which we choose as

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

for all i and j.

Figure 5.1 shows the values of (5.27) for several constellations and estimators. The top left plot belongs to the current case of $N = 30$ and includes the simulated rMSEs for three different estimators. The solid line (1) represents our proposed credibility estimator, i.e. $\hat{T}$ is made positive semidefinite, $\widehat{\tau^{-1}}$ does not include the factor $\frac{N-p-2}{N-1}$ and $\hat{\beta}_0$ is the weighted sample mean (5.24). The dotted line (2) belongs to the credibility estimator which involves the factor $\frac{N-p-2}{N-1}$ in $\widehat{\tau^{-1}}$. Line (3) uses the unweighted sample mean of the PMLEs (3.23) for the estimation of the $\mathbb{E}[\tilde{\beta}_i]$. In addition, the dashed line

represents

$$\frac{\sum_{k=1}^{m} \operatorname{tr}\left(\hat{A}_i(\omega_k)\hat{S}_i(\omega_k)\right)}{\sum_{k=1}^{m} \operatorname{tr}\left(\hat{S}_i(\omega_k)\right)} \tag{5.28}$$

that estimates the right hand side of (5.26). 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 estimator (2) is slightly better for clusters with a small number of trials w_i but get worse as i and thus w_i increase. The line (3) 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 $\hat{\hat{B}}_i$ 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 $\hat{T}$. The simulated mean squared errors without this modification are extremely large, see bottom left plot in Figure 5.1. In the worst case it reaches the value 15.40.

The top right and bottom right plots in Figure 5.1 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%. Furthermore, the distance to (5.28) has decreased and $\frac{\operatorname{tr}(A_i S_i)}{\operatorname{tr}(S_i)}$ seems to be a good approximation for the theoretical rMSE. When we take a look at the modified estimators in the bottom right plot, we observe the same behavior like in the case $N = 30$. Estimators (2) and (3) are almost identical to (1) but a non positive semidefinite $\hat{T}$ should be avoided, see bottom right plot.

Poisson case. Next, we consider a Poisson-CGLM with $N = 30$ and $n = 25$. The assumptions on the B_i and the X_{ij} remain the same as for the Binomial case. We study the effect of different exposure terms E_{ij}, which we choose as

$$E_{ij} = i,$$

and the resulting offsets $\log E_{ij}$. The linear predictor for the log-Poisson parameters are thus given by $\log E_{ij} + X_j B_i$, for all i and j. The calculated values of (5.27) are presented in Figure 5.2. The top left figure shows the simulated (lines (2) and (3)) and theoretical values (dashed line) of the ratio of the mean squared errors. As one would intuitively expect, the smaller the exposure terms the more powerful are the credibility estimations. For cluster $i = 1$, which has offsets $\log 1 = 0$, the rMSE is about 88% and matches with that of the setting $(n, N) = (25, 30)$ in Table 3.1. With increasing exposure, the advantage rapidly decreases from 12% down to about 1%. Compared to the Binomial case, the gap between the simulated rMSE and the approximated theoretical rMSE is small. A possible explanation might be the improved estimation of the conditional covariance matrix $\operatorname{Cov}(\tilde{\beta}_{in} \mid B_i)$, which has been discussed in Section 4.1. Further estimators have been also considered. The dotted line, which is numbered as (3), belongs to the credibility estimator that uses the non-weighted sample mean instead of $\hat{\beta}_0$. Again, the non-weighted sample mean is very close to the weighted sample mean and therefore leads to almost identical values of the rMSE. That is why line (3) is practically invisible. Nevertheless, the weighted sample mean has a lower empirical variance. The bottom

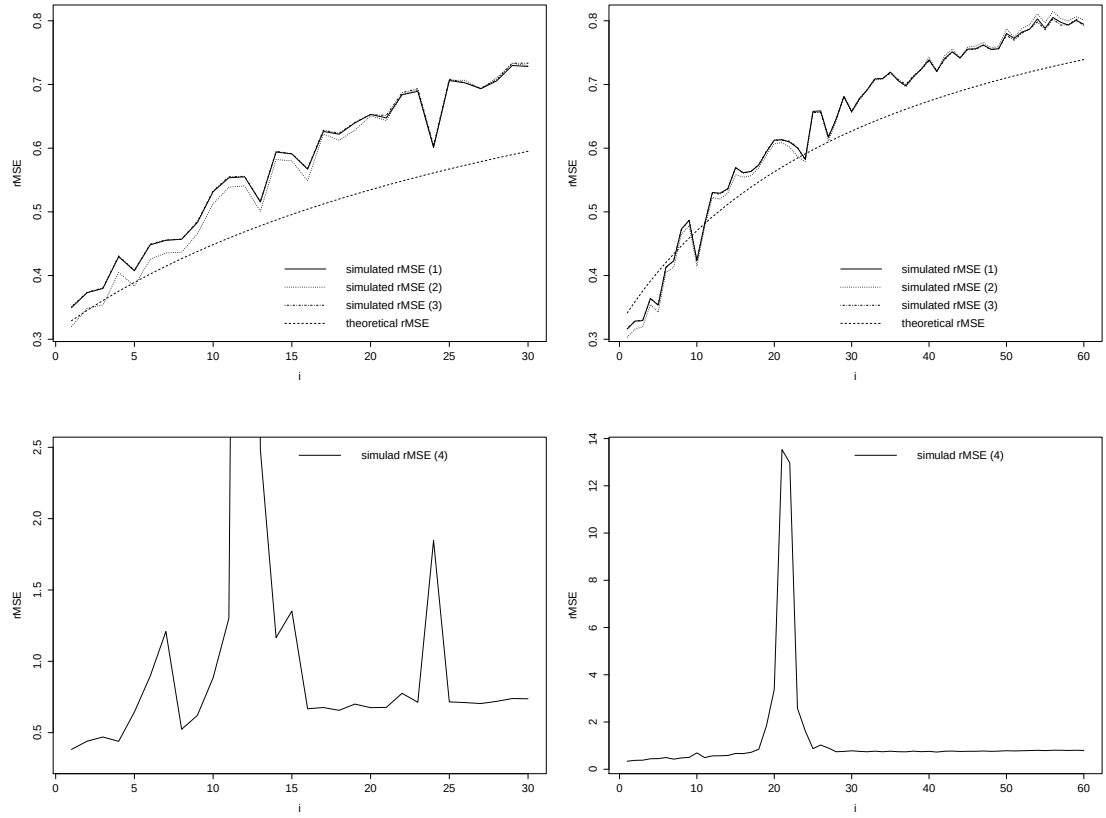

FIGURE 5.1. Plots of the rMSE for a Binomial-CGLM with $N = 30$ (top and bottom left) and $N = 60$ (top and bottom right). The lines correspond to the following estimators.

simulated rMSE (1): The proposed credibility estimator with positive semidefinite $\hat{T}$, $\widehat{\tau^{-1}}$ without factor $\frac{N-p-2}{N-1}$ and weighted sample mean $\hat{\beta}_0$.

simulated rMSE (2): The credibility estimator using $\frac{N-p-2}{N-1}\widehat{\tau^{-1}}$ instead.

simulated rMSE (3): The credibility estimator using the unweighted sample mean instead.

simulated rMSE (4): The credibility estimator using a non positive semidefinite $\hat{T}$ instead.

theoretical rMSE: (5.28)

left plot shows (5.27) for estimator (2) which involves the additional factor $\frac{N-p-2}{N-1}$. Unlike the Binomial case, the results now clearly differ. High exposure clusters are evaluated as highly credible and their estimated credibility matrices $\hat{A}_i$ are close to the identity matrix. The credibility matrix used by estimator (2) is $\frac{N-p-2}{N-1}\hat{A}_i$ and falsely classifies cluster i to be less credible. That explains the monotone increasing structure of (2). Finally, concerning $\hat{T}$, $\hat{T}$ was always positive semidefinite in each of the 10000 simulations. The analog plots on the right hand side of Figure 5.2 belong to the portfolio extended to $N = 60$. Credibility estimators perform slightly better but the overall behavior is the same as for the case $N = 30$.

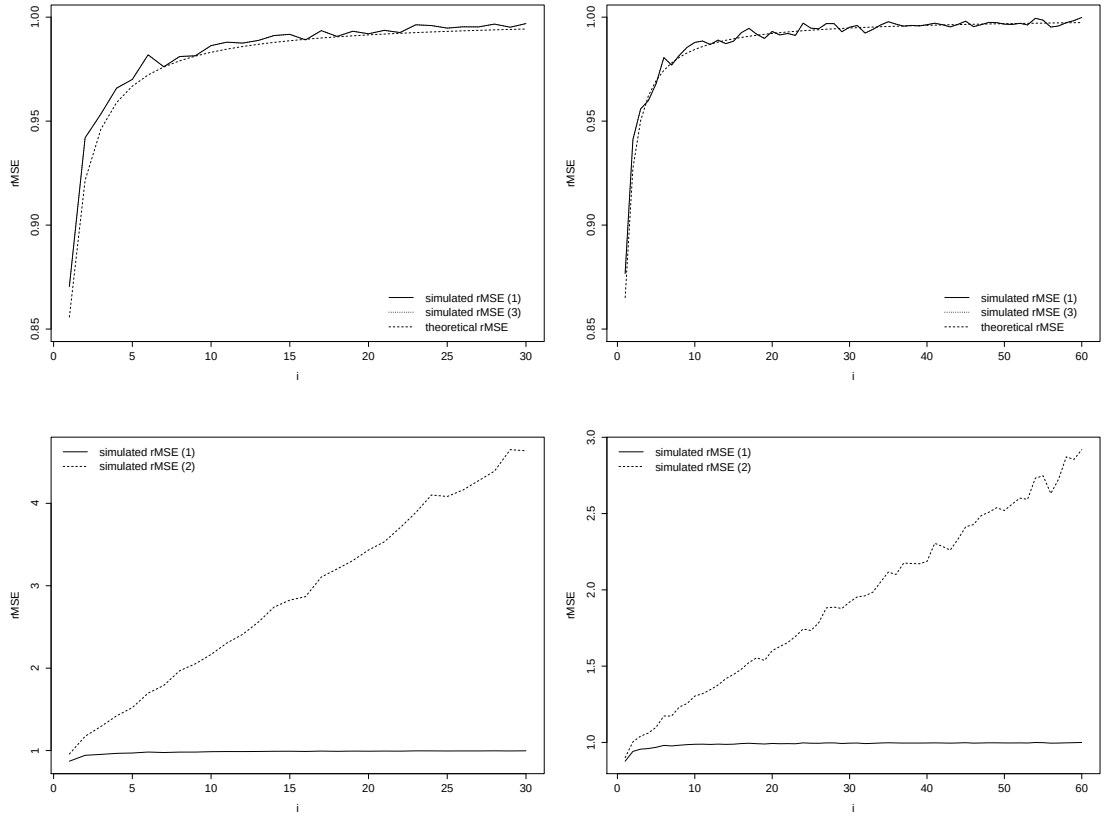

FIGURE 5.2. Plots of the rMSE for a Poisson-CGLM with $N = 30$ (top and bottom left) and $N = 60$ (top and bottom right). The lines correspond to the following estimators.

simulated rMSE (1): The proposed credibility estimator with positive semidefinite $\hat{T}$, $\widehat{\tau^{-1}}$ without factor $\frac{N-p-2}{N-1}$ and weighted sample mean $\hat{\beta}_0$.

simulated rMSE (2): The credibility estimator using $\frac{N-p-2}{N-1}\widehat{\tau^{-1}}$ instead.

simulated rMSE (3): The credibility estimator using the unweighted sample mean instead.

theoretical rMSE: (5.28)

The case with clusters of different sample sizes n_i, $i = 1, \ldots, N$, leads to similar results: Clusters with small sample sizes show greater improvements when the credibility estimator is used but the advantages vanish as the n_i grow. We therefore abstain from presenting further plots.

CHAPTER 6

Application to mortality data

This chapter deals with actuarial applications of a CGLM and in particular, we will study mortality data, which is for instance used for pricing life insurance products. We will discuss the advantages and possible difficulties of credibility models on the basis of the constructed model.

6.1. The model

We first give a short introduction to mortality models for single countries. An extension to models for multiple countries using the CGLM framework will follow. We observe mortality statistics for people of ages $x = x_1, \dots, x_m$ with m being the number of age groups. The data are taken from calendar years numbered consecutively from $t = 1$ to T. For each age x and year t, the Human-Mortality-Database (2014) provides death counts $D_x(t)$ as well as the initial exposure-to-risk $E_x(t)$. As we have already motivated, Poisson models and Poisson-GLMs are the 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

$$\log \mathbb{E}[D_x(t)] = \log E_x(t) + \kappa_1(t) + \kappa_2(t)(x - \bar{x}) \tag{6.1}$$

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 (6.1) 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

$$X^{(1)} = I_T \otimes 1_m = I_T \otimes \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and} \quad X^{(2)} = I_T \otimes \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_m - \bar{x} \end{pmatrix}. \tag{6.2}$$

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 (6.1). The GLM can now be solved for the unknown period parameters $\kappa_1(t)$ and $\kappa_2(t)$. Further mortality models are for instance the polynomial model introduced by Hatzopoulos and Haberman (2009) and the age-period-cohort model (APC model) by Osmond (1985). The polynomial model extends the CBD model (6.1) by incorporating

orthonormal polynomials of higher orders in the covariate vectors. A different approach is followed by the APC model which is given by

$$\log \mathbb{E}[D_x(t)] = \log E_x(t) + \beta_x + \kappa_t + \gamma_{t-x}.$$

It includes, as the name suggests, age effects β_x, period effects κ_t and cohort effects γ_{t-x} and all effects do not interact. Both models can be expressed as a GLM by choosing proper design matrices. See Currie (2013) and Appelt (2014) for details and further models.

More and more research papers study models for multiple countries instead of the just presented single country models. A particular example is that of Hatzopoulos and Haberman (2013) which is an extension of the polynomial model and also of the CBD model as its special case. Consider a portfolio of countries labeled $i = 1, \dots, N$ and denote by $D_{ix}(t)$ and $E_{ix}(t)$ the data for country i. An aggregate mortality statistics is created by summing over all i, i.e.

$$D_{\bullet x}(t) := \sum_{i=1}^{N} D_{ix}(t) \quad \text{and} \quad E_{\bullet x}(t) := \sum_{i=1}^{N} E_{ix}(t),$$

and they are assumed to satisfy (6.1) such that

$$\log \mathbb{E}[D_{\bullet x}(t)] = \log E_{\bullet x}(t) + \kappa_{\bullet 1}(t) + \kappa_{\bullet 2}(t)(x - \bar{x}).$$

Given the estimators $\hat{\kappa}_{\bullet 1}(t)$ and $\hat{\kappa}_{\bullet 2}(t)$, the N individual models

$$\log \mathbb{E}[D_{ix}(t)] = \log E_{ix}(t) + (\kappa_{\bullet 1}(t) + \kappa_{i1}(t)) + (\kappa_{\bullet 2}(t) + \kappa_{i2}(t))\,(x - \bar{x}). \tag{6.3}$$

are considered. In this stage of fit, estimation proposed by Hatzopoulos and Haberman (2013) accounts for the number of parameters in the model such that some $\hat{\kappa}_{ik}(t)$ are set to zero. The final model (6.3) allows coherent mortality forecasts with a common mortality trend for all countries and is therefore of great interest for insurance companies with a worldwide portfolio. Coherent forecasts can be achieved by extrapolating $\hat{\kappa}_{i1}(t)$ and $\hat{\kappa}_{i2}(t)$ with zero-drift processes. However, the advantage for companies acting only on a national level is not clear. This is exactly where CGLMs come into play since they target to improve estimation for single countries, especially for small ones.

We adjust the CBD model (6.1) to fit into the credibility framework described by (W1) to (W3) in Section 5.2. 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

$$\log \mathbb{E}[D_{ix}(t) \mid B_i] = \log E_{ix}(t) + K_{i1}(t) + K_{i2}(t)(x - \bar{x}), \tag{6.4}$$

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 $K_{i2} = (K_{i2}(t))_{t=1}^{T}$. We can reformulate structure (6.4) like (5.4) by using the design matrix $X = (X^{(1)}, X^{(2)})$ as defined in (6.2).

6.2. The fit

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. This is the reason why Germany is separately considered in its eastern and western part. We fit the CBD model for the ages 30 to 90 and the calender years 1980 to 2000, thus $m = 31$ and $T = 21$. The data for the years 2001 to 2009 are used later on for model validation. First, individual maximum likelihood estimators are obtained by fitting the conditional model, which is (6.1), 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 (5.23). We are especially interested in the countries where the credibility estimates clearly differ from the conditional maximum likelihood estimates. The top left plot in Figure 6.1 shows the quadratic deviations $\|\hat{\tilde{B}}_i - \hat{\beta}_i\|^2$. We take a closer look at Finland as its estimators record the largest deviation. Its estimators for $K_{i1}(t)$, $t = 1, \dots, T$, as well as the corresponding components of the collective estimator $\hat{\beta}_0$ are presented in the top right plot. Contrary to the expectation, the credibility estimator produces a wild zigzag line.

The reason behind is the sample covariance matrix $\hat{\tau}$, which appears in estimator (5.21) in form of

$$\widehat{\tau_i^{-1}} = \widehat{\mathrm{Cov}(\tilde{\beta}_i)}^{-1} = \left(\hat{T} + \hat{S}_i\right)^{-1} = \left(\hat{\tau} - \frac{1}{N}\sum_{l=1}^{N} \hat{S}_l + \hat{S}_i\right)^{-1}.$$

It may happen like in the case of iid clusters that

$$\hat{S}_i = \frac{1}{N}\sum_{l=1}^{N} \hat{S}_l \tag{6.5}$$

such that $\widehat{\tau_i^{-1}} = \hat{\tau}^{-1}$. But since $N = 36 < 42 = p$, $\hat{\tau}$ is not of full rank and cannot be inverted. In fact for Finland, the left and right hand side matrices of (6.5) are very close to each other, i.e. cluster Finland contains an average amount of information. The resulting matrix $\hat{T} + \hat{S}_i$ has an eigenvalue 10^{-4}, which is the closest eigenvalue to zero even under all countries in fit. Although the inverse matrix $\widehat{\tau_i^{-1}}$ exists, its condition number $\|\widehat{\tau_i^{-1}}\|\|\hat{T}+\hat{S}_i\|$ is very large. Such matrices are called ill-conditioned and typically have entries of different scales within the same row or column. The fact that $\widehat{\tau_i^{-1}}$ is ill-conditioned is also strongly connected to the model structure (6.4). Since no assumption about the dependence structure of K_{i1} and K_{i2} is made, $\mathrm{Cov}(\tilde{\beta}_i)$ has a block structure with non-zero off diagonal blocks. Structure (5.23) of the credibility formula then allows the credibility estimator for K_{i1} to depend on the individual estimator for K_{i2}, too. However, empirical evidence shows that the processes K_{i1} and K_{i2} are independent.

We refit the credibility CBD model under this additional independence assumption, i.e. we assume that the first T and last T components of

$$B_i = \begin{pmatrix} K_{i1}(1) & \dots & K_{i1}(T) & K_{i2}(1) & \dots & K_{i2}(T) \end{pmatrix}'$$

are independent. Now, credibility estimators are calculated separately for K_{i1} and K_{i2} by applying the credibility formula only to the corresponding components of $\hat{\beta}_i$. Separating

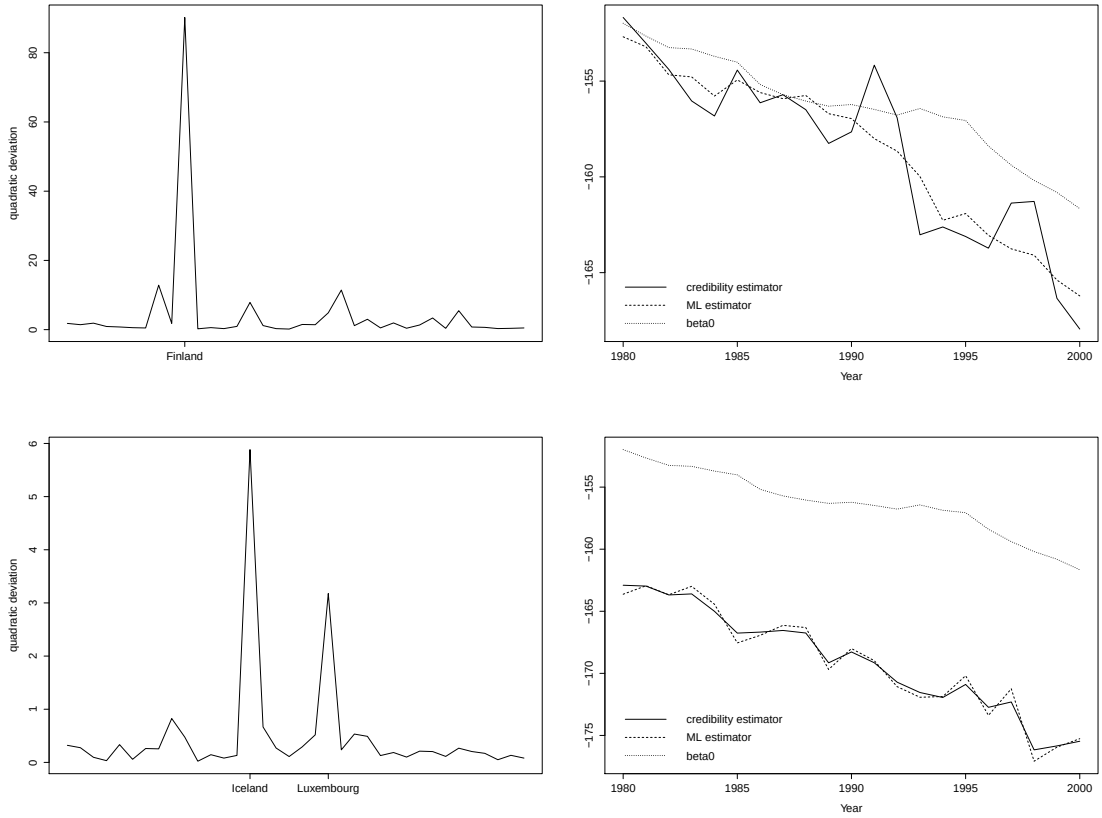

FIGURE 6.1. Several plots for the credibility CBD model: The plots on the left show the quadratic deviations between the credibility and ML estimators without (top left) and with (bottom left) the independence assumption of K_{i1} and K_{i2}. The top right plot presents the estimators for K_{i1} of Finland without the independence assumption and the bottom right one belongs to those of Iceland under the independence assumption.

the procedure also avoids the issue $N < p$ since we now have $p = T = 21$. The final credibility matrices $\hat{A}_i$ for B_i have a block diagonal structure

$$\hat{A}_i = \begin{pmatrix} \hat{A}_{i1} & 0 \\ 0 & \hat{A}_{i2} \end{pmatrix},$$

where $\hat{A}_{i1}$ and $\hat{A}_{i2}$ are the credibility matrices of K_{i1} and K_{i2} respectively. The bottom left plot in Figure 6.1 shows the quadratic deviations between $\hat{\hat{B}}_i$ and $\hat{\beta}_i$. No longer Finland but Iceland and Luxembourg clearly stand out. Since these two countries have the smallest population of all N countries in the fit the result is natural and coincides with the propose of credibility estimation. The bottom right plot of Figure 6.1 displays the estimators for K_{i1} of Iceland. The estimators for K_{i2} are not shown since they almost agree for all countries in the fit.

6.3. Benefits of credibility estimation

Let $\hat{K}_{i1}^{(\text{cred})}(t)$ and $\hat{K}_{i1}^{(\text{ML})}(t)$, $t = 1, \ldots, T$, be the estimators for the age-independent period effects. We already know that the credibility estimator is more favorable than the MLE in context of the mean squared error. This property reflects in form of smoother paths of $(\hat{K}_{i1}^{(\text{cred})}(t))_{t=1}^{T}$ whereas the MLEs better fit the particular realization but tend to overfit the data. We use the sample standard deviation of the innovations $(\hat{K}_{i1}(t+1) - \hat{K}_{i1}(t))_{t=1}^{T-1}$ to measure the smoothness of both estimated paths. The calculated values are presented in Figure 6.2 and we have an relative improvement by about 9% on average. The largest improvements can be observed for Iceland ($i = 15$) and Luxembourg ($i = 21$) with values 41% and 47% respectively.

Period effects with smooth paths are indeed a high desired property in estimating and forecasting mortality and several papers explicitly target for such models. Currie et al. (2004), for example, uses a Poisson-GLM with a penalty term incorporated in the likelihood function. The penalty has the structure

$$\lambda \sum_{t=2}^{T-1} (\beta_{t-1} - 2\beta_t + \beta_{t+1})^2, \tag{6.6}$$

where β_t is the regression coefficient for year t, and accounts for smoothness over adjacent calender years. A smoothing parameters $\lambda > 0$ has to be specified by the user. Russolillo et al. (2011), on the other hand, considers a multi-country model for the three-way array of log death rates $\log m_{ix}(t) = \log D_{ix}(t) - \log E_{ix}(t)$ for country i, age$x$ and year t. The array is decomposed into its first principal components β_x, γ_i and $\kappa(t)$ such that

$$\log m_{ix}(t) = \alpha_{ix} + \beta_x \gamma_i \kappa(t) + \epsilon_{ix}(t) \tag{6.7}$$

with centering term α_{ix} and modeling error $\epsilon_{ix}(t)$. The country specific period effects $(\gamma_i \kappa(t))_{t=1}^{T}$ can be interpreted as being perfectly smoothed over all countries since irregularities may only appear in either all or none of the countries. A smoothing over calender years does not take place. Our credibility model incorporates smoothing in both dimensions, calender year and country. Individual paths of period effects are mixed with the path of the portfolio and more precisely for $t = 1, \ldots, T$,

$$\hat{K}_{i1}^{(\text{cred})}(t) = \hat{K}_{i1}^{(\text{col})}(t) + \sum_{s=1}^{T} \hat{A}_{i1}(t,s) \left(\hat{K}_{i1}^{(\text{MLE})}(s) - \hat{K}_{i1}^{(\text{col})}(s) \right)$$

with $\hat{K}_{i1}^{(\text{col})}(t)$ being the corresponding component of the collective estimator $\hat{\beta}_0$. Of course, the structure itself does not ensure smoothness but the property naturally follows from the credibility weights $\hat{A}_{i1}$, which are optimally chosen. Unlike model (6.7), peaks in country specific period effects may remain if these are evaluated to be credible. Figure 6.3 shows the first four rows of $\hat{A}_{\text{Iceland},1}$ in order to provide clarity about the smoothing procedure. The maximum weight is always allocated to the current years, i.e. $t = 1, \ldots, 4$ respectively, and weights decrease to zero as the time lag increases in both directions (positive and negative). Moreover, compared to (6.6), there is no smoothness parameter involved and the "degree of smoothness" naturally results from the constellation of the data.

Example 6.1 (Credibility CBD model for high ages). Figure 6.4 belongs to the credibility CBD model fitted to the ages 90 to 100, years 1980 to 2000 and the same portfolio of countries. The two plots show credibility and ML estimators for K_{i1} of

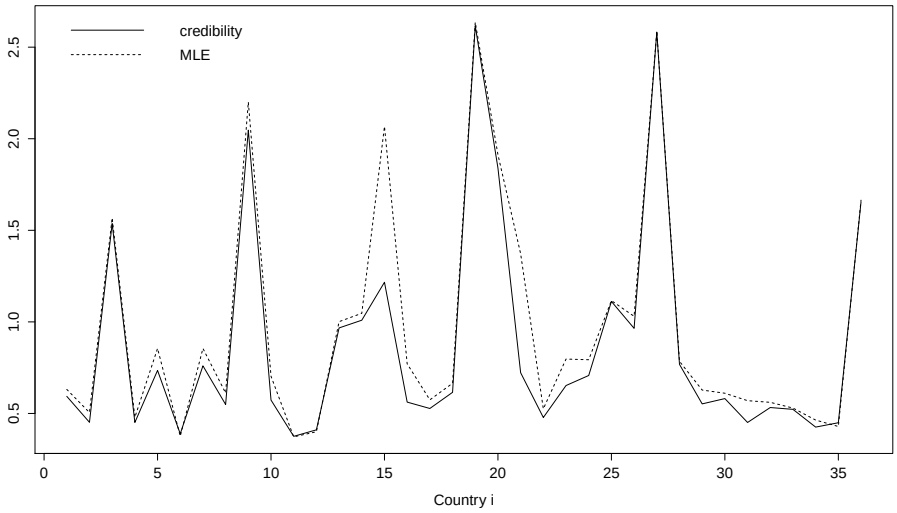

FIGURE 6.2. The standard deviations of the innovation terms as an indicator for smoothness.

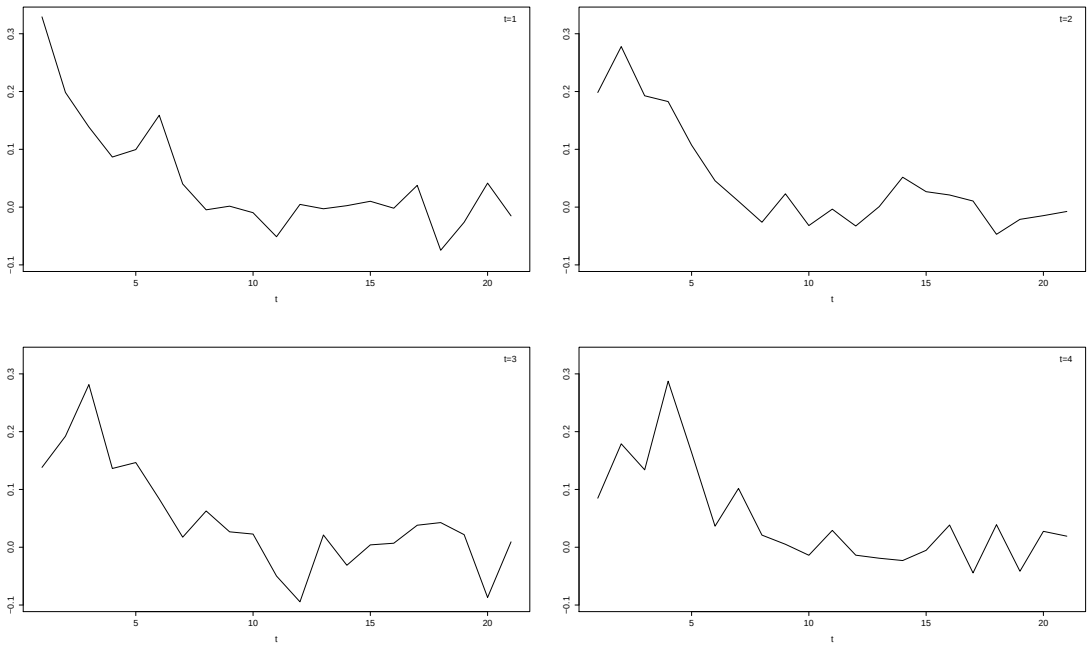

FIGURE 6.3. The credibility weights of the first four years for Iceland.

Iceland (left) and Luxembourg (right). Since we have less age groups in fit and also with smaller exposure terms, the credibility estimators are expected to profit more from the other clusters. In fact, the mentioned smoothing effect is visible more clearly. 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.

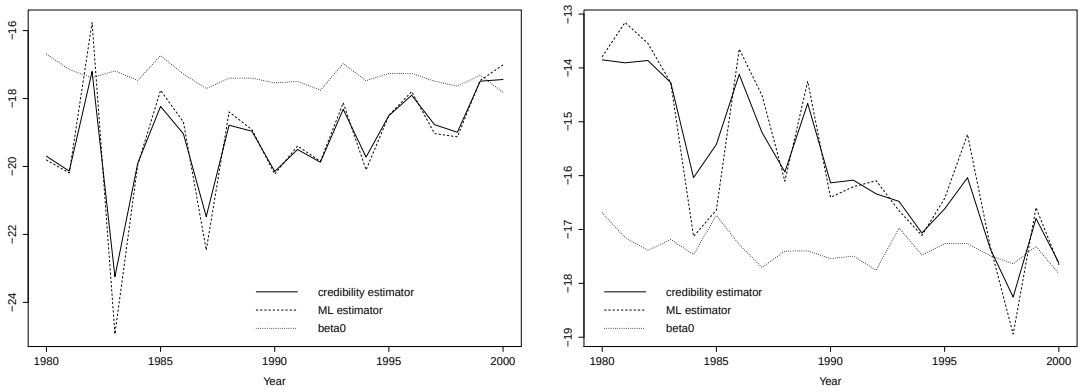

FIGURE 6.4. The credibility CBD model for high ages, estimators for K_{i1} of Iceland (left) and Luxembourg (right).

The same incidence happens in Luxembourg, 1984. In that year, every 22 of the 96 years 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 successfully smooth these peaks out.

6.4. Consequences for mortality forecasting

Finally, we study the estimation results for the ages 30 to 90 by means of the remaining life expectancies at age 30. We assume that these are outcomes of ARIMA processes with Gaussian white noise and use the Box-Jenkins methodology to find the best fits. The ARIMA orders (p, d, q) with $p, q \in \{0, 1, 2\}$ and $d \in \{0, 1\}$ are considered and evaluated via the AICc. Details on the methodology follow in Section 7.3 and the focus at this point is on the resulting forecasts. Since the path obtained from the credibility estimators is more smooth, we expect different estimators for the ARIMA parameters and even different ARIMA orders are possible. This is indeed the case as Table 6.1 shows. The future mortality tables are then constructed from the point forecasts of the ARIMA processes and period life expectancies directly follow. See Chapter 8 for the explicit formulas. The forecast life expectancies together with those calculated from the observed mortality tables are presented in Figure 6.5 (left). One can immediately recognize that the observed life expectancies of the validation years are much better approximated by the forecasts produced by the credibility estimators. To be more precise, the quadratic deviations to the observed values equal 0.451 whereas the MLE-forecasts deviate by 1.350. Furthermore, we study the impact of different variances of the Gaussian error terms, see last columns in Table 6.1. We calculate 95% confidence intervals of period life expectancies for the years 2001 to 2009 and check whether they include the observed values. The results are shown in the right plot of Figure 6.5. Since the estimated variance for K_{i1} is much smaller in the credibility case, we obtain narrower confidence intervals (solid lines compared to dashed lines). Nevertheless, the observed life expectancies (dotted line) are fully contained.

data	order (p,d,q)	AR1	MA1	MA2	intercept	σ^2
$\hat{K}_{i1}^{(\text{cred})}$	$(0,1,1)$	-	-1	-	-0.666	0.664
$\hat{K}_{i1}^{(\text{ML})}$	$(0,1,2)$	-	-1.975	1	-0.663	1.089
$\hat{K}_{i2}^{(\text{cred})}$	$(1,0,0)$	0.454	-	-	62.416	6.235
$\hat{K}_{i2}^{(\text{ML})}$	$(1,0,0)$	0.451	-	-	62.416	6.279

TABLE 6.1. Parameter estimates for K_{i1} and K_{i2}.

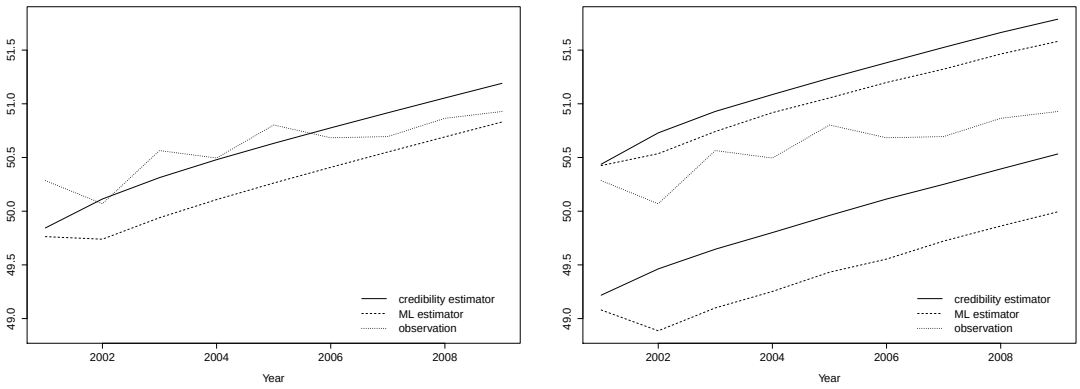

FIGURE 6.5. Forecasts of period life expectancies, point predictions (left) and confidence intervals (right).

6.5. Selection of the cluster structure

An important basis for credibility estimation is the structure of the underlying data set. The learning effect, which appeared as a smoothing effect in context of mortality estimation, is the essential component in estimation. While it is also a relevant information that a cluster is fully credible, the resulting estimator equals the individual MLE and does not justify the increased model complexity. The credibility formula (5.14) involves additional parameters and one needs to find the right balance between model complexity and model fit. In particular, the learning effect must be clearly evident. To clarify the importance of the cluster structure, recall the mortality setting where clusters were the countries in fit. We could increase the number of countries N by considering the countries of United Kingdom (UK) separately. The Human Mortality Database has indeed individual data for England & Wales, Scotland and Northern Ireland. Even finer regional partitions, not necessarily of UK, are conceivable as long as data is available. Increasing the portfolio size would improve the estimation of the structural parameters but additional parameters have to be estimated, too. Moreover, if the partitions are very similar, the separation will not provide any new insights. Choosing the proper cluster structure thus needs special attention.

Let (B_i, Y_i), $i = 1, \dots, N$, be a portfolio of N independent clusters. Furthermore, let (B_i, Y_i), $i = 1, \dots, (N-1) + \tilde{N}$ with $\tilde{N} \geq 2$, be the same portfolio where a particular cluster is finer partitioned into $\tilde{N}$ clusters that are without loss of generality numbered $i = 1, \dots, \tilde{N}$. For example, partitioning UK into the $\tilde{N} = 3$ mentioned countries would

increase the mortality portfolio by two clusters. In order to ensure that the new clusters contribute to the learning effect of the credibility estimator, their realizations of B_i has to differ significantly from each other. Otherwise their conditional distributions agree (apart from known weights and offset terms) and there is no need to distinguish between these clusters. Thus, the hypothesis is

$$\Theta_0 = \left\{ \begin{pmatrix} B_1 \\ \vdots \\ B_{\tilde{N}} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \otimes \beta : \beta \in \mathbb{R}^p \right\} \tag{6.8}$$

and the aim of this section is to develop a test procedure for Θ_0. For the particular example of UK in the credibility CBD model, the death counts and exposure terms of the three countries sum to these of UK respectively. If Θ_0 holds, the conditional Poisson rates $\mathbb{E}[D_{ix}(t)|B_i]$, $i = 1, \dots, 3$, will also add up to the UK rates but this identity is in general not satisfied under the alternative hypothesis. The B_i contain country/region specific effects which are not reflected in the aggregate model that is UK. The test for Θ_0 is based on the MLEs $\hat{\beta}_i$ which contain the essential information about the clusters. Intuitively, we will reject Θ_0 if the deviation

$$\sum_{i=1}^{\tilde{N}} \|\hat{\beta}_i - \hat{\beta}\|^2$$

from an estimator $\hat{\beta}$ of β is large. In fact, this expression can be extended to a $\mathcal{X}^2$-test.

We begin with some preliminary results. Let $Z_1, \dots, Z_N$ be iid random variables with $Z_1 \sim \mathcal{N}(0, 1)$. It is a known fact that

$$\sum_{i=1}^{N} Z_i^2 \sim \mathcal{X}^2(N), \tag{6.9}$$

where $\mathcal{X}^2(N)$ denotes a chi-squared distribution with N degrees of freedom. Considering $Z = (Z_1, \dots, Z_N)$ as a multivariate Normal variable, i.e. $Z \sim \mathcal{N}(0, I_N)$, allows us to write

$$\|Z\|^2 = Z'Z = \sum_{i=1}^{N} Z_i^2 \sim \mathcal{X}^2(N). \tag{6.10}$$

Similarly, relation (6.9) can be extended to multivariate Normal random variables as Pearson (1900) shows.

Theorem 6.2 (Pearson, 1900). *Let $Z_1, \dots, Z_N$ be independent random vectors with $Z_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, where $\mu_i \in \mathbb{R}^p$ and $\Sigma_i \in \mathbb{R}^{p,p}$ denote the mean vectors and covariance matrices respectively. Then, if the Σ_i are invertible,*

$$\sum_{i=1}^{N} (Z_i - \mu_i)' \Sigma_i^{-1} (Z_i - \mu_i) \sim \mathcal{X}^2(Np).$$

Alternative proof. Pearson's original proof shows the identity of the distribution functions. We present a more probabilistic approach instead. Fix $1 \leq i \leq N$. We show that

$$(Z_i - \mu_i)' \Sigma_i^{-1} (Z_i - \mu_i) \sim \mathcal{X}^2(p). \tag{6.11}$$

Then, the claim immediately follows by the additivity of the chi-squared distribution. By (6.10), it suffices to find a $W \sim \mathcal{N}(0, I_N)$ such that

$$(Z_i - \mu_i)'\Sigma_i^{-1}(Z_i - \mu_i) \stackrel{d}{=} W'W.$$

Since Σ_i is a covariance matrix, it is orthogonally diagonalizable with transformation

$$P'\Sigma_i P = \Lambda, \tag{6.12}$$

where Λ is a diagonal matrix with positive diagonal entries. Let $Q = P'\Lambda^{-1/2}P$ and choose

$$W = Q(Z_i - \mu_i).$$

As $Q\Sigma_i Q' = I_p$, W follows a multivariate standard Normal distribution and furthermore by (6.12),

$$\begin{aligned} W'W &= (Z_i - \mu_i)'P'\Lambda^{-1/2}PP'\Lambda^{-1/2}P(Z_i - \mu_i) \\ &= (Z_i - \mu_i)'\Sigma_i^{-1}(Z_i - \mu_i) \end{aligned}$$

and the claim follows. □

A common technique in chi-squared tests for univariate samples $(Z_1, \dots, Z_n)$ is to use the sample mean $\bar{Z}$ instead of the unknown mean μ. Then,

$$\sum_{i=1}^{N} \left(Z_i - \bar{Z}\right)^2 \sim \mathcal{X}^2(N-1), \tag{6.13}$$

i.e. the replacement costs one degree of freedom. An analogue result holds in the multivariate case, where we lose one degree of freedom in each of the p components.

Theorem 6.3 (Cochran, 1934). *Let $Z_i \sim \mathcal{N}(\mu, I_p)$, $i = 1, \dots, N$, be iid p-variate Normal vectors. Then,*

$$\sum_{i=1}^{N} \left(Z_i - \bar{Z}\right)' \left(Z_i - \bar{Z}\right) \sim \mathcal{X}^2((N-1)p),$$

where $\bar{Z} = \left(\bar{Z}_1', \dots, \bar{Z}_p'\right)'$ is the sample mean vector.

The two results can now be used to construct a $\mathcal{X}^2$-test for the sub-portfolio of partitions $(B_i, Y_i)_{i=1}^{\tilde{N}}$. Under Θ_0, each cluster follows a GLM with unknown GLM parameter β. Their MLEs $\hat{\beta}_{in}$ are asymptotically Normal distributed, such that for large enough n

$$\hat{\beta}_{in} \sim \mathcal{N}(\beta, F_{in}(\hat{\beta}_{in})).$$

Theorem 6.2 and Theorem 6.3 directly imply the following corollary.

Corollary 6.4 ($\mathcal{X}^2$-test). *Let $\bar{\hat{\beta}}$ be the sample mean of $\hat{\beta}_1, \dots, \hat{\beta}_{\tilde{N}}$. The expression*

$$D = \sum_{i=1}^{\tilde{N}} \left(\hat{\beta}_i - \bar{\hat{\beta}}\right)' F_{in}^{-1}(\hat{\beta}_i) \left(\hat{\beta}_i - \bar{\hat{\beta}}\right)$$

is asymptotically chi-squared distributed with $(\tilde{N}-1)p$ degrees of freedom.

The hypothesis Θ_0 will be rejected with significance $(1-\alpha) \in (0,1)$ if D exceeds the α-quantile of a $\mathcal{X}^2((\tilde{N}-1)p)$-distribution.

Example 6.5 (United Kingdoms). The quantity D for the $\tilde{N} = 3$ countries England & Wales, Scotland and Northern Ireland is given by

$$D = 48.856$$

whereas the 5% quantile of a $\mathcal{X}^2(2p)$, $p = 21$, is given by 106.395. Hence, the hypothesis Θ_0 is not rejected and we keep UK in the portfolio. In fact, using the extended portfolio does not change the credibility estimation for Iceland. The quadratic deviation between the two credibility estimators is 0.009 on the scale $\|\hat{\hat{B}}^{(N=36)}_{\text{Iceland}}\|^2 = 682348$.

6.6. Conclusion

We have now intensively dealt with the application to mortality data but CGLMs are generally suitable for a wide range of actuarial problems. From a pure theoretical point of view, every GLM can be extended to a CGLM as far as data is available. However, if the portfolio structure is not naturally given like in the case of mortality data, the data has to be first partitioned into N clusters. In order to ensure that the sample covariance matrix is of full rank, it must hold that $N \geq p$. Furthermore, it is not guaranteed that the MLEs $\hat{\beta}_i$, $i = 1, \ldots, N$, of the resulting clusters can be uniquely determined. This, for instance, applies for categorical covariates, where the design matrix X_i will contain a column full of zeros if the corresponding category is not observed within cluster i.

Once a CGLM is applicable, it improves conventional estimators not only in theory but also in practice. The benefit of credibility estimation is especially large, if a cluster contains only a little statistical information in terms of population size or death counts close to zero. An improved mean squared error reflected in form of a learning effect which produced smoother paths. Although it was not observable in our analysis, this learning effect could have also worked into the other direction. More precisely, peaks or humps appearing in the path of the collective estimator can be added to individual paths if the calculated credibilities force to do so. In any case, the credibility paths for the CBD model look reasonable and also the forecasts make sense. The model actually delivers useful results for Iceland and Luxembourg. In addition, most of the countries were evaluated as being highly credible, which is as well an important discovery. We conclude that CGLM is a highly promising model. Further investigation besides mortality data was unfortunately limited due to the lack of suitable data.

Part 2

Evolutionary Credibility models of ARMA type

CHAPTER 7

A credibility model for mortality projection

The content of this and the upcoming chapter is also presented in Schinzinger et al. (2014) in a more demographic context. The present thesis picks up the ideas and methodologies of that paper but concentrates more on the mathematical aspects. Empirical illustration on the basis of Belgian mortality data is identical to Schinzinger et al. (2014).

7.1. Evolutionary credibility models

A particular risk $(Y_t, \Delta_t)_{t=1}^T$ with a w-variate observation process (Y_t) and a v-variate risk profile process (Δ_t) on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is considered. This corresponds to a portfolio $(Y_i, B_i)_{i=1}^N$ of the previous credibility models but notation differs to have a clear distinction between the models. The unobservable process (Δ_t) follows a known dynamics and the premiums $\mathbb{E}[Y_t \mid \Delta_t]$ change with the passage of time. The aim is to predict future Δ_{T+k} or $\mathbb{E}[Y_{T+k} \mid \Delta_{T+k}]$, $k \in \mathbb{N}$, by means of past observations. In this connection, one defines for $t \geq s$

$$\mu_{t|s} := \mathrm{Pro}\left(\Delta_t \mid L(1, Y_1, \ldots, Y_s)\right), \tag{7.1}$$

where

$$L(1, Y_1, \ldots, Y_s) = \left\{ a + \sum_{k=1}^{N} A_k Y_k : a \in \mathbb{R}^v, A_k \in \mathbb{R}^{v,w} \right\},$$

and the corresponding error covariance matrix

$$Q_{t|s} := \mathbb{E}\left[(\Delta_t - \mu_{t|s})(\Delta_t - \mu_{t|s})'\right]. \tag{7.2}$$

The Y_t are required to be square integrable so that these quantities are all well-defined. Then, $\mu_{t|s}$ is the best linear predictor of Δ_t in terms of all components of $1, Y_1, \ldots, Y_s$. Unlike in the previous credibility models, the Δ_t are not necessarily independent. More precisely, (Y_t, Δ_t) is assumed to have a state-space representation of the form

$$\Delta_{t+1} = F\Delta_t + V_t, \tag{7.3}$$

$$Y_t = G\Delta_t + W_t \tag{7.4}$$

with $F \in \mathbb{R}^{v,v}$, $G \in \mathbb{R}^{w,v}$ and white noise processes (V_t) and (W_t). The two white noise processes are serially uncorrelated and also uncorrelated with each other. Their joint covariance matrix thus has the structure

$$\mathrm{Cov}\left(\begin{pmatrix} V_t \\ W_t \end{pmatrix}, \begin{pmatrix} V_s \\ W_s \end{pmatrix}\right) = \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \tag{7.5}$$

$Q \in \mathbb{R}^{v,v}$ and $R \in \mathbb{R}^{w,w}$, if $t = s$ and zero else. Under these assumptions, the credibility estimators $\mu_{T+k|T}$ for Δ_{T+k} can be calculated in a recursive way.

Theorem 7.1. *Assume that* (7.3) *to* (7.5) *hold. Starting with the initial values*

$$\mu_{1|0} = \mathbb{E}[\Delta_1],$$
$$Q_{1|0} = \mathrm{Cov}(\Delta_1) = Q,$$

we have the following recursion formula. Provided an additional observation Y_t, (7.1) *and* (7.2) *are recursively given by*

$$\mu_{t|t} = \mu_{t|t-1} + A_t(Y_t - G\mu_{t|t-1}),$$
$$Q_{t|t} = (I_v - A_t G)Q_{t|t-1}$$

with a $(v \times w)$*-credibility matrix*

$$A_t = Q_{t|t-1}G' \left(GQ_{t|t-1}G' + R\right)^{-1}.$$

The movements from t *to* $t+1$ *are given by*

$$\mu_{t+1|t} = F\mu_{t|t},$$
$$Q_{t+1|t} = FQ_{t|t}F' + Q.$$

Proof. See (Bühlmann and Gisler, 2005, Theorem 10.3). □

The credibility estimator $\mu_{T+1|T}$ results by iterating this procedure for $t = 1, \ldots, T$. Finally, $\mu_{T+k|T}$ and the credibility estimator for $\mathbb{E}[Y_{T+k} \mid \Delta_{T+k}] = G\Delta_{T+k}$ are given by

$$\mu_{T+k|T} = F^{k-1}\mu_{T+1|T},$$
$$\mathrm{Pro}\left(G\Delta_{T+k} \mid L(1, Y_1, \ldots, Y_T)\right) = G\mu_{T+k|T}$$

respectively. Theorem 7.1 is also known as the Kalman recursion or the Kalman filter algorithm, cf. Brockwell and Davis (2006), and is implemented in the statistical software R.

7.2. Mortality improvement rates

Age-specific improvement rates. We assume that we observe age-specific mortality statistics over an age range of x_1 to x_n. Here, n is the number of age groups included in the analysis. The mortality data relates to calendar years 1 to T, and we are now at the beginning of the year $T+1$. For each age $x = x_1, \ldots, x_n$ and year $t = 1, \ldots, T$, we calculate the crude death rate $m_x(t)$ as the ratio of the number of deaths over the initial exposure-to-risk. Our aim is to project future rates $m_x(T+1), m_x(T+2), \ldots$ from the observed rates $m_x(1), \ldots, m_x(T)$.

Our proposed evolutionary credibility model is based on log mortality rate changes rather than levels. Specifically, define

$$r_{xt} := \log m_x(t) - \log m_x(t-1)$$

as the log improvement rate in mortality at age x from year $t-1$ to year t. Then, we decompose r_{xt} into

$$r_{xt} = \beta_x \Delta_t + \epsilon_{xt} \tag{7.6}$$

where Δ_t is Normal distributed with mean δ and variance σ_Δ^2, written $\Delta_t \sim \mathcal{N}(\delta, \sigma_\Delta^2)$. We assume that Δ_t appearing in (7.6) obeys some time series model and the error terms ϵ_{xt} are independent with $\epsilon_{xt} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Furthermore, the errors ϵ_{xt} are supposed to be independent from the time factor $\Delta = (\Delta_t)_{t \in \mathbb{N}}$.

By interpreting the latent random variables Δ_t as a risk profile process, our model (7.6) has the natural form of an evolutionary credibility model with $(r_{x_1t}, \ldots, r_{x_nt})'$ being an x_n-variate observation process. Conditional on Δ_t, the mortality improvement rates $(r_{xt})_{x=x_1}^{x_n}$ are independent for fixed t, but without conditioning on Δ_t, they are serially correlated. Credibility estimation of future mortality rates depend on past observation through the recursive credibility formula. In the particular setting of Normal error terms, we even obtain a predictive distribution, see Section 8.1. The primary reason for the Normal assumption is to determine the dynamics of Δ. We will in particular use a likelihood based information criterion to perform an ARMA model selection procedure and this requires the specification of the underlying distribution. Notice that the recursive credibility formula is only based on structures (7.3) to (7.5) and abstains from further distributional conditions.

The time factor Δ reflects the general variation of mortality. The non-negative parameters β_x measure the sensitivity of mortality at age x with respect to calendar time. As the specification (7.6) is not identifiable, some constraints are needed. This is why we adopt in the remainder the standard constraint

$$\sum_{x=x_1}^{x_n} \beta_x = 1. \tag{7.7}$$

Considering (7.6) and the assumptions made so far, we see that the correlation of mortality improvements at different ages

$$\operatorname{corr}(r_{x_1t}, r_{x_2t}) = \frac{\beta_{x_1}\beta_{x_2}\sigma_\Delta^2}{\sqrt{\beta_{x_1}^2\sigma_\Delta^2 + \sigma_\epsilon^2}\sqrt{\beta_{x_2}^2\sigma_\Delta^2 + \sigma_\epsilon^2}},$$

covers the entire range $[0, 1]$ when σ_Δ^2 and σ_ϵ^2 vary.

Aggregate mortality improvement rates. If we sum the age-specific mortality improvement rates, the coefficients β_x disappear because they add up to 1 according to (7.7). More precisely, we define the aggregate errors

$$\epsilon_{\bullet t} := \sum_{x=x_1}^{x_n} \epsilon_{xt},$$

which obey a Normal distribution with zero mean and variance

$$\sigma_\bullet^2 := n\sigma_\epsilon^2.$$

Aggregate errors are mutually independent and independent of Δ_t. Considering (7.7), summing over x the identity (7.6) gives

$$r_{\bullet t} := \sum_{x=x_1}^{x_n} r_{xt} = \Delta_t + \epsilon_{\bullet t} \tag{7.8}$$

and it immediately follows that $r_{\bullet t} \sim \mathcal{N}(\delta, \sigma_\Delta^2 + \sigma_\bullet^2)$. In the empirical illustrations, we first consider the observed $r_{\bullet 1}, \ldots, r_{\bullet T}$ and we fit model (7.8). The advantage of this approach is that we are allowed to study the dynamics of Δ_t describing improvement rates from the aggregate (7.8) involving the global improvement $r_{\bullet t}$ and not the detailed age structure β_x.

Covariance structure. The stochastic process $\Delta = (\Delta_t)_{t\in\mathbb{N}}$ is assumed to be stationary Gaussian. The random vector $(\Delta_1, \ldots, \Delta_T)'$ thus obeys the multivariate Normal distribution with mean vector

$$\delta \mathbf{1}_T = (\delta, \ldots, \delta)',$$

where $\mathbf{1}_T = (1, \ldots, 1)' \in \mathbb{R}^T$, and covariance matrix of the following Toeplitz form:

$$\mathrm{Cov}(\Delta_t, \Delta_s) = \rho_{|t-s|}\sigma_\Delta^2 \tag{7.9}$$

for correlation parameters $\rho_h \in [-1, 1]$, $h = 1, 2, \ldots$, and $\rho_0 = 1$. In matrix notation (7.9) reads

$$\sigma_\Delta^2 \boldsymbol{C}_T = \sigma_\Delta^2 \begin{pmatrix} 1 & \rho_1 & \ldots & \rho_{T-1} \\ \rho_1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_1 \\ \rho_{T-1} & \ldots & \rho_1 & 1 \end{pmatrix} \tag{7.10}$$

with correlation matrix $\boldsymbol{C}_T \in \mathbb{R}^{T,T}$. Note that the specification (7.9) has been also proposed by Sundt (1981) in a credibility context with an autoregressive structure, i.e. assuming

$$\rho_h = \rho^h, \quad h \in \mathbb{N}, \tag{7.11}$$

for some correlation parameter $\rho \in [-1, 1]$.

The model specification (7.8) directly implies that $(r_{\bullet 1}, \ldots, r_{\bullet T})'$ is multivariate Normal with mean vector

$$\delta \mathbf{1}_T = (\delta, \ldots, \delta)' \tag{7.12}$$

and covariance matrix

$$\sigma_\bullet^2 I_T + \sigma_\Delta^2 \boldsymbol{C}_T. \tag{7.13}$$

In (7.13), I_T denotes the $T \times T$ identity matrix. However, the covariance structure (7.13) may not be identifiable or in other words, not one-to-one with respect to its variance parameters. Let us make this point clear in the following example.

Example 7.2. Assume that the time factor Δ obeys the MA(1)-process

$$\Delta_t = Z_t + \theta Z_{t-1}$$

with independent innovation terms $Z_t \sim \mathcal{N}(0, \sigma_Z^2)$. Then, as

$$\mathrm{Var}(\Delta_t) = (1 + \theta^2)\sigma_Z^2 = \sigma_\Delta^2,$$

we see that σ_Δ^2 is implicitly given through the error variance σ_Z^2 and the MA-parameter θ. Moreover,

$$\mathrm{Cov}(\Delta_{t-1}, \Delta_t) = \theta\sigma_Z^2 = \frac{\theta}{1+\theta^2}\sigma_\Delta^2$$

so that

$$\rho_h = \begin{cases} \frac{\theta}{1+\theta^2} & \text{if } h = 1, \\ 0 & \text{if } h \geq 2. \end{cases}$$

If we replace σ_Δ^2 by $\tilde{\sigma}_\Delta^2$, the triples $(\theta, \sigma_\Delta^2, \sigma_\bullet^2)$ and $(\tilde{\theta}, \tilde{\sigma}_\Delta^2, \tilde{\sigma}_\bullet^2)$ with

$$\tilde{\sigma}_\bullet^2 = \sigma_\bullet^2 + \sigma_\Delta^2 - \tilde{\sigma}_\Delta^2$$

and $\tilde{\theta}$ as the solution of

$$\frac{\tilde{\theta}}{1+\tilde{\theta}^2} = \frac{\theta}{1+\theta^2}\frac{\sigma_\Delta^2}{\tilde{\sigma}_\Delta^2}$$

will produce the same covariance matrix (7.13) provided that $\tilde{\sigma}_\bullet^2 > 0$. In fact, it is easy to verify that

$$\mathrm{Cov}(r_{\bullet t-1}, r_{\bullet t}) = \mathrm{Cov}(\Delta_{t-1}, \Delta_t) = \frac{\tilde{\theta}}{1+\tilde{\theta}^2}\tilde{\sigma}_\Delta^2 = \frac{\theta}{1+\theta^2}\sigma_\Delta^2$$

and

$$\mathrm{Cov}(r_{\bullet t}, r_{\bullet t}) = \tilde{\sigma}_\bullet^2 + \tilde{\sigma}_\Delta^2 = \left(\sigma_\bullet^2 + \sigma_\Delta^2 - \tilde{\sigma}_\Delta^2\right) + \tilde{\sigma}_\Delta^2 = \sigma_\bullet^2 + \sigma_\Delta^2.$$

There may exist identifiable ARMA specifications for the time factor Δ. However, it is not clear whether such specifications appropriately explain the true dynamics of $r_\bullet$. In order to solve the identifiability issue, we now consider both genders together, as explained next.

Gender-combined mortality improvement factors. Instead of considering males and females separately, we now integrate both gender-specific improvement rates into a single model. In addition to ensuring identifiability of the covariance structure, this approach also enforces consistency between genders. In insurance applications, it allows the actuary to evaluate potential diversification benefits between male and female future mortality improvements.

Our model specification is as follows. Let $r_{xt}^{(m)}$ denote the mortality improvement rates for males and let $r_{xt}^{(f)}$ denote the corresponding mortality improvement rates for females from the same country. We now assume that the model (7.6) applies to both genders, i.e. that

$$r_{xt}^{(i)} = \beta_x^{(i)}\Delta_t^{(i)} + \epsilon_{xt}^{(i)} \quad \text{for } i \in \{m, f\} \tag{7.14}$$

holds with a certain dependence structure between the two $\Delta^{(i)}$-processes specified later on. The parameters $\beta_x^{(i)}$ add up to 1 for each $i \in \{m, f\}$ in accordance with (7.7). Furthermore, the error terms $\epsilon_{xt}^{(i)}$ in (7.14) are assumed to be mutually independent with distribution $\mathcal{N}(0, \sigma_{i\epsilon}^2)$. The corresponding aggregate structure is then given by

$$r_{\bullet t}^{(i)} = \Delta_t^{(i)} + \epsilon_{\bullet t}^{(i)} \quad \text{for } i \in \{m, f\}, \tag{7.15}$$

where $\Delta_t^{(i)} \sim \mathcal{N}(\delta_i, \sigma_{i\Delta}^2)$.

State-space representation. We consider the multivariate process

$$\boldsymbol{r}_{\bullet t} = \begin{pmatrix} r_{\bullet t}^{(m)} \\ r_{\bullet t}^{(f)} \end{pmatrix} = \begin{pmatrix} \Delta_t^{(m)} \\ \Delta_t^{(f)} \end{pmatrix} + \begin{pmatrix} \epsilon_{\bullet t}^{(m)} \\ \epsilon_{\bullet t}^{(f)} \end{pmatrix}$$

in state-space form in order to express it as an evolutionary credibility model. Both time factors $\Delta^{(m)}$ and $\Delta^{(f)}$ are supposed to obey some autoregressive moving average (ARMA) dynamics. Common state-space representations of ARMA processes can be easily extended to that of $(\boldsymbol{r}_{\bullet t})$. First, consider (Y_t) following an ARMA(p, q) process with autoregressive (AR) parameters $\phi_1, \ldots, \phi_p$, moving average (MA) parameters

$\theta_1, \ldots, \theta_q$ and innovations Z_t. Let $d = \max\{p, q+1\}$ and set $\phi_k = 0$ for $k > p$ and $\theta_k = 0$ for $k > q$. Hamilton (1994) showed that we can represent Y_t as

$$Y_t = \begin{pmatrix} 1 & \theta_1 & \cdots & \theta_{d-1} \end{pmatrix} \begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-d+1} \end{pmatrix}$$

with state equation

$$\begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-d+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \cdots & \phi_{d-1} & \phi_d \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-d} \end{pmatrix} + \begin{pmatrix} Z_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Let us now extend this representation to $\boldsymbol{r}_{\bullet t}$. To this end, define

$$d := \max\{p_m, q_m + 1, p_f, q_f + 1\},$$

where (p_i, q_i) is the ARMA order of $\Delta^{(i)}$. The gender specific ARMA parameters are denoted by an additional superscript (i). Then, we have observation equation

$$\begin{aligned} \begin{pmatrix} r_{\bullet t}^{(m)} \\ r_{\bullet t}^{(f)} \end{pmatrix} - \begin{pmatrix} \delta_m \\ \delta_f \end{pmatrix} &= \begin{pmatrix} 1 & \theta_1^{(m)} & \cdots & \theta_{d-1}^{(m)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & \theta_1^{(f)} & \cdots & \theta_{d-1}^{(f)} \end{pmatrix} \begin{pmatrix} X_t^{(m)} \\ \vdots \\ X_{t-d+1}^{(m)} \\ X_t^{(f)} \\ \vdots \\ X_{t-d+1}^{(f)} \end{pmatrix} + \begin{pmatrix} \epsilon_{\bullet t}^{(m)} \\ \epsilon_{\bullet t}^{(f)} \end{pmatrix} \\ &=: \boldsymbol{G}\boldsymbol{X}_t + \boldsymbol{W}_t \end{aligned} \tag{7.16}$$

with state equation

$$\begin{aligned} \begin{pmatrix} X_t^{(m)} \\ \vdots \\ X_{t-d+1}^{(m)} \\ X_t^{(f)} \\ \vdots \\ X_{t-d+1}^{(f)} \end{pmatrix} &= \begin{pmatrix} \phi_1^{(m)} & \cdots & \phi_{d-1}^{(m)} & \phi_d^{(m)} & 0 & \cdots & 0 & 0 \\ 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 0 & \phi_1^{(f)} & \cdots & \phi_{d-1}^{(f)} & \phi_d^{(f)} \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1}^{(m)} \\ \vdots \\ X_{t-d}^{(m)} \\ X_{t-1}^{(f)} \\ \vdots \\ X_{t-d}^{(f)} \end{pmatrix} + \begin{pmatrix} Z_t^{(m)} \\ 0 \\ \vdots \\ 0 \\ Z_t^{(f)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \\ &=: \boldsymbol{F}\boldsymbol{X}_{t-1} + \boldsymbol{V}_t. \end{aligned} \tag{7.17}$$

From a strict formal point of view, the process $(\mathbf{X}_t)$ is the actual Δ-process in sense of the model structure (7.3) and (7.4). In terms of our model structure (7.15), we have

$$\begin{pmatrix} \Delta_t^{(m)} \\ \Delta_t^{(f)} \end{pmatrix} = \boldsymbol{G}\boldsymbol{X}_t + \begin{pmatrix} \delta_m \\ \delta_f \end{pmatrix}$$

and this allows a more natural interpretation of the state process. Equations (7.16) and (7.17) combine the state-space representations of $\Delta^{(m)}$ and $\Delta^{(f)}$ to a joint structure

and dependencies between the gender stem from the innovation errors $Z_t^{(i)} \sim \mathcal{N}(0, \sigma_{iZ}^2)$. They are correlated through a gender correlation parameter $\gamma \in [-1, 1]$, i.e.

$$\operatorname{Cov}\left(Z_t^{(m)}, Z_t^{(f)}\right) = \gamma\sqrt{\sigma_{mZ}^2\sigma_{fZ}^2}.$$

This defines the covariance matrix of the random vector $(\boldsymbol{V}_t', \boldsymbol{W}_t')'$ as

$$\begin{pmatrix} \boldsymbol{Q} & 0 \\ 0 & \boldsymbol{R} \end{pmatrix},$$

where

$$\boldsymbol{Q} := \begin{pmatrix} \sigma_{mZ}^2 & 0 & \cdots & 0 & \gamma\sqrt{\sigma_{mZ}^2\sigma_{fZ}^2} & 0 & \cdots & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \gamma\sqrt{\sigma_{mZ}^2\sigma_{fZ}^2} & 0 & \cdots & 0 & \sigma_{fZ}^2 & 0 & \cdots & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \end{pmatrix}, \tag{7.18}$$

$$\boldsymbol{R} := \begin{pmatrix} \sigma_{m\bullet}^2 & 0 \\ 0 & \sigma_{f\bullet}^2 \end{pmatrix}. \tag{7.19}$$

Equations (7.16) to (7.19) define the joint state space representation of our credibility model and this model has been also proposed by Schinzinger et al. (2014). The model is kept general and allows for different degrees of homogeneity between the genders. A higher degree of homogeneity can be achieved by making one or more of the following simplifying assumptions.

- (S1) $\Delta^{(m)}$ and $\Delta^{(f)}$ are ARMA processes of the same order (p, q) and share the common ARMA parameters $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$.
- (S2) The gender correlation parameter is given by $\gamma = 1$.
- (S3) The parameters δ_i, $\sigma_{i\Delta}^2$ (or equivalently σ_{iZ}^2), $\sigma_{i\bullet}^2$ do not depend on the gender i.

All these assumptions can be easily incorporated into the above state-space representation. Activating all of them leads to the particular case $\Delta_t^{(m)} = \Delta_t^{(f)}$ in which the mortality improvement factors applying to males and females are both functions of a single Δ_t, i.e.

$$r_{xt}^{(i)} = \beta_x^{(i)}\Delta_t + \epsilon_{xt}^{(i)}, \quad i \in \{m, f\}. \tag{7.20}$$

This case is of particular interest as Carter and Lee (1992) suggested to use the same time index for both genders. To avoid long-run divergence in gender-specific mortality forecasts, Li and Lee (2005) further proposed to use the same β_x for all groups. Here, we nevertheless allow for gender-specific sensitivities $\beta_x^{(i)}$ and leave the final decision to the user.

Specifying the joint dynamics of $(r_{\bullet t}^{(m)}, r_{\bullet t}^{(f)})$ now solves the identifiability issue.

Lemma 7.3. *Assume that (S1) holds, i.e. we have gender common ARMA parameters. Furthermore, suppose that the $\Delta^{(i)}$ are causal. Then,*

$$\operatorname{Cov}(\Delta_t^{(m)}, \Delta_s^{(f)}) = \gamma \rho_{|t-s|} \sigma_\Delta^2, \tag{7.21}$$

where

$$\sigma_\Delta^2 = \sqrt{\sigma_{m\Delta}^2 \sigma_{f\Delta}^2}$$

and ρ_h given through the gender individual covariance structure (7.9)*.*

Proof. First assume that both $\Delta^{(i)}$ follow a $\mathrm{MA}(q)$ process with parameters $\theta_0, \dots, \theta_q$ with $\theta_0 = 1$. Then, for $h \in \mathbb{N}$ and $h \leq q$, the covariance structure (7.18) of the innovation terms yields that

$$\begin{aligned}
\operatorname{Cov}\left(\Delta_{t+h}^{(i)}, \Delta_t^{(i)}\right) &= \operatorname{Cov}\left(\sum_{j=0}^{q} \theta_j Z_{t+h-j}^{(i)}, \sum_{j=0}^{q} \theta_j Z_{t-j}^{(i)}\right) \\
&= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k \operatorname{Cov}\left(Z_{t+h-j}^{(i)}, Z_{t-k}^{(i)}\right) \\
&= \sum_{j=0}^{q-h} \sum_{k=0}^{q} \theta_{j+h} \theta_k \operatorname{Cov}(Z_{t-j}^{(i)}, Z_{t-k}^{(i)}) \\
&= \left(\sum_{j=0}^{q-h} \theta_{j+h} \theta_j\right) \sigma_{iZ}^2
\end{aligned}$$

Taking $h = 0$ gives

$$\sigma_{i\Delta}^2 = \operatorname{Var}\left(\Delta_t^{(i)}\right) = \left(\sum_{j=0}^{q} \theta_j^2\right) \sigma_{iZ}^2$$

and therefore by $\operatorname{Cov}(\Delta_{t+h}^{(i)}, \Delta_t^{(i)}) = \rho_h \sigma_{i\Delta}^2$,

$$\rho_h = \left(\sum_{j=0}^{q-h} \theta_{j+h} \theta_j\right) \left(\sum_{j=0}^{q} \theta_j^2\right)^{-1}.$$

Similarly, since $\operatorname{Cov}(Z_t^{(m)}, Z_t^{(f)}) = \gamma \sqrt{\sigma_{mZ}^2 \sigma_{fZ}^2}$,

$$\begin{aligned}
\operatorname{Cov}\left(\Delta_{t+h}^{(m)}, \Delta_t^{(f)}\right) &= \sum_{j=0}^{q-h} \sum_{k=0}^{q} \theta_{j+h} \theta_k \operatorname{Cov}(Z_{t-j}^{(m)}, Z_{t-k}^{(f)}) \\
&= \gamma \left(\sum_{j=0}^{q-h} \theta_{j+h} \theta_j\right) \sqrt{\sigma_{mZ}^2 \sigma_{fZ}^2} \\
&= \gamma \left(\sum_{j=0}^{q-h} \theta_{j+h} \theta_j\right) \left(\sum_{j=0}^{q} \theta_j^2\right)^{-1} \sqrt{\left(\sum_{j=0}^{q} \theta_j^2\right) \sigma_{mZ}^2 \left(\sum_{j=0}^{q} \theta_j^2\right) \sigma_{fZ}^2} \\
&= \gamma \rho_h \sigma_\Delta^2.
\end{aligned}$$

The case $(-h) \in \mathbb{N}$ works in the same way such that (7.21) holds for any MA(q) process.

The result can be generalized using the moving average representation of ARMA(p, q) models. Since the $\Delta^{(i)}$ are causal, there exist constants $\psi_j \in \mathbb{R}$, $j \in \mathbb{N}$, with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that

$$\Delta_t^{(i)} - \delta_i = \sum_{j=0}^{\infty} \psi_j Z_{t-j}^{(i)} \tag{7.22}$$

almost surely and in $\mathcal{L}^2$. See Brockwell and Davis (2006) for details. Thus, (7.21) follows by taking $q \to \infty$ and replacing θ_j by ψ_j in the above calculations. □

Rewriting (7.21) in matrix notation, the covariance matrix between the vectors $(\Delta_1^{(m)}, \dots, \Delta_T^{(m)})'$ and $(\Delta_1^{(f)}, \dots, \Delta_T^{(f)})'$ is $\gamma\sigma_\Delta^2 \boldsymbol{C}_T$, where $\boldsymbol{C}_T$ has been defined in (7.10). The structure directly follows from Lemma 7.3. Then, the gender-combined random vector

$$\boldsymbol{r}_\bullet = (r_{\bullet 1}^{(m)}, \dots, r_{\bullet T}^{(m)}, r_{\bullet 1}^{(f)}, \dots, r_{\bullet T}^{(f)})'. \tag{7.23}$$

of past observed aggregate improvement rates is multivariate Normal with mean vector

$$\boldsymbol{\delta}_\bullet = (\delta_m, \dots, \delta_m, \delta_f, \dots, \delta_f)' \tag{7.24}$$

and covariance matrix

$$\boldsymbol{\Sigma}_\bullet = \begin{pmatrix} \sigma_{m\bullet}^2 I_T + \sigma_{m\Delta}^2 \boldsymbol{C}_T & \gamma\sigma_\Delta^2 \boldsymbol{C}_T \\ \gamma\sigma_\Delta^2 \boldsymbol{C}_T & \sigma_{f\bullet}^2 I_T + \sigma_{f\Delta}^2 \boldsymbol{C}_T \end{pmatrix}. \tag{7.25}$$

The additional parameter γ now solves the identifiability problem of $\boldsymbol{\Sigma}_\bullet$ and the variance parameters.

Theorem 7.4. *In addition to the conditions of Lemma 7.3, assume that the ARMA order (p, q) satisfies $T - 1 > p + q$. If $\gamma \neq 0$, then $\boldsymbol{\Sigma}_\bullet$ is uniquely determined by the model parameters, i.e $\boldsymbol{\Sigma}_\bullet = \tilde{\boldsymbol{\Sigma}}_\bullet$ implies*

$$(\sigma_{m\bullet}^2, \sigma_{f\bullet}^2, \sigma_{m\Delta}^2, \sigma_{f\Delta}^2, \gamma, \rho_1, \dots, \rho_p) = (\tilde{\sigma}_{m\bullet}^2, \tilde{\sigma}_{f\bullet}^2, \tilde{\sigma}_{m\Delta}^2, \tilde{\sigma}_{f\Delta}^2, \tilde{\gamma}, \tilde{\rho}_1, \dots, \tilde{\rho}_p)$$

where $\tilde{\boldsymbol{\Sigma}}_\bullet$ is the covariance matrix corresponding to the alternative parameters.

Proof. Assume that $\boldsymbol{\Sigma}_\bullet = \tilde{\boldsymbol{\Sigma}}_\bullet$ and recall structure (7.10) of the correlation matrix $\boldsymbol{C}_T$. The structure of the diagonal block matrices of (7.25) implies that

$$\sigma_{i\Delta}^2 \rho_h = \tilde{\sigma}_{i\Delta}^2 \tilde{\rho}_h \tag{7.26}$$

for $i \in \{m, f\}$ and $h = 1, \dots, T - 1$. The ρ_h and $\tilde{\rho}_h$ are the entries of $\boldsymbol{C}_T$ and $\tilde{\boldsymbol{C}}_T$ respectively. Multiplying (7.26) for males and females and taking the square roots provide the identities

$$\sqrt{\sigma_{m\Delta}^2 \sigma_{f\Delta}^2}\rho_h = \sqrt{\tilde{\sigma}_{m\Delta}^2 \tilde{\sigma}_{f\Delta}^2}\tilde{\rho}_h, \tag{7.27}$$

$h = 1, \dots, T - 1$. On the other hand, also the off diagonal matrices $\gamma\sigma_\Delta^2 \boldsymbol{C}_T$ and $\tilde{\gamma}\tilde{\sigma}_\Delta^2 \tilde{\boldsymbol{C}}_T$ agree. Therefore,

$$\gamma\sqrt{\sigma_{m\Delta}^2 \sigma_{f\Delta}^2}\rho_h = \tilde{\gamma}\sqrt{\tilde{\sigma}_{m\Delta}^2 \tilde{\sigma}_{f\Delta}^2}\tilde{\rho}_h,$$

$h = 1, \dots, T - 1$, and it follows from (7.27) and $\gamma \neq 0 \neq \tilde{\gamma}$ that $\gamma = \tilde{\gamma}$. The condition $T - 1 > p + q$ enures that $\rho_{h_1} \neq \rho_{h_2}$ and $\tilde{\rho}_{h_1} \neq \tilde{\rho}_{h_2}$ for some $h_1 \neq h_2$. The identities of

the further parameters follow step by step. Considering the diagonal entries of $\gamma\sigma_\Delta^2 \boldsymbol{C}_T$ and its counterpart, we obtain

$$\gamma\sqrt{\sigma_{m\Delta}^2\sigma_{f\Delta}^2} = \tilde{\gamma}\sqrt{\tilde{\sigma}_{m\Delta}^2\tilde{\sigma}_{f\Delta}^2}$$

Since $\gamma = \tilde{\gamma}$,

$$\sqrt{\sigma_{m\Delta}^2\sigma_{f\Delta}^2} = \sqrt{\tilde{\sigma}_{m\Delta}^2\tilde{\sigma}_{f\Delta}^2}$$

holds, which in turn proves

$$\rho_h = \tilde{\rho}_h, \quad h = 1, \ldots, T-1, \tag{7.28}$$

using (7.27). Relation (7.26) then provides

$$\sigma_{i\Delta}^2 = \tilde{\sigma}_{i\Delta}^2, \quad i \in \{m, f\}$$

and $\sigma_{i\bullet}^2 = \tilde{\sigma}_{i\bullet}^2$ follows from

$$\sigma_{i\bullet}^2 + \sigma_{i\Delta}^2 = \tilde{\sigma}_{i\bullet}^2 + \tilde{\sigma}_{i\Delta}^2,$$

which are the diagonal entries of $\sigma_{i\bullet}^2 I_T + \sigma_{i\Delta}^2 \boldsymbol{C}_T$ and that of the alternative parameters respectively. □

Both results rely on the simplifying assumption (S1) which is indeed justified by the upcoming empirical below. Therefore, we do not provide a general covariance structure for the case that ARMA dynamics for males and females differ.

7.3. Application to Belgian mortality data

Presentation of the data. We consider mortality data for Belgian males and females available from Statistics Belgium (`http://statbel.fgov.be/`). Many insurance applications consider ages after retirement to study various issues about pension benefits. The data set considered here thus consists of ages $x_1 = 65$ to $x_n = 99$ ($n = 35$) observed in the time period from 1970 to 2010. Observations before 1970 are not included since there is a structural break in the 70s as documented in Coelho and Nunes (2011). Thus, $t = 1$ corresponds to the mortality improvement from calendar year 1970 to 1971 whereas T corresponds to that from 2009 to 2010. In a later stage of our analysis, we supplement these observations with calendar years 2011 and 2012 to study the robustness over successive forecasts. However, notice that the model selection procedure is generally applicable for any choices of ages and years in fit.

Figure 7.1 displays the observed age-aggregated mortality improvements $r_{\bullet t}^{(i)}$ for males and females. Both series appear to be strongly correlated. Mortality statistics depicted in Figure 7.1 indicate negative correlation between $r_{\bullet t}^{(i)}$ and $r_{\bullet t+1}^{(i)}$. This property is a consequence of the typical zigzag pattern, i.e. large improvements in mortality rates are followed by small improvements (or even declines) and vice verse. This apparent behavior also rules out time-invariant random effects in (7.6), i.e. $\Delta_t^{(i)} = \Delta^{(i)}$ for all $t \in \mathbb{N}$, as this specification implies $\mathrm{Cov}(r_{\bullet t}^{(i)}, r_{\bullet s}^{(i)}) = \mathrm{Var}(\Delta^{(i)}) > 0$ for all $t \neq s$. Hence, $\Delta_t^{(i)} = \Delta^{(i)}$ constraints $r_{\bullet t}^{(i)}$ and $r_{\bullet s}^{(i)}$ to be positively correlated among all years t and s which contradicts empirical evidence in Figure 7.1

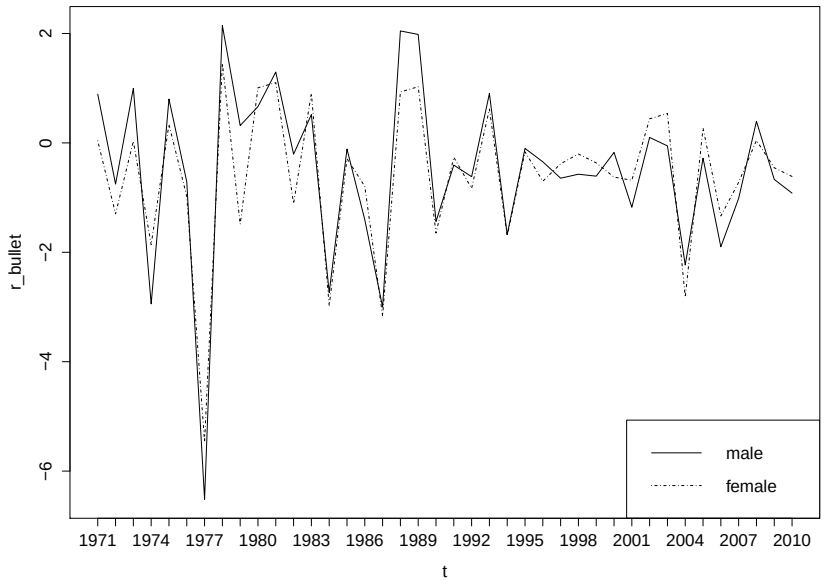

FIGURE 7.1. Age-aggregated mortality improvements $r^{(m)}_{\bullet t}$ for Belgian males and $r^{(f)}_{\bullet t}$ for females.

Model assumptions. First, we check whether the Belgian data satisfies the model assumptions specified in Section 7. Specifically, we investigate whether the data agree with the assumptions of Normality and stationarity in the age-aggregate model (7.8).

We test both $r^{(m)}_{\bullet t}$ and $r^{(f)}_{\bullet t}$ for Normality using the Shapiro-Wilk test. The p-values for both genders are 0.0038 and 0.0036 so that the hypotheses of Normal distribution are rejected. However, if we omit the observations with the largest deviation from the sample mean in both series, we get the p-values 0.5752 and 0.2281 respectively. The hypotheses cannot be rejected. The Q-Q plots in Figure 7.3 confirm this observation. Concerning stationarity, the augmented Dickey-Fuller test provides small p-values for the hypothesis of non-stationarity. The calculated p-values are 0.023 and 0.01037 for males and females respectively, where the lag has been chosen as $(T-1)^{1/3}$. Thus, the hypothesis of non-stationarity is rejected for both genders.

Model selection. As our models are fully specified, with Normal distributed components, the maximum likelihood approach is expected to deliver accurate estimations. As motivated, we first fit the age-aggregate model to study period and age-effects separately. Let $\boldsymbol{r}_\bullet$ gather the observed aggregate improvement rates as defined in (7.23). The log-likelihood function

$$\log L = -\frac{1}{2}\log|\boldsymbol{\Sigma}_\bullet| - \frac{1}{2}\left(\boldsymbol{r}_\bullet - \boldsymbol{\delta}_\bullet\right)\boldsymbol{\Sigma}_\bullet^{-1}\left(\boldsymbol{r}_\bullet - \boldsymbol{\delta}_\bullet\right)' \tag{7.29}$$

of a $2T$-variate Normal distribution with mean vector $\boldsymbol{\delta}_\bullet$ and covariance matrix $\boldsymbol{\Sigma}_\bullet$ has to be maximized with respect to the model parameters. Notice that the input variables for the optimization are the ARMA parameters and the innovation variances σ^2_{iZ} rather than the correlation parameters ρ_h and the variance terms $\sigma^2_{i\Delta}$. The latter quantities

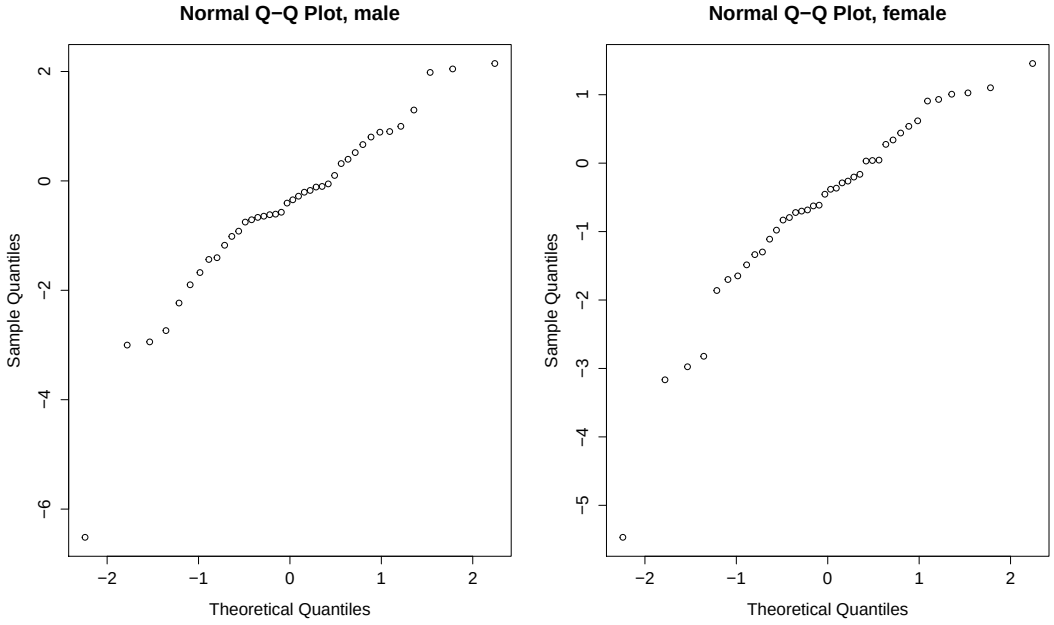

FIGURE 7.2. Normal Q-Q plots of the age-aggregated mortality improvement rates.

can be deduced from the input variables using the identities

$$\rho_h = \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2} \tag{7.30}$$

and

$$\sigma_{i\Delta}^2 = \sigma_{iZ}^2 \sum_{j=0}^{\infty} \psi_j^2,$$

where $(\psi_j)_{j\in\mathbb{N}}$ comes from the corresponding MA(∞)-representation, see (Brockwell and Davis, 2006, Section 3.3). The actual evaluation of the log-likelihood function is performed with the `FKF`-package in `R`. Instead of (7.29), the likelihood is calculated by means of the fitted residuals

$$\begin{pmatrix} \epsilon_{\bullet t}^{(m)} \\ \epsilon_{\bullet t}^{(f)} \end{pmatrix} = \begin{pmatrix} r_{\bullet t}^{(m)} - \hat{r}_{\bullet t}^{(m)} \\ r_{\bullet t}^{(f)} - \hat{r}_{\bullet t}^{(f)} \end{pmatrix},$$

where $\hat{r}_{\bullet t}^{(i)}$ are the best predictions obtained by the Kalman filter. This method avoids the computation and the inversion of the huge covariance matrix $\mathbf{\Sigma}$. Its implementation only requires the specification of a state-space representation instead of manually defining $\mathbf{\Sigma}$.

The model selection procedure then follows a backward approach. We start from a first model allowing for dynamics specific to each gender and we simplify it by activating the assumptions (S1) to (S3) step by step. Each time, we evaluate the candidate model by using the Akaike information criterion with correction for finite samples (AICc). Recall that

$$\text{AICc} = -2 \log L + 2k + \frac{2k(k+1)}{T-k-1},$$

	δ_m	δ_f	$\sigma^2_{m\Delta}$	$\sigma^2_{f\Delta}$	$\sigma^2_{m\bullet}$	$\sigma^2_{f\bullet}$	ϕ_1	θ_1	γ
Model 3, MA(1)	-0.504	-0.605	2.139	1.812	0.347	10^{-7}	-	-0.459	1
Model 5, ARMA(1,1)	-0.554		1.920		0.180		0.429	-0.999	1

TABLE 7.1. Parameter estimates for Models 3 and 5.

where k is the dimension of the parameter space.

Model 1. We first fit ARMA models for both genders separately, i.e. we model $(r^{(m)}_{\bullet 1}, \dots, r^{(m)}_{\bullet T})$ and $(r^{(f)}_{\bullet 1}, \dots, r^{(f)}_{\bullet T})$ in isolation following the Box-Jenkins methodology. Although the marginal models face the identifiability problem described earlier and maximum likelihood estimates are not unique, the AICc values can still be taken for comparison. We restrict our analysis to ARMA(p, q) models with $0 \leq p + q \leq 5$. This also covers the previously mentioned AR(1) structure characterized by the exponentially-decaying correlations (7.11). Models of higher orders are not shown here since additional parameters were not significant and the corresponding AICc values did not show any improvement over those of the lower-order ARMA models.
Minimum AICc values are attained at ARMA(1,1) for males and MA(1) for females. The joint likelihood under the gender-combined structure yields an AICc of 228.5117.

Model 2, (S1). Let us now assume that $r^{(m)}_{\bullet t}$ and $r^{(f)}_{\bullet t}$ follow the same ARMA(p, q) dynamics with common ARMA parameters, i.e. that (S1) holds. Several ARMA models are tested by maximizing the likelihood of the joint series of gender-specific aggregate improvement factors, correlated through the γ parameter. It turns out that the MA(1) structure is optimal, with an AICc of 221.5098 outperforming the preceding 228.5117.

Model 3, (S1) and (S2). As the estimated γ in Model 2 appears to be close to 1, we now consider a fixed gender-correlation parameter $\gamma = 1$. The optimal model is again MA(1) and setting $\gamma = 1$ impacts on AICc, which changes from 221.5098 to 218.5918. The estimated model is described in Table 7.1.

Model 4, (S1) and (S3). Now, Model 2 is fitted with gender-common parameters δ, σ^2_Δ and $\sigma^2_\bullet$. The optimal AICc value of 220.9682 then corresponds to the ARMA(1,1) model.

Model 5, (S1) to (S3). We now set γ equal to 1 in Model 4 so that (7.20) holds. The ARMA(1,1) model fits best with an AICc of 218.3188. The estimations of model parameters are displayed in Table 7.1.

Model validation. Models 3 and 5 are the AICc-best models. We compare them by checking their fitted residuals obtained through the Kalman filter specified by the state-space representation (7.16) to (7.19). The fitted residuals are viewed as realizations of $\epsilon^{(i)}_{\bullet t}$ and are tested for Normality, independence and stationarity. For that purpose, we use the Shapiro-Wilk test, the Box-Ljung test and the ADF-test respectively. The results are given in Table 7.2. The second hypothesis "Normality*" describes the Shapiro-Wilk test, where the absolutely largest residuals were left out. All of the three properties hold when taking a significance level of 0.1. Therefore, we decide to proceed with both models.

	Model 3, MA(1)		Model 5, ARMA(1,1)	
Hypothesis	m	f	m	f
Normality	10^{-4}	0.1324	10^{-4}	0.2129
Normality*	0.5581	-	0.5819	-
Independence	0.7976	0.3647	0.7488	0.6278
Non-stationarity	0.0103	0.0623	0.0584	0.0851

TABLE 7.2. p-values for various tests

7.4. Comparison with the Lee-Carter model

Table 7.3 demonstrates estimates for the age-aggregate mortality improvement model implicitly given by the Lee-Carter model. Recall that in the Lee-Carter framework, the log death rates $\log m_x^{(i)}(t)$, $i \in \{m, f\}$, are decomposed by a principal component analysis into $\alpha_x^{(i)} + \beta_x^{(i)}\kappa_t^{(i)}$ where the time factor $\kappa_t^{(i)}$ obeys an ARIMA dynamics. Therefore,

$$\sum_{x=x_1}^{x_n} \big(\log m_x^{(i)}(t) - \log m_x^{(i)}(t-1)\big) = \sum_{x=x_1}^{x_n} \beta_x^{(i)}(\kappa_t^{(i)} - \kappa_{t-1}^{(i)}) = \kappa_t^{(i)} - \kappa_{t-1}^{(i)}$$

for $i \in \{m, f\}$. As in the majority of empirical studies conducted with the Lee-Carter model, we assume that $\kappa_t^{(i)}$ obeys the random walk with drift model

$$\kappa_t^{(i)} - \kappa_{t-1}^{(i)} = \delta_{i\kappa} + S_t^{(i)}$$

with independent error components $S_t^{(i)} \sim \mathcal{N}(0, \sigma_{i\kappa}^2)$. Furthermore, the residual variance between the observed and fitted model is denoted by $\sigma_{i\circ}^2$ which is the analog term to $\sigma_{i\bullet}^2$ in our model. Even though both models are based on a similar structure, the differences in the estimated values are remarkable. Drift parameters clearly vary from those of the Lee-Carter model. What is even more important is how the total variances $\sigma_{i\bullet}^2 + \sigma_{i\Delta}^2$ and $\sigma_{i\circ}^2 + \sigma_{i\kappa}^2$ are allocated in the two models. While the Lee-Carter mortality improvement model gives more weight to the measurement variance $\sigma_{i\circ}^2$, the innovation variance $\sigma_{i\Delta}^2$ is dominating in our model. Notice that the innovation error affects all ages through the sensitivity factor $\beta_x^{(i)}$.

7.5. Age-specific structure

Given the parameters of the age-aggregate model, we can calibrate the age-specific coefficients $\beta_x^{(i)}$ and the residual variances $\sigma_{i\epsilon}^2$ appearing in the age-specific model (7.14), which was

$$r_{xt}^{(i)} = \beta_x^{(i)}\Delta_t^{(i)} + \epsilon_{xt}^{(i)}.$$

Notice that for each age x, the random vector $(r_{x1}^{(i)}, \ldots, r_{xT}^{(i)})$ is multivariate Normal with mean vector

$$\beta_x^{(i)}\delta_i \mathbf{1}_T = (\beta_x^{(i)}\delta_i, \ldots, \beta_x^{(i)}\delta_i)'$$

and covariance matrix

$$\sigma_{i\epsilon}^2 I_T + \beta_x^2 \sigma_{i\Delta}^2 \boldsymbol{C}_T.$$

	Model 3, MA(1) male	Model 5, ARMA(1,1) male	Lee-Carter male
δ_i	-0.504	-0.554	-0.434
$\sigma^2_{i\Delta}$	2.139	1.920	0.330
$\sigma^2_{i\bullet}$	0.347	0.180	1.7904
	female	female	female
δ_i	-0.605	-0.554	-0.522
$\sigma^2_{i\Delta}$	1.812	1.920	0.6637
$\sigma^2_{i\bullet}$	10^{-7}	0.180	0.6994

TABLE 7.3. Estimated mean and variance parameters. For the Lee-Carter models, the values are the estimates for $\delta_{i\kappa}$, $\sigma^2_{i\kappa}$ and $\sigma^2_{i\circ}$ respectively.

The corresponding Normal log-likelihood function can thus be maximized with respect to the mean $\beta^{(i)}_x\delta_i$ for each age x separately, which gives

$$\widehat{\beta^{(i)}_x\delta_i} = \frac{1}{T}\sum_{t=1}^{T} r^{(i)}_{xt}.$$

As the analysis of the aggregate mortality improvement rates $r^{(m)}_{\bullet t}$ and $r^{(f)}_{\bullet t}$ gives

$$\widehat{\delta}_i = \frac{1}{T}\sum_{t=1}^{T} r^{(i)}_{\bullet t},$$

we finally choose a plug-in estimator for $\beta^{(i)}_x = \frac{\beta^{(i)}_x\delta_i}{\delta_i}$ as

$$\widehat{\beta}^{(i)}_x = \left(\sum_{t=1}^{T} r^{(i)}_{\bullet t}\right)^{-1} \sum_{t=1}^{T} r^{(i)}_{xt}, \tag{7.31}$$

which add up to 1. Hence, constraint (7.7) is satisfied. The estimated parameters are displayed in Figure 7.3 together with their smoothing splines.

For $\sigma^2_{i\epsilon} = \frac{1}{n}\sigma^2_{i\bullet}$, we choose

$$\widehat{\sigma}^2_{i\epsilon} = \frac{1}{n}\widehat{\sigma}^2_{i\bullet}. \tag{7.32}$$

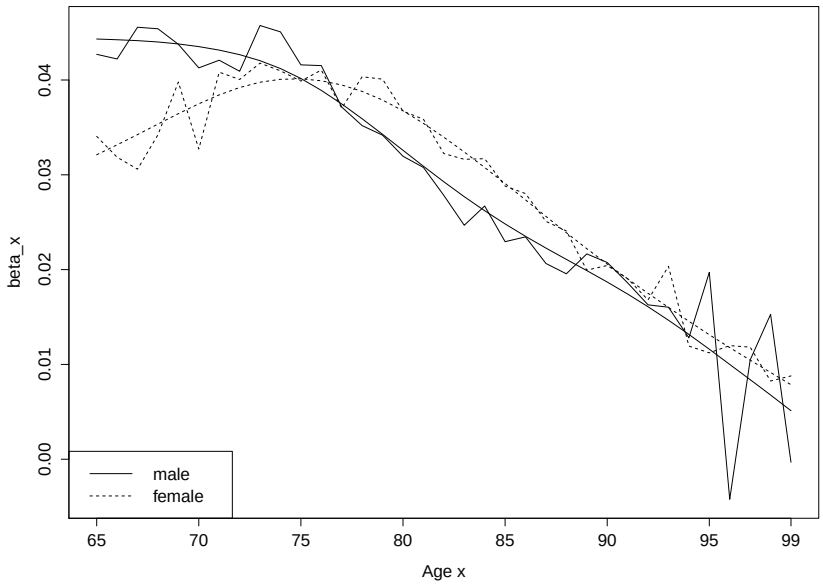

FIGURE 7.3. Estimated age effects $\beta_x^{(i)}$

CHAPTER 8

Mortality forecasting

8.1. Predictive distribution

In this section, we assume that $\Delta^{(m)}$ and $\Delta^{(f)}$ are ARMA processes of the same order with gender-specific parameters δ_i, $\sigma^2_{i\Delta}$ and $\sigma^2_{i\bullet}$. This corresponds to assumption (S1), i.e. Model 2 in Section 7.3. The predictive distributions for Models 3 to 5 are then easily obtained by making the parameters gender-common and/or setting γ to 1. Model 1 is not of practical importance.

In applications, we are interested in the prediction of the future k years

$$\left(\Delta^{(m)}_{T+1},\dots,\Delta^{(m)}_{T+k},\Delta^{(f)}_{T+1},\dots,\Delta^{(f)}_{T+k}\right)$$

given the past observed aggregate mortality improvement factors $(r^{(m)}_{\bullet 1},\dots,r^{(m)}_{\bullet T},r^{(f)}_{\bullet 1},\dots,r^{(f)}_{\bullet T})$. Their credibility estimators follows from the recursive formula presented in Theorem 7.1. Since we have a full parametric model, we can even specify the predictive distribution. The main tool is the following famous result, which can be found, for example, in Giri (1977).

Lemma 8.1. *Let $N_i \sim \mathcal{N}(\mu_i,\Sigma_i)$, $i=1,2$, with $\mathrm{Cov}(Y_1,Y_2)=\Sigma_{12}$. Then, conditional on $Y_2=y_2$, Y_1 follows a Normal distribution with mean vector*

$$\mu_1+\Sigma_{12}\Sigma_2^{-1}(y_2-\mu_2)$$

and covariance matrix

$$\Sigma_1-\Sigma_{12}\Sigma_2^{-1}\Sigma_{12}'.$$

In order to apply Lemma 8.1, we require the distribution of the random vector

$$\left(r^{(m)}_{\bullet 1},\dots,r^{(m)}_{\bullet T},r^{(f)}_{\bullet 1},\dots,r^{(f)}_{\bullet T},\Delta^{(m)}_{T+1},\dots,\Delta^{(m)}_{T+k},\Delta^{(f)}_{T+1},\dots,\Delta^{(f)}_{T+k}\right) \tag{8.1}$$

gathering past aggregate mortality improvement factors and future time indices. The $T\times k$ correlation matrix $\boldsymbol{C}_{T,k}$ of the past $(\Delta^{(i)}_1,\dots,\Delta^{(i)}_T)$ and the future $(\Delta^{(i)}_{T+1},\dots,\Delta^{(i)}_{T+k})$ up to horizon $T+k$ is given by

$$\boldsymbol{C}_{T,k}=\begin{pmatrix}\rho_T & \rho_{T+1} & \cdots & \rho_{T+k-1}\\ \rho_{T-1} & \rho_T & \cdots & \rho_{T+k-2}\\ \vdots & \vdots & \cdots & \vdots\\ \rho_1 & \rho_2 & \cdots & \rho_k\end{pmatrix}.$$

The (t,l)-th entry in $\boldsymbol{C}_{T,k}$ is $\mathrm{corr}(\Delta^{(i)}_t,\Delta^{(i)}_{T+l})=\rho_{T+l-t}$ and can be calculated by equation (7.30).

Further, define $\boldsymbol{C}_{k,T} = \boldsymbol{C}'_{T,k}$. The random vector (8.1) is multivariate Normal with mean vector $\delta\mathbf{1}_{2T+2k}$ and covariance matrix

$$\begin{pmatrix} \sigma^2_{m\bullet} I_T + \sigma^2_{m\Delta}\boldsymbol{C}_T & \gamma\sigma^2_\Delta \boldsymbol{C}_T & \sigma^2_{m\Delta}\boldsymbol{C}_{T,k} & \gamma\sigma^2_\Delta \boldsymbol{C}_{T,k} \\ \gamma\sigma^2_\Delta \boldsymbol{C}_T & \sigma^2_{f\bullet} I_T + \sigma^2_{f\Delta}\boldsymbol{C}_T & \gamma\sigma^2_\Delta \boldsymbol{C}_{T,k} & \sigma^2_{f\Delta}\boldsymbol{C}_{T,k} \\ \sigma^2_{m\Delta}\boldsymbol{C}_{k,T} & \gamma\sigma^2_\Delta \boldsymbol{C}_{k,T} & \sigma^2_{m\Delta}\boldsymbol{C}_k & \gamma\sigma^2_\Delta \boldsymbol{C}_k \\ \gamma\sigma^2_\Delta \boldsymbol{C}_{k,T} & \sigma^2_{f\Delta}\boldsymbol{C}_{k,T} & \gamma\sigma^2_\Delta \boldsymbol{C}_k & \sigma^2_{f\Delta}\boldsymbol{C}_k \end{pmatrix}$$

and Lemma 8.1 can now be applied.

Theorem 8.2. *The predictive distribution for* $(\Delta^{(m)}_{T+1}, \dots, \Delta^{(m)}_{T+k}, \Delta^{(f)}_{T+1}, \dots, \Delta^{(f)}_{T+k})$ *given* $(r^{(m)}_{\bullet 1}, \dots, r^{(m)}_{\bullet T}, r^{(f)}_{\bullet 1}, \dots, r^{(f)}_{\bullet T})$ *is multivariate Normal with mean vector*

$$\begin{pmatrix} \delta_m \mathbf{1}_k \\ \delta_f \mathbf{1}_k \end{pmatrix} + \begin{pmatrix} \sigma^2_{m\Delta}\boldsymbol{C}_{k,T} & \gamma\sigma^2_\Delta \boldsymbol{C}_{k,T} \\ \gamma\sigma^2_\Delta \boldsymbol{C}_{k,T} & \sigma^2_{f\Delta}\boldsymbol{C}_{k,T} \end{pmatrix} \begin{pmatrix} \sigma^2_{m\bullet} I_T + \sigma^2_{m\Delta}\boldsymbol{C}_T & \gamma\sigma^2_\Delta \boldsymbol{C}_T \\ \gamma\sigma^2_\Delta \boldsymbol{C}_T & \sigma^2_{f\bullet} I_T + \sigma^2_{f\Delta}\boldsymbol{C}_T \end{pmatrix}^{-1} \begin{pmatrix} r^{(m)}_\bullet - \delta_m \mathbf{1}_T \\ r^{(f)}_\bullet - \delta_f \mathbf{1}_T \end{pmatrix} \tag{8.2}$$

and covariance matrix

$$\begin{pmatrix} \sigma^2_{m\Delta}\boldsymbol{C}_k & \gamma\sigma^2_\Delta \boldsymbol{C}_k \\ \gamma\sigma^2_\Delta \boldsymbol{C}_k & \sigma^2_{f\Delta}\boldsymbol{C}_k \end{pmatrix} - \begin{pmatrix} \sigma^2_{m\Delta}\boldsymbol{C}_{k,T} & \gamma\sigma^2_\Delta \boldsymbol{C}_{k,T} \\ \gamma\sigma^2_\Delta \boldsymbol{C}_{k,T} & \sigma^2_{f\Delta}\boldsymbol{C}_{k,T} \end{pmatrix} \begin{pmatrix} \sigma^2_{m\bullet} I_T + \sigma^2_{m\Delta}\boldsymbol{C}_T & \gamma\sigma^2_\Delta \boldsymbol{C}_T \\ \gamma\sigma^2_\Delta \boldsymbol{C}_T & \sigma^2_{f\bullet} I_T + \sigma^2_{f\Delta}\boldsymbol{C}_T \end{pmatrix}^{-1} \begin{pmatrix} \sigma^2_{m\Delta}\boldsymbol{C}_{T,k} & \gamma\sigma^2_\Delta \boldsymbol{C}_{T,k} \\ \gamma\sigma^2_\Delta \boldsymbol{C}_{T,k} & \sigma^2_{f\Delta}\boldsymbol{C}_{T,k} \end{pmatrix}. \tag{8.3}$$

Furthermore, (8.2) *is also the credibility estimator for* $(\Delta^{(m)}_{T+1}, \dots, \Delta^{(m)}_{T+k}, \Delta^{(f)}_{T+1}, \dots, \Delta^{(f)}_{T+k})$.

Proof. We only have to prove the second claim. Since (8.2) equals

$$\mathbb{E}\left[\left(\Delta^{(m)}_{T+1}, \dots, \Delta^{(m)}_{T+k}, \Delta^{(f)}_{T+1}, \dots, \Delta^{(f)}_{T+k} \right) \middle| \, r^{(m)}_{\bullet 1}, \dots, r^{(m)}_{\bullet T}, r^{(f)}_{\bullet 1}, \dots, r^{(f)}_{\bullet T} \right],$$

it is the Bayes estimator for the future Δ-variables. Furthermore, it is also an affine function of $\boldsymbol{r}_\bullet$, i.e. a particular element of the Hilbert space $L(1, \boldsymbol{r}_{\bullet 1}, \dots, \boldsymbol{r}_{\bullet T})$, so that (8.2) is exactly the credibility estimator. □

Given the predictive distribution for the future $\Delta^{(i)}_{T+k}$, one can obtain forecasts for the future death rates $m^{(i)}_{x,T+k}$ and the corresponding one-year death probabilities $q^{(i)}_{x,T+k}$. Precisely, by iterating the relationship

$$m^{(i)}_x(t) = m^{(i)}_x(t-1) \exp\left(\beta^{(i)}_x \Delta^{(i)}_t + \epsilon^{(i)}_{xt} \right),$$

we get

$$m^{(i)}_x(T+k) = m^{(i)}_x(T) \exp\left(\sum_{j=1}^{k} \left(\beta^{(i)}_x \Delta^{(i)}_{T+j} + \epsilon^{(i)}_{x,T+j} \right) \right). \tag{8.4}$$

We can use the credibility estimators to get point predictions of $m^{(i)}_x(T+k)$,

$$\widehat{m}^{(i)}_x(T+k) = m^{(i)}_x(T) \exp\left(\sum_{j=1}^{k} \widehat{\beta}^{(i)}_x \mathbb{E}\left[\Delta^{(i)}_{T+j} \mid r^{(m)}_{\bullet 1}, \dots, r^{(m)}_{\bullet T}, r^{(f)}_{\bullet 1}, \dots, r^{(f)}_{\bullet T} \right] \right). \tag{8.5}$$

Paths of future $m_x^{(i)}(T+k)$ can be simulated by (8.4). The corresponding one-year death probabilities $q_{x,T+k}^{(i)}$ and one-year survival probabilities $p_{x,T+k}^{(i)}$ are easily obtained from

$$q_{x,T+k}^{(i)} = 1 - p_{x,T+k}^{(i)} = 1 - \exp\big(-m_x^{(i)}(T+k)\big).$$

Any quantity of interest can then be computed from these life tables.

8.2. Period life expectancies

We illustrate the forecasts of our mortality model on the basis of period life expectancies. Using the predictive distribution (8.2)-(8.3) of the future $\Delta_{T+k}^{(i)}$ and the procedure described above, predictions for future mortality rates are derived. The predicted period life expectancy $\hat{e}_{65}(T+k)$ at age 65 in calendar year $T+k$ can then be calculated using the formula

$$\hat{e}_{65}(T+k) = \frac{1}{2} + \sum_{j\geq 1} \prod_{l=0}^{j-1} \hat{p}_{65+l,T+k}. \tag{8.6}$$

The predicted mortality improvements are applied on the last observation $m_x^{(i)}(2010)$. Table 8.1 shows point predictions $\hat{e}_{65}(2050)$ using formula (8.5) and also the point forecasts obtained by the Lee-Carter model are shown for comparison. We see that smoothing the estimated age effects $\beta_x^{(i)}$ has little impact on the life expectancy. Considering the Lee-Carter forecast, applying the mortality reduction factors to the last observations $m_x(2010)$ greatly affects the projected $e_{65}(T+k)$. In the remainder, all calculations are done with smoothed $\beta_x^{(i)}$ and Lee-Carter forecasts use $m_x^{(i)}(T)$ as an initial value instead of the offsets $\alpha_x^{(i)}$. In this case, the forecasts roughly agree.

Next, $e_{65}(2050)$ has been calculated for 3000 scenarios of simulated life tables for year 2050. To stress the role of the age-common processes $\Delta_t^{(i)}$ and their Lee-Carter counterparts $(\kappa_t^{(i)} - \kappa_{t-1}^{(i)})$, the noise terms in both models have been set to zero. The empirical standard deviations of the simulated $e_{65}(2050)$ are listed in Table 8.2. Although $\Delta_t^{(i)}$ have larger standard deviations than their Lee-Carter counterparts, the opposite is the case for the life expectancies. This might be counter-intuitive at first sight but it is a consequence of the underlying ARMA structure. As the estimated autocorrelation function of the time index is negative for lags of size one, large deviations of $\Delta_t^{(i)}$ are likely to be followed by $\Delta_{t+1}^{(i)}$ going into the opposite direction. By (8.4), the deviations cancel out. On the other hand, mortality improvements are independent under the Lee-Carter model. Thus, outliers remain and strongly impact the future life expectancy.

8.3. Robustness over successive forecasts

To conclude, let us show that the model proposed in the present thesis solves the robustness issue mentioned in the introduction, when applied sequentially over the years. To this end, we fit the model using data up to 2010 and update the predictive distribution by using data up to years 2011 and 2012. This provides three forecasts of future mortality that we compare together as well as to the Lee-Carter forecasts and the three official forecasts published by Statistics Belgium over the same period. As Table 8.3 shows,

	Mo.3, MA(1)	Mo.3, MA(1) smoothed β_x	Mo.5, ARMA(1,1)	Mo.5, ARMA(1,1) smoothed β_x
male	22.2386	22.2610	22.5540	22.5793
female	25.7778	25.7691	25.3981	25.4147
	Lee-Carter with $m_x(2010)$	Lee-Carter with α_x		
male	21.9329	19.5003		
female	25.3079	23.2666		

TABLE 8.1. Point forecasts of period life expectancy in 2050

	Mo.3, MA(1)	Mo.5, ARMA(1,1)	Lee-Carter
Male	0.9193	0.2848	0.7140
Female	0.6797	0.2425	0.8117

TABLE 8.2. Standard deviations of simulated life expectancies

	Mo.3, MA(1) male	Mo.5, ARMA(1,1) male	Lee-Carter male	Official male
up to 2012	22.29	22.63	21.86	23.04
up to 2011	22.38	22.64	22.12	23.21
up to 2010	22.26	22.57	21.93	22.91
	Mo.3, MA(1) female	Mo.5, ARMA(1,1) female	Lee-Carter female	Official female
up to 2012	25.71	25.41	25.16	25.12
up to 2011	25.84	25.43	25.41	25.29
up to 2010	25.79	25.40	25.30	25.00

TABLE 8.3. Predicted period life expectancies in year 2050 for different observation periods.

our estimates are more stable than the other two. This is again a consequence of the underlying ARMA structure, i.e. mortality improvements not being independent in time.

This effect is further illustrated in Figure 8.1. We have displayed there the forecasts for $e_{65}(T+1), \ldots, e_{65}(2015)$ with $T = 2010$, 2011 and 2012, starting from the latest available $e_{65}(T)$. It can be clearly seen that differences in the initial values are stabilized over time for our model, whereas forecasts by Lee-Carter are just straight lines starting from the different initial values.

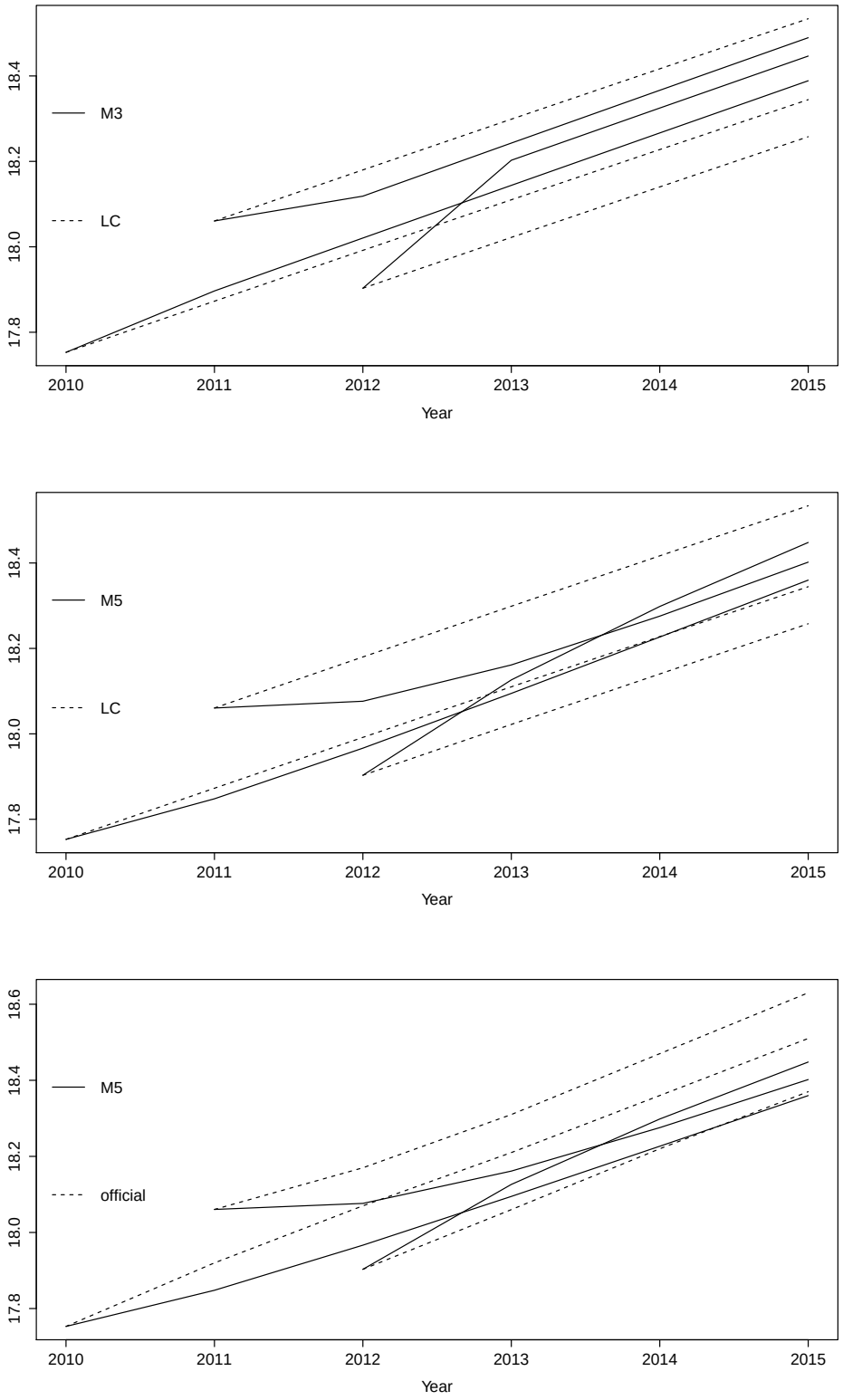

FIGURE 8.1. Comparison of predicted paths of period life expectancies. From top to bottom, we compare predictions of Model 3, MA(1), to Lee-Carter, Model 5, ARMA(1,1), to Lee-Carter and Model 5, ARMA(1,1), to Statistics Belgium. The three consecutive lines are based on data up to 2010, 2011 and 2012 respectively.

Bibliography

Aleksic, M.-C. and M. Börger (2011). Coherent projections of age, period, and cohort dependent mortality improvements. Technical report, Discussion Paper, University of Ulm.

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

Anderson, D. A. and M. Aitkin (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society. Series B (Methodological) 47*(2), pp. 203–210.

Appelt, M. (2014). Mehrpopulationen Sterblichkeitsschätzungen unter Verwendung von Credibility GLMs. Master's thesis, Universität Ulm.

Bichsel, F. (1964). Erfahrungstarifierung in der Motorfahrzeug-Haftpflichtversicherung. *Bulletin of Swiss Association of Actuaries*.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Brillinger, D. R. (1986). A biometrics invited paper with discussion: the natural variability of vital rates and associated statistics. *Biometrics*, 693–734.

Brockwell, P. J. and R. A. Davis (2006). *Time series: theory and methods*. Springer.

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.

Buhlmann, H. and E. Straub (1972). Credibility for loss ratios. *Actuarial Research Clearing House 2*.

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.

Carter, L. R. and R. D. Lee (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting 8*(3), 393–411.

Christiansen, M. and E. Schinzinger (2015). Credibility estimation for generalized linear models and application in mortality modeling. Preprint submitted.

Coelho, E. and L. C. Nunes (2011). Forecasting mortality in the event of a structural change. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*(3), 713–736.

Currie, I. D. (2013). Fitting models of mortality with generalized linear and non-linear models.

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

Czado, C., A. Delwarde, and M. Denuit (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics 36*(3), 260–284.

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

Debón, A., F. Montes, and F. Martínez-Ruiz (2011). Statistical methods to compare mortality for a group with non-divergent populations: an application to spanish regions. *European Actuarial Journal 1*(2), 291–308.

Delwarde, A., M. Denuit, M. Guillen, A. Vidiella, et al. (2006). Application of the Poisson log-bilinear projection model to the G5 mortality experience. *Belgian Actuarial Bulletin 6*(1), 54–68.

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.

Fahrmeir, L., G. Tutz, and W. Hennevogl (1994). *Multivariate statistical modelling based on generalized linear models*, Volume 2. Springer New York.

Fang, Y., K. A. Loparo, and X. Feng (1994). Inequalities for the trace of matrix product. *Automatic Control, IEEE Transactions on 39*(12), 2489–2490.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). Introducing Markov chain Monte Carlo. *Markov chain Monte Carlo in practice 1*, 19.

Giri, N. C. (1977). *Multivariate statistical inference*, Volume 65. Academic Press New York.

Girosi, F. and G. King (2008). *Demographic forecasting.* Princeton University Press.

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.

Hamilton, J. D. (1994). *Time series analysis*, Volume 2. Princeton university press Princeton.

Hatzopoulos, P. and S. Haberman (2009). A parameterized approach to modeling and forecasting mortality. *Insurance: Mathematics and Economics 44*(1), 103–123.

Hatzopoulos, P. and S. Haberman (2013). Common mortality modeling and coherent forecasts. An empirical analysis of worldwide mortality data. *Insurance: Mathematics and Economics 52*(2), 320–337.

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).

Hunt, A. and D. Blake (2014). A general procedure for constructing mortality models. *North American Actuarial Journal 18*(1), 116–138.

Jewell, W. S. (1973). Multi-dimensional credibility. Technical report, DTIC Document.

Jewell, W. S. (1974). Credible means are exact Bayesian for exponential families. *Astin Bulletin 8*(01), 77–90.

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

Kogure, A., K. Kitsukawa, and Y. Kurachi (2009). A Bayesian comparison of models for changing mortalities toward evaluating longevity risk in Japan. *Asia-Pacific Journal of Risk and Insurance 3*(2), 1–22.

Kogure, A. and Y. Kurachi (2010). A Bayesian approach to pricing longevity risk based on risk-neutral predictive distributions. *Insurance: Mathematics and Economics 46*(1), 162–172.

Lee, R. D. and L. R. Carter (1992). Modeling and forecasting us mortality. *Journal of the American statistical association 87*(419), 659–671.

Lee, Y. and J. Nelder (1996). Hierarchical generalized linear models. *J. R. Statistical Soc. B 58*(4), 619–678.

Li, J. (2013). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population studies 67*(1), 111–126.

Li, J. (2014). An application of MCMC simulation in mortality projection for populations with limited data. *Demographic Research 30*(1), 1–48.

Li, N. and R. Lee (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography 42*(3), 575–594.

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

Mitchell, D., P. Brockett, R. Mendoza-Arriaga, and K. Muthuraman (2013). Modeling and forecasting mortality rates. *Insurance: Mathematics and Economics 52*(2), 275–285.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), pp. 370–384.

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.

Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology 14*(1), 124–129.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50*(302), 157–175.

Pedroza, C. (2006). A Bayesian forecasting model: Predicting US male mortality. *Biostatistics 7*(4), 530–550.

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

Revuz, D. and M. Yor (1999). *Continuous martingales and Brownian motion*, Volume 293. Springer Science & Business Media.

Russolillo, M., G. Giordano, and S. Haberman (2011). Extending the Lee–Carter model: a three-way decomposition. *Scandinavian Actuarial Journal 2011*(2), 96–117.

Santner, T. J. and D. E. Duffy (1989). *The statistical analysis of discrete data*. Springer-Verlag New York.

Schinzinger, E., M. Denuit, and M. Christiansen (2014). An evolutionary credibility model for mortality improvement rates. Preprint submitted.

Seber, G. A. and A. J. Lee (2012). *Linear regression analysis*, Volume 936. John Wiley & Sons.

Sheldon, R. et al. (2002). *A first course in probability*. Pearson Education India.

Sundt, B. (1981). Recursive credibility estimation. *Scandinavian Actuarial Journal 1981*(1), 3–21.

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

Yang, S. S. and C.-W. Wang (2013). Pricing and securitization of multi-country longevity risk with mortality dependence. *Insurance: Mathematics and Economics 52*(2), 157–169.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American statistical association 86*(413), 79–86.

Zhou, R., J. S.-H. Li, and K. S. Tan (2013). Pricing standardized mortality securitizations: A two-population model with transitory jump effects. *Journal of Risk and Insurance 80*(3), 733–774.

Acknowledgments

This thesis appears in its current form due to the help of several people. First and foremost, I want to thank my PhD supervisor Assoc. Prof. Marcus Christiansen for his great support during the last years. He gave me the opportunity to write this thesis, drew my attention to actuarial mathematics and taught me how to develop and express my ideas. We had a lot of discussions and I benefitted substantially from his experience and valuable comments.
He also organized funding for my research after the expiration of the scholarship of the DFG Research Training Group 1100. In this connection, I also want to thank Prof. An Chen and Prof. Hajo Zwiesler for the financial support I have received from the Institute of Insurance Science. Of course, I acknowledge the help of all members of the institute.

I am very grateful to Prof. Ulrich Stadtmüller for accepting to be the second examiner of this thesis and his continuous support during my PhD studies. Prof. Denuit provided me the opportunity for a research visit in Louvain-la-Neuve. I want to thank him for the productive and enjoyable time in Belgium. With his expert knowledge in the field of mortality modeling he greatly contributed to this project.

I acknowledge the funding of the DFG Research Training Group 1100. Special thanks go to my friends and colleagues of the RTG for the wonderful and unforgettable time in Ulm. It was a great pleasure to be a part of this group.

Last but not least, I would like to thank my family for their constant and unconditional support.

Die Inhalte der Seiten wurden aus Gründen des Datenschutzes entfernt.

Erklärung

Hiermit versichere ich, Edo Schinzinger, dass ich die vorliegende Arbeit selbständig angefertigt habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe. Ich erkläre außerdem, dass diese Arbeit weder im In- noch im Ausland in dieser oder ähnlicher Form in einem anderen Promotionsverfahren vorgelegt wurde.

Ulm, den 19. Februar 2015

(Edo Schinzinger)