이학박사 학위논문

Statistical Inference for Random Unknowns

관측할 수 없는 변량 효과에 관한 통계적 추론

2023년 8월

서울대학교 대학원

통계학과

이항빈

Statistical Inference for Random Unknowns

관측할 수 없는 변량 효과에 관한 통계적 추론

지도교수 임 요 한

이 논문을 이학박사 학위논문으로 제출함

2023년 7월

서울대학교 대학원

통계학과

이 항 빈

이항빈의 이학박사 학위논문을 인준함

2023년 7월

위 원 장	정 성 규	(인)
부위원장	임 요 한	(인)
위 원	이 영 조	(인)
위 원	Myunghee Cho Paik	(인)
위 원	이 동 환	(인)

Statistical Inference for Random Unknowns

by

Hangbin Lee

A Thesis
submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
August, 2023

Abstract

Statistical Inference for Random Unknowns

Hangbin Lee

Department of Statistics

The Graduate School

Seoul National University

This thesis is composed of six topics related to statistical inference on unobserved random effects, each centered around the concept of extended likelihood that incorporates information about the random unknowns. The first two topics focus on the theoretical properties of confidence distribution, whose density can be interpreted as an extended likelihood. The latter four topics reformulate the hierarchical likelihood, as an extended likelihood at specific scale, and investigate its theoretical properties, as well as its applications to deep learning.

In the first topic, an epistemic confidence of the observed confidence intervals is introduced. Furthermore, the relevant subset problem associated is explained by incorporating the existence of betting markets into the Ramsey-De Finetti's Dutch book argument. It is demonstrated that the epistemic confidence is free from such issues. In the second topic, it is revealed that the existence of a point mass in confidence distribution plays an important role in maintaining the essential properties of the confidence distribution. The point mass has been considered paradoxical in Stein's paradox and satellite collision problems, but in fact, it gives an advantage to the confidence distribution. The proposed confidence distribution is free from the false confidence for the

parameter of interest, and it maintains the confidence feature in both Stein's problem and the satellite conjunction problem.

The third topic introduces the reformulation of hierarchical likelihood (h-likelihood) and establishes the theoretical properties for h-likelihood inference. This novel hierarchical likelihood can provide maximum likelihood estimators for fixed parameters and asymptotically the best unbiased estimators for random parameters, resolving the ambiguity of Lee and Nelder's (1996) original hierarchical likelihood. The last three topics deal with applications of the h-likelihood approach to deep learning models. While the most deep learning models implicitly assume the independence of data, real-world large scale data is often clustered with temporal-spatial correlations. In such cases, prediction performance of deep learning models can be improved by introducing random effects via h-likelihood. The fourth topic deals with deep learning models for continuous data with temporal-spatial correlations, and the fifth topic focuses on deep learning models for count data with non-Gaussian random effects. The sixth topic proposes h-likelihood approach to semi-parametric deep neural networks with gamma frailty for analyzing clustered censored data. In all three topics, the proposed methods improve prediction performance by introducing random effects into the existing deep learning models.

Keywords: random effects, hierarchical likelihood, deep learning, confidence, epistemic confidence, confidence distribution, repeated measures, spatiotemporal data, survival analysis, frailty model

Student Number: 2015-20310

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we investigate six topics on the statistical inference for random unknowns. The first two topics focus on the theoretical properties of confidence distribution, whose density can be interpreted as an extended likelihood. The third topic reformulates the hierarchical likelihood (h-likelihood), which is an extended likelihood at specific scale, and establish its theoretical properties. The last three topics deal with applications of the h-likelihood approach to deep learning models.

In Chapter 2, we define confidence to be epistemic if it applies to an observed confidence interval. Epistemic confidence is unavailable – or even denied – in orthodox frequentist inference, as the confidence level is understood to apply to the procedure. Yet there are obvious practical and psychological needs to think about the uncertainty in the observed interval. We extend the Dutch Book argument used in the classical Bayesian justification of subjective probability to a stronger market-based version, which prevents external agents from exploiting unused information in any relevant subset. We previously showed

that confidence is an extended likelihood, and the likelihood principle states that the likelihood contains all the information in the data, hence leaving no relevant subset. Intuitively, this implies that confidence associated with the full likelihood is protected from the Dutch Book, and hence is epistemic. Our goal is to validate this intuitive notion through theoretical backing and practical illustrations.

In Chapter 3, we focus on the existence of a point mass in confidence distribution (CD). To make probabilistic inference on multi-parameter cases, several methods have been proposed, such as generalized fiducial distribution (GFD) and the reference posterior (RP). However, Stein's (1959) problem highlights a fundamental deficiency in probabilistic inference, called the probability dilution, in high-dimensional cases. Furthermore, in the context of satellite conjunction problem, poor-quality data with increasing variance results in a dilution of collision probability. Additionally, we highlight the ambiguity of coverage probability in an observed frequentist confidence interval (CI), and the CD provides valuable information on the ambiguous coverage probability of an observed CI. However, due to the presence of a point mass, the CD can maintain the confidence feature and avoid probability dilution in both Stein's problem and satellite conjunction problem. A point mass in CD has been considered paradoxical, but in fact, it gives an advantage rather than a drawback.

In recent research, another deficiency in probabilistic inference was reported, called the false confidence, which can be mitigated by adopting the consonant belief (CB). It was claimed that additional consonant feature is crucial for overcoming the deficiencies in probabilistic inferences. However, we

demonstrate that the CD can be free from the false confidence for the proposition of interest. We further introduces the null belief theorem, which implies a fundamental deficiency of CB in statistical inferences. Our findings demonstrate that the CD outperforms the GFD, RP, and CBs. Therefore, it is a confidence feature, not a consonant one, that successfully avoids deficiencies of probabilistic inference and overcome difficulties in Stein's problem and the satellite conjunction problem.

In Chapter 4, we reformulate the hierarchical likelihood (h-likelihood) and establish its theoretical properties. Maximum likelihood procedure is widely use for statistical inferences. The maximum h-likelihood procedure gives maximum likelihood estimators for fixed parameters and asymptotically best unbiased predictors for random parameters. We provide theoretical foundations for the h-likelihood inference, which extends classical likelihood theories to broad classes of statistical models with random parameters. We introduce an generalized Cramer-Rao lower bound, which can be applied to both fixed and random parameters, and show that the maximum h-likelihood estimators asymptotically achieve the lower bound. We also study asymptotic theory when the consistency of either the fixed parameter estimation or the random parameter prediction is not guaranteed.

In Chapter 5, we propose the h-likelihood approach for DNN with temporal-spatial random effects. DNN is one of the most powerful tools for prediction, but many of them implicitly assume that the data are statistically independent. However, in the real world, it is common for large-scale data to be clustered with temporal-spatial correlation structures. Variational approaches and integrated likelihood approaches have been proposed to obtain approximate max-

imum likelihood estimators (MLEs) for correlated data. However, due to the large size of data, they cannot provide exact MLEs. In this study, we propose a new hierarchical likelihood approach to DNNs with correlated random effects for clustered data. By jointly optimizing the the negative h-likelihood loss, we can provide exact MLEs for both mean and dispersion parameters, as well as the best linear unbiased predictors for the random effects. Moreover, the hierarchical likelihood allows a computable procedure for restricted maximum likelihood estimators of dispersion parameters. The proposed two-step algorithm enables online learning for the neural networks, whereas the integrated likelihood cannot decompose like a widely-used loss function in DNNs. The proposed h-likelihood approach offers several advantages, which we demonstrate through numerical studies and real data analyses.

In Chapter 6, we introduce gamma random effects into the Poisson DNN for clustered count data via h-likelihood. Poisson deep neural networks (DNNs) have been developed to predict count data. To improve the predictions for clustered data, there has been a growing interest in subject-specific predictions of DNNs. In this chapter, we propose a new hierarchical likelihood approach for introducing gamma random effects into the Poisson DNNs. The h-likelihood approach simultaneously yields maximum likelihood estimators for fixed parameters and best unbiased predictors for random effects by optimizing a single objective function, enabling a fast end-to-end Poisson-gamma DNN. It enhances prediction performance by capturing both nonlinear effects of input variables and subject-specific cluster effects. We have observed that the local minima problem can lead to poor predictions when the Poisson DNN reflects subject-specific cluster effects. To address this issue, we propose an

adjustment to the random effects to enhance the prediction performance. In addition, we introduce a method-of-moments-based estimator to pretrain the dispersion parameter. Experimental studies and real data analyses confirm that the Poisson-gamma DNN improves the prediction performance of the existing methods. In particular, real data analyses illustrate that incorporating random subject-specific cluster effects helps to identify the nonlinear effects of the input variables, which cannot be found by the Poisson DNN.

In Chapter 7, we propose the DNN-based semi-parametric frailty model for prediction of clustered time-to-event data. An advantage of the proposed model is that the joint maximization of the h-likelihood provides maximum likelihood estimators for fixed parameters and best unbiased predictors for random frailties. Thus, the proposed DNN-FM is trained by using a negative profiled h-likelihood as a loss function, constructed by profiling out the non-parametric baseline hazard. Experimental studies show that the proposed method enhances the prediction performance of the existing methods. A real data analysis shows that the inclusion of subject-specific frailties helps to improve prediction of the DNN-based Cox proportional hazard model.

Chapter 2

Epistemic confidence in the observed confidence interval

2.1 Introduction

Given data $Y = y$ of arbitrary size or complexity, generated from a model $p_\theta(y)$ indexed with the scalar parameter of interest θ, a confidence interval CI(y) is computed with coverage probability

$$P_\theta(\theta \in \text{CI(Y)}) = \gamma.$$

We are interested in the *epistemic confidence*, defined as the sense of confidence *in the observed CI*(y)**.** For simplicity, we shall often drop the explicit dependence on y from the CI. Arguably, this is what we want from a CI, but the orthodox frequentist view is emphatic that the probability γ applies not to the observed interval CI(y), but to the procedure. In the confidence

interval theory, the coverage probability is called the confidence level. So, in the frequentist theory, 'confidence' has no separate meaning from probability as well as no epistemic property. Strictly speaking, we do not automatically have 95% confidence in the observed 95% CI. Schweder and Hjort (2016) and Schweder (2018) have been strong proponents of interpreting confidence as 'epistemic probability.' However, their view is not commonly accepted. Traditionally, only the Bayesians have no problem in stating that their subjective probability is epistemic. How do they achieve that? Is there a way to make the non-Bayesian confidence epistemic? Our aim is to show a way to achieve that.

Frequentists interpret probability as either a long-term frequency or a propensity of the generating mechanism. So, for them, unique events, such as the next toss of a coin or the true status of an observed CI, do not have a probability. On the other hand, Bayesians can attach their subjective probability to such unique events. But what does 'attach' mean? One standard interpretation is made based on a logical device called the Dutch Book. As classically proposed by Ramsey (1926) and De Finetti (1931), your subjective probability of an event E is defined as the personal betting price that you put on the event. Though subjective, the price is not arbitrary, but it follows a normative rational consideration; it is a price that is protected from the Dutch Book, i.e., no external agent can make a risk-free profit off you. Let's call your prices for a collection of bets a betting strategy. Then it is irrational to use a betting strategy that is guaranteed to lose.

Thus we conceptually define confidence to be epistemic if it is protected from the Dutch Book, but crucially we assume that there is a betting market of

a crowd of independent and intelligent players. In this market, bets are like a commodity with supply and demand from among the players. Assuming a perfect market condition – for instance, full competition, perfect information and no transaction cost – in accordance with the Arrow-Debreu theorem (Arrow and Debreu, 1954), there is an equilibrium price at which there is balance between supply and demand. 'Perfect information' means all players have access to the generic data y and the statistical model $p_\theta(y)$. For the betting market in particular, the fundamental theorem of asset pricing (Ross et al., 1976) states that, assuming a statistical model, the Dutch Book cannot be made if the price is determined by the objective probability.

It is worth emphasizing the difference between our setup and the classical Dutch Book argument used to establish the subjective Bayesian probability. In the latter, because it does not presume the betting market, bets are made only between two persons, you and me. To avoid the Dutch Book, you have to make your bets internally consistent by following probability laws. However, even if your bets are internally consistent, if your prices do not match the market prices, I can make a risk-free profit by playing between you and the market; see Example 2.1. So, the presence of the market imposes a stronger requirement for epistemic probability. We shall avoid the terms 'subjective' and 'objective'; one might consider 'epistemic' to be subjective since it refers to a personal decision-making based on a unique event, but the market consideration makes it impersonal.

Our question is when or under what condition the confidence, as measured by the coverage probability, apply to the observed interval. One way to judge this is whether you are willing to bet on the true status of the CI using the

confidence level as your personal price. Normatively, this should be the case if you know there is no better price. Intuitively, this is when you're sure that you have used all the available information in the data, so nobody can exploit you, i.e., construct a Dutch Book against you. Theoretically, to construct the Dutch Book, an external agent must exploit unused information in the form of a relevant subset, conditional on which he can get a different coverage probability.

Pawitan and Lee (2021) showed that the confidence is an extended likelihood (Lee et al., 2017). The extended likelihood principle (Bjørnstad, 1996) states that the extended likelihood contains all the information in the data. Intuitively, this implies that the extended likelihood leaves no relevant subset, and is thus protected from the Dutch Book. In other words, we can attach the degree of confidence to the observed CI, i.e., confidence is epistemic, provided it is associated with the full likelihood. Our aim is to establish the theoretical justification for this intuitive notion and to provide clear illustrative examples.

To summarize briefly and highlight the plan of the chapter, we describe three key concepts: relevant subset, confidence and ancillary statistic. We prove the main theorem that there are no relevant subsets if confidence is associated with the full likelihood. This condition is easily satisfied if the confidence is based on a sufficient statistic. When there is no sufficient statistic, but there exists a maximal ancillary statistic, then this ancillary defines relevant subsets; the confidence is conditional on the ancillary, but there are no further relevant subsets.

2.2 Main theory

2.2.1 Relevant subsets

Intuitively, we could use the coverage probability γ as a betting price if there is no better price given the data at hand. So the question is, are there any features of the data that can be used to improve the price? If they exist, such features are said to be relevant. Formally, a statistic $R(y)$ is defined to be relevant (cf. Buehler, 1959) if the conditional coverage probability given $R(y)$ is non-trivially biased in one direction. That is, for a positive bias, there is $\epsilon > 0$ free of θ and some y, such that

$$P_\theta(\theta \in \mathrm{CI}(Y)|R(y)) \geq \gamma + \epsilon \quad \text{for all } \theta. \tag{2.1}$$

If it exists, the feature $R(y)$ can be used to construct a Dutch Book: Suppose you and I are betting, and I notice that the event $R(y)$ occurs. If you set the price at γ, then I would buy the bet from you and then sell it in the betting market at $\gamma + \epsilon$ to make a risk-free profit of ϵ. Similarly, for the negative bias, the relevant $R(y)$ has the property

$$P_\theta(\theta \in \mathrm{CI}(Y)|R(y)) \leq \gamma - \epsilon \text{ for all } \theta. \tag{2.2}$$

Technically, $R(y)$ induces subsets of the sample space, known as the 'relevant subsets,' so the terms 'relevant statistic' and 'relevant subset' are interchangeable. If there is a relevant subset, the confidence level γ is not epistemic. Conversely, if there are no relevant subsets, the betting price determined by the confidence level is protected from the Dutch Book. So, mathematically,

we establish epistemic confidence by showing that it corresponds to a coverage probability that is free of relevant subsets.

Example 2.1. Let $y \equiv (y_1, y_2)$ be an iid sample from a uniform distribution on $\{\theta-1, \theta, \theta+1\}$, where the parameter θ is an integer. Let $y_{(1)}$ and $y_{(2)}$ be the minimum and maximum values of y_1 and y_2. We can show that the confidence interval $\text{CI}(y) \equiv [y_{(1)}, y_{(2)}]$ has a coverage probability

$$P_\theta(\theta \in \text{CI}) = 7/9 = 0.78.$$

For example, on observing $y_{(1)} = 3$ and $y_{(2)} = 5$, the interval $[3, 5]$ is formally a 78% CI for θ. But, if we ponder a bit, in this case we can actually be sure that the true $\theta = 4$. So, the probability of 7/9 is clearly a wrong price for this interval. This is a typical example justifying the frequentist objection to attaching the coverage probability as a sense of confidence in an observed CI.

Here the range $R \equiv R(y) \equiv y_{(2)} - y_{(1)}$ is relevant. If $R = 2$ we know for sure that θ is equal to the midpoint of the interval, so the CI will always be correct. But if $R = 0$, the CI is equal to the point y_1, and it falls with equal probability at the integers $\{\theta - 1, \theta, \theta + 1\}$. So, for all θ,

$$\begin{aligned} P_\theta(\theta \in \text{CI} | R = 2) &= 1 > 7/9 \\ P_\theta(\theta \in \text{CI} | R = 1) &= 1 > 7/9 \\ P_\theta(\theta \in \text{CI} | R = 0) &= 1/3 < 7/9. \end{aligned}$$

In the betting market, the range information will be used by the intelligent players to settle prices at these conditional probabilities. For example, if $y_1 = 3$

and $y_2 = 5$, the intelligent players will not use 7/9 as the price and will instead use 1.00. So, the information can be used to construct a Dutch Book against anyone who ignores R. How do we know that there is a relevant subset in this case? Moreover, given R, how do we know if there is no further relevant subset?

To contrast with the classical Ramsey-de Finetti Dutch Book argument, suppose $y_1 = y_2 = 3$. If, for whatever subjective reasons, you set the price 7/9 for $[\theta \in \text{CI}]$, you are being internally consistent as long as you set the price 2/9 for $[\theta \notin \text{CI}]$, since the two numbers constitute a valid probability measure. Internal consistency means that I cannot make a risk-free profit from you based on this single realization of y. Even if I know that 1/3 is a better price, I cannot take any advantage of you because there is no betting market. So 7/9 is a valid subjective probability. □

2.2.2 Confidence distribution

Let $t \equiv T(y)$ be a statistic for θ, and define the right-side P-value function

$$C_m(\theta; t) \equiv P_\theta(T \geq t). \tag{2.3}$$

Assuming that, for each t, it behaves formally like a proper cumulative distribution function, $C_m(\theta; t)$ is called the confidence distribution of θ. The subscript m is used to indicate that it is a 'marginal' confidence, as it depends on the marginal distribution of T. For continuous T, at the true parameter, the random variable $C_m(\theta; T)$ is standard uniform. For continuous θ, the

corresponding confidence density is

$$c_m(\theta) \equiv c_m(\theta; t) \equiv \partial C_m(\theta; t)/\partial\theta. \tag{2.4}$$

The functions $C_m(\theta; t)$ and $c_m(\theta)$ across θ are realized statistics, which depend on both the data and the model, but not on the true unknown parameter θ_0. We can view the confidence distribution simply as the collection of P-values or CIs. We define

$$C_m(\theta \in \text{CI}) \equiv \int_{\text{CI}} c_m(\theta) d\theta \tag{2.5}$$

to convey the 'confidence of θ belonging in the CI'. Fisher (1930, 1933) called $C_m(\theta; t)$ the fiducial distribution of θ, but he required T to be sufficient. However, the recent definition of the confidence distribution (Schweder and Hjort, 2016) requires only $C_m(\theta; T)$ to be uniform at the true parameter, thus guaranteeing a correct coverage probability. Lemma 2.1 below establishes when Fisher's fiducial probability $C_m(\theta; t)$ becomes a frequentist coverage probability, which requires T to be continuous. When T is discrete, the equality is only achieved asymptotically; see Appendix 2.5.2 for an example.

Assume Condition 2.1 in Section 2.2.3 that for any $\alpha \in (0, 1)$, the quantile function $q_\alpha(\theta)$ of T is a strictly increasing function of θ. Then the frequentist procedure based on T gives a γ-level CI defined by

$$\text{CI}_\gamma(T) = \left(q_{\gamma_2}^{-1}(T), q_{\gamma_1}^{-1}(T)\right) \tag{2.6}$$

for some $\gamma_2 > \gamma_1 > 0$ with $\gamma_2 - \gamma_1 = \gamma$, to have a coverage probability

$$P_\theta(\theta \in \mathrm{CI}_\gamma(T)) = P_\theta\Big[T \in (q_{\gamma_1}(\theta), q_{\gamma_2}(\theta))\Big] = \gamma_2 - \gamma_1 = \gamma.$$

Here the coverage probability is a frequentist probability based on the distribution of unobserved future data T, whereas given observed data t, the confidence is for the observed interval $\mathrm{CI}(t)$ based on the confidence density of θ. The confidence becomes

$$\begin{aligned} C_m(\theta \in \mathrm{CI}_\gamma(t); t) &= C_m(\theta = q_{\gamma_1}^{-1}(t); t) - C_m(\theta = q_{\gamma_2}^{-1}(t); t) \\ &= P_{\theta = q_{\gamma_1}^{-1}(t)}(T \geq t) - P_{\theta = q_{\gamma_2}^{-1}(t)}(T \geq t) \\ &= (1 - \gamma_1) - (1 - \gamma_2) = \gamma = P_\theta(\theta \in \mathrm{CI}_\gamma(T)). \end{aligned}$$

Thus, we have the following lemma.

Lemma 2.1. *Under Condition 2.1,*

$$P_\theta(\theta \in CI(T)) = C_m(\theta \in CI(t); t). \tag{2.7}$$

where $CI(t)$ is the observed interval of confidence procedure $CI(T)$ defined in (2.6).

However, as shown in Example 2.1, a correct coverage probability does not rule out relevant subsets. This means that the current definition of confidence distribution does not guarantee epistemic confidence. The key step is to define a confidence distribution that uses the full information. Motivated by the

Bayesian formulation and Efron (1993), let's define the implied prior as

$$c_0(\theta) \equiv c_0(\theta; t) \equiv m(t)\frac{c_m(\theta; t)}{L(\theta; t)}, \tag{2.8}$$

where $m(t)$ cancels out all the terms not involving θ in $c_m(\theta; t)/L(\theta; t)$. Then define the *full confidence density* as

$$c_f(\theta) \equiv c_f(\theta; y) \propto c_0(\theta)L(\theta; y). \tag{2.9}$$

The subscript f is now used to indicate that the confidence density is associated with the full likelihood based on the whole data. When necessary for clarity, the dependence of the confidence density and the likelihood on t and on the whole data y will be made explicit. $c_f(\theta)$ is defined only up to a constant term to allow it to integrate to one. Obviously, if T is sufficient, then $c_m(\theta) = c_f(\theta)$, but in general they are not equal. In Section 2.3, we show a more convenient way to construct $c_f(\theta)$. The confidence function parallel to (2.5) can be denoted by $C_f(\cdot)$. Thus, the full confidence density looks like a Bayesian posterior. However, the implied prior is not subjectively selected, and can be improper or data-dependent.

2.2.3 Main theorem

The full confidence density $c_f(\theta)$ can be used in general to compute the degree of confidence γ to any observed CI(y) as

$$\gamma = \int_{CI(y)} c_f(\theta)d\theta.$$

The CI has a coverage probability, which may or may not be equal to γ. We say that $c_f(\theta)$ has no relevant subsets, if there is no $R(y)$ such that the conditional coverage probability is biased in one direction according to (2.1) or (2.2). For our main theorem, we assume the following regularity conditions.

Condition 2.1. $T = T(Y)$ is a continuous scalar statistic whose quantile function $q_\alpha(\theta)$, defined by $P_\theta(T \leq q_\alpha(\theta)) = \alpha$, is strictly increasing function of θ for any $\alpha \in (0, 1)$.

Condition 2.2. There exists a function $g(\theta) > 0$ free of y, such that for any given $Y = y$,

$$E_{\theta|y}\left(\frac{g(\theta)}{c_0(\theta; y)}\Big|\theta \in \mathrm{CI}(y)\right) \leq E_{\theta|y}\left(\frac{g(\theta)}{c_0(\theta; y)}\right) < \infty \tag{2.10}$$

where $E_{\theta|y}(\cdot)$ is the expectation under confidence density $c(\theta; y)$, and $c_0(\theta; y)$ is the implied prior.

If the implied prior does not depend on the data, $c_0(\theta; y) = c_0(\theta)$, then the choice $g(\theta) = c_0(\theta)$ leads both sides of (2.10) to be 1. Thus, Condition 2.2 holds for any data-free implied prior even if it is improper. If the implied prior is data-dependent, $g(\theta)$ would be a function which diverges near the boundary of Θ. We shall illustrate this in Example 2.3 in Section 2.3.3. For the general single-parameter exponential family,

$$p_\theta(y) = h(y) \exp(\theta T(y) - A(\theta)) \tag{2.11}$$

where the parameter space Θ and the sample space of T are identical, the

choice $g(\theta) = h(\theta)e^{A(\theta)}$ leads to a sufficient condition for Condition 2.2:

$$P_{\theta=t}(T \in \mathrm{CI}(t)) \leq P_{\theta=t}(\theta \in \mathrm{CI}(T)) = \gamma.$$

The quantity $1 - P_{\theta=t}(T \in \mathrm{CI}(t))$ is the significance level for testing the null hypothesis $\theta = t$ with acceptance region $\mathrm{CI}(t)$. $P_{\theta=t}(\theta \in \mathrm{CI}(T))$ is the frequentist coverage probability. This inequality states the usual relationship between the hypothesis testing and the confidence interval.

Theorem 2.1. *Consider the full confidence density $c_f(\theta) \propto c_0(\theta)L(\theta; y)$ with $c_0(\theta)$ being the implied prior (2.8). Let γ be the degree of confidence for the observed CI(y) such that*

$$\int_{CI(y)} c_f(\theta)d\theta = \gamma \quad \textit{for all } y.$$

Under Conditions 2.1 and 2.2, $c_f(\theta)$ has no relevant subsets.

Proof: We first prove the positively biased case, which presumes that there exists a positively-biased relevant subset R. Equation (2.1) can be expressed as

$$\int_R I(\theta \in \mathrm{CI}_\gamma(y))f_\theta(y)dy \geq (\gamma + \epsilon)\int_R f_\theta(y)dy \quad \text{for any } \theta \in \Theta.$$

Consider a function $g(\theta) > 0$ from Condition 2.2, then we have

$$A = \int_\Theta \int_R I(\theta \in \mathrm{CI}_\gamma(y))f_\theta(y)dyg(\theta)d\theta \geq (\gamma + \epsilon)\int_\Theta \int_R f_\theta(y)dyg(\theta)d\theta = B.$$

On both sides the integrands are non-negative, so the order of integration can

be interchanged. Then the left-hand-side becomes

$$\begin{aligned}
A &= \int_R \int_\Theta I(\theta \in \mathrm{CI}_\gamma(y)) f_\theta(y) g(\theta) d\theta dy = \int_R \left[\int_{CI_\gamma(y)} g(\theta) f_\theta(y) d\theta \right] dy \\
&= \int_R \left[\int_{CI_\gamma(y)} c(\theta; y) \frac{g(\theta)}{c(\theta; y)/f_\theta(y)} d\theta \right] dy \\
&= \int_R m(y) \left[E_{\theta|y} \left(\frac{g(\theta)}{c_0(\theta; y)} \Big| \theta \in \mathrm{CI}_\gamma(y) \right) \int_{\mathrm{CI}_\gamma(y)} c(\theta; y) d\theta \right] dy \\
&= \gamma \int_R m(y) \left[E_{\theta|y} \left(\frac{g(\theta)}{c_0(\theta; y)} \Big| \theta \in \mathrm{CI}_\gamma(y) \right) \right] dy,
\end{aligned}$$

while the right-hand-side becomes

$$\begin{aligned}
B &= (\gamma + \epsilon) \int_R \int_\Theta f_\theta(y) g(\theta) d\theta dy \\
&= (\gamma + \epsilon) \int_R m(y) \left[E_{\theta|y} \left(\frac{g(\theta)}{c_0(\theta; y)} \right) \right] dy.
\end{aligned}$$

Since $\gamma + \epsilon > \gamma$, $m(y) > 0$, and $E_{\theta|y} \left(\frac{g(\theta)}{c_0(\theta;y)} \right) \geq E_{\theta|y} \left(\frac{g(\theta)}{c_0(\theta;y)} \Big| \theta \in \mathrm{CI}_\gamma(y) \right)$, we get $A < B$, which is a contradiction. Hence there is no positively-biased relevant subset.

Now suppose that there exists a negatively biased relevant subset R^*. Let $R = (R^*)^C$ be the complementary set of R^*, then

$$\begin{aligned}
\gamma &= P_\theta(\theta \in \mathrm{CI}_\gamma(T(Y))) \\
&= P_\theta(\theta \in \mathrm{CI}_\gamma(T(Y)) | Y \in R) P_\theta(Y \in R) \\
&\quad + P_\theta(\theta \in \mathrm{CI}_\gamma(T(Y)) | Y \in R^*)(1 - P_\theta(Y \in R)) \\
&< P_\theta(\theta \in \mathrm{CI}_\gamma(T(Y)) | Y \in R) P_\theta(Y \in R) + (\gamma - \epsilon)(1 - P_\theta(Y \in R)),
\end{aligned}$$

which leads to

$$P_\theta(\theta \in \mathrm{CI}_\gamma(T(Y))|Y \in R) > \gamma + \epsilon(1 - P_\theta(Y \in R))/P_\theta(Y \in R).$$

Hence R becomes a positively-biased relevant subset, which is shown above to lead to a contradiction. Therefore, overall there is no relevant subset. □

Note that we now have two ways of computing the price of an observed CI: using $C_f(\theta \in \mathrm{CI})$ or using $P_\theta(\theta \in \mathrm{CI})$. The latter has the desired coverage probability, but not guaranteed to be free of relevant subsets; the former is free of relevant subset, but not guaranteed to match the coverage probability. If the two are equal, we have a confidence that corresponds to a coverage probability free of relevant subsets, hence epistemic. If T is sufficient and satisfies Condition 2.1, Lemma 2.1 implies that the frequentist CI satisfies

$$P_\theta(\theta \in \mathrm{CI}(Y)) = C_m(\theta \in \mathrm{CI}(y)) = C_f(\theta \in \mathrm{CI}(y)) = \gamma \text{ for all } \theta \text{ and } y$$

Thus, we can summarize the first key result in the following corollary:

Corollary 2.1. *Under Conditions 2.1 and 2.2, if T is sufficient, the confidence based on $c_m(\theta; t)$ has a correct coverage probability and no relevant subsets, hence it is epistemic.*

We note that $P_\theta(\theta \in \mathrm{CI}(Y)) = C_f(\theta \in \mathrm{CI}(y))$ holds asymptotically, regardless whether y is continuous or discrete. Corollary 1 specifies the conditions where it is true in finite samples.

If $c_0(\theta)$ is a proper probability density that does not depend on y, then $c_f(\theta)$ is a Bayesian posterior density, shown already by Robinson's (1979) Proposi-

tion 7.4 not to have relevant subsets. For proper priors, Condition 2.2 trivially holds, so the theorem extends his result to improper and data-dependent priors. Moreover, there is a significant difference in the interpretation. If you use a proper but arbitrary $c_0(\theta)$ that is not the same as the implied prior, and there is a betting market, your price γ will differ from the market price. So, as illustrated in Example 2.1, I can construct a Dutch Book against you. While, assuming no betting market, the theorem is meaningful only for two people betting repeatedly against each other, with gains or losses expressed in terms of expected value or long-term average. This is the setting described by Buehler (1959) and Robinson (1979). Crucially, the presence of relevant subsets does not guarantee an external agent a risk-free profit from a single bet. So, it does not satisfy our original definition of epistemic confidence.

Lindley (1958) showed that, assuming T is sufficient, Fisher's fiducial probability – hence the marginal confidence – is equal to the Bayesian posterior if and only if the family $p_\theta(y)$ is transformable to a location family. However, his proof assumed $c_0(\theta)$ to be free of y. Condition 2.2 of the main theorem allows $c_0(\theta)$ to depend on the data, so our result is not limited to the location family.

2.2.4 Ancillary statistics

The current definition of confidence distribution (Schweder and Hjort, 2016) only requires $C_m(\theta; T)$ to follow uniform distribution. However, if T is not sufficient, the marginal confidence is not epistemic, because it does not use the full likelihood, so it is not guaranteed free of relevant subsets. Limiting ourselves to models with sufficient statistics to get epistemic confidence is

overly restrictive, since sufficient statistics exist at arbitrary sample sizes in the full exponential family only (Pawitan, 2001). Using non-sufficient statistics implies a potential loss of efficiency and epistemic property. Further progress depends on the ancillary statistic, a feature of the data whose distribution is free of the unknown parameter (Ghosh et al., 2010). We first have a parallel development for the conditional confidence distribution given the ancillary $A(y) = a$:

$$C_c(\theta; t|a) \equiv P_\theta(T \geq t|a) \quad \text{and} \quad c_c(\theta; t|a) \equiv \partial C_c(\theta; t|a)/\partial\theta.$$

We have immediately the following corollary from Lemma 2.1. Condition 2.1 needs a little modification, where it refers to the conditional statistic $T|a$ for each a.

Corollary 2.2. *Under Condition 2.1,*

$$P_\theta(\theta \in CI|a) = C_c(\theta \in CI; t|a). \tag{2.12}$$

where CI is the confidence interval based on the conditional distribution of $T|a$.

Furthermore, define the implied prior as

$$c_0(\theta) \equiv c_0(\theta; t|a) \equiv m(t, a)\frac{c_c(\theta; t|a)}{L(\theta; t|a)}, \tag{2.13}$$

where $m(t, a)$ cancels out all the terms not involving θ in $c_c(\theta; t|a)/L(\theta; t|a)$. As before, the full confidence is $c_f(\theta) \propto c_0(\theta)L(\theta; y)$.

Suppose $T(y) = t$ is not sufficient but (t, a) is, where a is an ancillary statistic. In this case, a is called an ancillary complement, and in a qualitative

sense it is a maximal ancillary, because

$$L(\theta; y) = L(\theta; t, a) \propto p_\theta(t|a)p(a) \propto p_\theta(t|a) = L(\theta; t|a). \tag{2.14}$$

Thus, conditioning a non-sufficient statistic by a maximal ancillary has recovered the lost information and restored the full-data likelihood. In particular, the conditional confidence becomes the full confidence: $c_c(\theta; t|a) = c_f(\theta)$. Note that (2.14) holds for any maximal ancillary, so if a maximal ancillary exists, then the full likelihood is automatically equal to the conditional likelihood given any maximal ancillary statistic. In its sampling theory form, when t is the MLE $\widehat{\theta}$, full information can be recovered from $p_\theta(\widehat{\theta}|a)$, whose approximation was studied by Barndorff-Nielsen (1983).

In conditional inference (Reid, 1995), we condition on the ancillary to make our inference more 'relevant' to the data at hand; in other words, more epistemic. But this is typically stated on an intuitive basis; the following corollary provides a mathematical justification. Since we already condition on $A(y)$, a further relevant subset $R(y)$ is such that the conditional probability $P_\theta(\theta \in \text{CI}|A(y), R(y))$ is non-trivially biased in one direction from $P_\theta(\theta \in \text{CI}|A(y))$ in the same manner as (2.1). Now we can state our second key result:

Corollary 2.3. *If $A(y) = a$ is maximal ancillary for $T(y)$, and CI is constructed from the conditional confidence density based on $T|a$, then under Conditions 2.1 and 2.2, the conditional confidence $C_c(\theta \in CI; t|a)$ has a correct coverage probability and no further relevant subsets. Hence the conditional confidence is epistemic.*

Because of (2.14), the confidence is epistemic for any choice of the maximal ancillary. However, maximal ancillary may not be unique; this is an issue traditionally considered most problematic in conditional inference. If it is not unique, then the conditional coverage probability might depend upon the choice. However, this does not affect the guaranteed absence of relevant subset in the corollary. We discuss this further in Section 2.4 and illustrate with an example in Appendix 2.5.3.

2.2.5 Computation of epistemic confidence

Our theory indicates that we get epistemic confidence from the full confidence density $c_f(\theta) \propto c_0(\theta)L(\theta; y)$. The corresponding coverage probability is either a marginal probability or a conditional probability given a maximal ancillary. The full likelihood $L(\theta; y)$ is almost always easy to compute. However, in order to get a correct coverage, the implied prior $c_0(\theta)$ must be computed using (2.8) or (2.13); in practice these can be difficult to evaluate. We illustrate through a series of examples some suitable approximations of $c_0(\theta)$ that are simpler to compute.

Suppose, for sample size $n = 1$, there is a statistic $t_1 \equiv T(y_1)$ that satisfies Condition 2.1, i.e. it allows us to construct a valid confidence density $c_m(\theta, t_1)$. The statistic t_1 trivially exists if y_1 itself leads to a valid confidence density. Then we can compute $c_0(\theta)$ based on $c_m(\theta; t_1)/L(\theta; t_1)$. First consider the case when $c_0(\theta)$ is free of the data. From the updating formula in Section 3 of

Pawitan and Lee (2021), the confidence density based on the whole data is

$$\begin{aligned} c_f(\theta; y) &\propto c_m(\theta; t_1) L(\theta; y_1|t_1) L(\theta; y_2 \cdots y_n) \\ &\propto c_0(\theta) L(\theta; t_1) L(\theta; y_1|t_1) L(\theta; y_2 \cdots y_n) \\ &= c_0(\theta) L(\theta; y_1) L(\theta; y_2 \cdots y_n) \\ &\propto c_0(\theta) L(\theta; y). \end{aligned} \tag{2.15}$$

Once $c_0(\theta)$ is available, (2.15) is highly convenient, since it is computationally straightforward. More importantly, as shown in some examples below, formula (2.15) works even when there is no sufficient estimate from the whole data for $n > 1$; see location-family model in Section 2.3.2 and the curved exponential model in Example 4.

When $c_0(\theta)$ depends on the data, it matters which y_i is used to compute it. In this case the updating formula is only an approximation. As long as the contribution of $\log c_0(\theta)$ to $\log c_f(\theta)$ is of order $O(1/n)$, we expect a close approximation as illustrated in Example 3.

2.3 Examples

2.3.1 Simple model

Example 4.1 (continued)**.** Based on y_1 alone, we have

$$c(\theta; y_1) \propto L(\theta; y_1) = 1 \quad \text{for } \theta \in \{y_1 - 1, y_1, y_1 + 1\},$$

so the implied prior $c_0(\theta) = 1$ for all θ. The full likelihood based on (y_1, y_2) is

$$L(\theta) = 1 \quad \text{for } \theta \in \{y_{(2)} - 1, y_{(1)} + 1\},$$

so, the full confidence density is $c_f(\theta) \propto L(\theta)$. For example, if $y_1 = 3$ and $y_2 = 5$, we do have 100% confidence that $\theta = 4$. And if $y_1 = y_2 = 3$, we only have 33.3% confidence for $\theta = 4$, though we have 100% confidence for $\theta \in \{2, 3, 4\}$. The MLE of θ is not unique, but we can choose $\widehat{\theta} = \bar{y}$ as the MLE. It is not sufficient, but $(\bar{y}, R)$ is, so R is a maximal ancillary. Indeed the full confidence values match the conditional probabilities given the range R. Furthermore, from Corollary 3, there is no further relevant subset, so the confidence is epistemic.□

2.3.2 Location family model

Suppose $y_1, \ldots, y_n$ are an iid sample from the location family with density

$$p_\theta(y_i) = f(y_i - \theta),$$

where $f(\cdot)$ is an arbitrary but known density. Based on y_1 alone,

$$c_m(\theta; y_1) = f(y_1 - \theta) = L(\theta; y_1),$$

so the implied prior is $c_0(\theta) = 1$. So, using formula (2.15), the full confidence density is

$$c_f(\theta) \propto L(\theta) = \prod_{i=1}^{n} f(y_i - \theta). \tag{2.16}$$

This is a remarkably simple way to arrive at the confidence density of θ and epistemic CIs without having to find the MLE and its distribution.

Without further specifications, the MLE $T \equiv \widehat{\theta}$ is not sufficient, so the marginal P-value $P_\theta(T \geq t)$ will not yield the full confidence. The distribution of the residuals $(y_i - \theta)$ are free of θ, so the set of differences $(y_i - y_j)$'s are ancillary. In his classic paper, Fisher (1934) showed that

$$p_\theta(\widehat{\theta}|a) = k(a)\frac{L(\theta)}{L(\widehat{\theta})},$$

where a is the set of differences from the order statistics $y_{(1)}, \ldots, y_{(n)}$. This means that the conditional likelihood based on $\widehat{\theta}|a$ matches the full likelihood (2.16), and the confidence level of CIs based on (2.16) will match the conditional coverage probability. Indeed, here $(\widehat{\theta}, a)$ is sufficient and a is maximal ancillary. Note however that (2.16) does not require any explicit knowledge or formula of the ancillary statistic.

2.3.3 Exponential family model

Let $y_1, \ldots, y_n$ be an iid sample from the exponential family with log-density

$$\log p_\theta(y_i) = \sum_{j=1}^{J} u_j(\theta)t_j(y_i) - A(\theta) + v(y_i). \tag{2.17}$$

The MLE is sufficient if $J = 1$, but not if $J > 1$. In the latter case, the family is called the curved exponential family. By Theorem 2.1, when $J = 1$ confidence statements based on the MLE will be epistemic. Our theory covers the continuous case in order to get exact coverage probabilities. Many important

members are discrete, which is more complicated because the definition of the P-value is not unique, and the coverage probability function is guaranteed not to match any chosen confidence level. We discuss an example in Appendix 2.5.2.

The standard evaluation of the confidence requires the tail probability of the distribution of the MLE, which in general has no closed form formula. Barndorff-Nielsen's (1983) approximate conditional density of the MLE $\widehat{\theta}$ is given by

$$p_\theta(\widehat{\theta}|a) = k|I(\widehat{\theta})|^{1/2}\frac{L(\theta)}{L(\widehat{\theta})} + O(n^{-1}), \tag{2.18}$$

where the MLE is the solution of $A'(\theta) = \sum_i \sum_j h'_j(\theta)t_j(y_i)$, a is the maximal ancillary and k is a normalizing constant that is free of θ. For $J = 1$ and the canonical parameter $h_1(\theta) = \theta$, the ancillary is null, and the approximation leads to the right-side P-value

$$P_\theta\{Z \geq r^*(\theta)\}, \quad r^*(\theta) \equiv r + \frac{1}{r}\log\frac{z}{r}, \tag{2.19}$$

where Z is the standard normal variate and

$$r = \text{sign}(\widehat{\theta} - \theta)\sqrt{w}, \quad z = |I(\widehat{\theta})|^{1/2}(\widehat{\theta} - \theta),$$

with $w = 2\log\{L(\widehat{\theta})/L(\theta)\}$ and $I(\widehat{\theta})$ the observed Fisher information. From the P-value we can get the corresponding confidence density and the implied prior.

Example 2.2. Let $y = (y_1, \cdots, y_n)$ be an iid sample from the gamma distri-

bution with mean one and shape parameter θ. The density is given by

$$p_\theta(y_i) = \frac{1}{\Gamma(\theta)}\theta^\theta y_i^{\theta-1} e^{-\theta y_i},$$

so we have an exponential family model with

$$t(y_i) = -y_i + \log y_i, \quad A(\theta) = \log\Gamma(\theta) - \theta\log\theta.$$

To use formula (2.15), we first find the implied prior density using $t_1 \equiv t(y_1)$ alone:

$$c_0(\theta) \propto \frac{c(\theta; t_1)}{L(\theta; t_1)}, \tag{2.20}$$

where $c_1(\theta) = \partial\{P_\theta(T_1 \geq t_1)\}/\partial\theta$ and $L(\theta; t_1) = p_\theta(y_1)$. The probability $P_\theta(T_1 \geq t_1)$ is an incomplete gamma integral, which is computed numerically. The implied prior is shown in Figure 2.1(a). So from (2.15), we get the confidence density

$$c_f(\theta) \propto c_0(\theta)L(\theta) = c_0(\theta)\prod_{i=1}^{n} p_\theta(y_i). \tag{2.21}$$

For an example with $n = 5$ and $\sum_i t(y_i) = -5.8791$, which corresponds to the MLE $\widehat{\theta} = 3$, the confidence density is given by the solid line in Figure 2.1(b). The normalized likelihood function is also shown by the dashed line, which is quite distinct from the confidence density.

By comparison, to get the marginal confidence density based on the P-value formula (2.19), we need

$$w = -2n\log\Gamma(\theta) + 2n\log\Gamma(\widehat{\theta}) + 2n(\theta\log\theta - \widehat{\theta}\log\widehat{\theta}) + 2(\theta - \widehat{\theta})\sum(\log y_i - y_i),$$

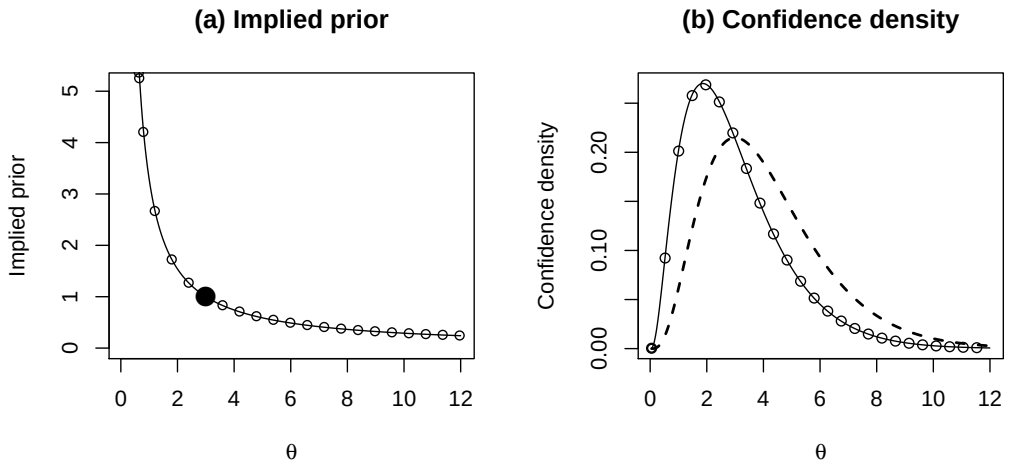


Figure 2.1: (a) Implied priors of θ computed from (2.20) (solid) and (2.19) ($\circ$). Both are normalized such that they are equal to one at the MLE ($\bullet$). (b) The normalized likelihood function (dashed) and the confidence densities based on a sample with $n = 5$ using (2.21) (solid) and (2.19) ($\circ$).

where $\widehat{\theta}$ is the solution of

$$n\psi(\theta) - n\log\theta - n = \sum t(y_i),$$

with $\psi(\theta) \equiv \partial \log\Gamma(\theta)/\partial\theta$, and the observed Fisher information is

$$I(\widehat{\theta}) = n\{\psi'(\widehat{\theta}) - 1/\widehat{\theta}\}.$$

For the data example, the MLE $\widehat{\theta}$ has to be computed numerically. The circle points in Figure 2.1(b) are the marginal confidence density based on the same sample above. As expected, this tracks almost exactly the one given by formula (2.21). The corresponding implied prior based on $c_m(\theta)/L(\theta)$ is given in Figure 2.1(a), also closely matching the implied prior based on $n = 1$. □

Example 2.3. This is an example where $c_0(\theta)$ is data dependent. Let $y = (y_1, \ldots, y_n)$ be iid sample from $N(\theta, \theta)$ for $\theta > 0$. The log-density is given by

$$\log p_\theta(y) = -\frac{n}{2}\log(2\pi\theta) - \frac{1}{2}\left(\sum_i y_i^2/\theta - 2\sum_i y_i + n\theta\right),$$

so this is a regular exponential family with sufficient statistic $T(y) = \sum_i y_i^2$. The marginal confidence density $c_m(\theta)$ can be computed based on the non-central χ^2 distribution for $T(y)$. For $n = 1$, $T(y_1) = t_1 = y_1^2$ is sufficient, and

$$\begin{aligned}
C(\theta; y_1) &= P_\theta(Y_1^2 \geq y_1^2) = 1 - \Phi\left(\frac{|y_1| - \theta}{\sqrt{\theta}}\right) + \Phi\left(\frac{-|y_1| - \theta}{\sqrt{\theta}}\right) \\
c(\theta; y_1) &= \frac{1}{2}\phi\left(\frac{|y_1| - \theta}{\sqrt{\theta}}\right)\left(\frac{|y_1|}{\theta\sqrt{\theta}} + \frac{1}{\sqrt{\theta}}\right) + \frac{1}{2}\phi\left(\frac{-|y_1| - \theta}{\sqrt{\theta}}\right)\left(\frac{|y_1|}{\theta\sqrt{\theta}} - \frac{1}{\sqrt{\theta}}\right) \\
L(\theta; y_1) &= \phi\left(\frac{y_1 - \theta}{\sqrt{\theta}}\right).
\end{aligned}$$

Note that the log-density is not of the form (2.11); here Condition 2.2 holds using $g(\theta) = \theta^{-3/2}e^{\theta/2}$. Since the implied prior is data-dependent, the full confidence density depends on which y_i is used to compute the implied prior:

$$c_{fi}(\theta) = c_0(\theta; y_i)L(\theta; y)$$

In Figure 2.2, for $n = 3$, we compare $c_{fi}(\theta)$ using three different versions of $c_0(\theta)$ based on y_i for $i = 1, 2, 3$. Two datasets are shown, where the first has a small variance and the second a large variance. These are also compared with the marginal confidence $c_m(\theta)$. As shown in the figure, even for such a small dataset, the effect of the data dependence in this case is negligible. □

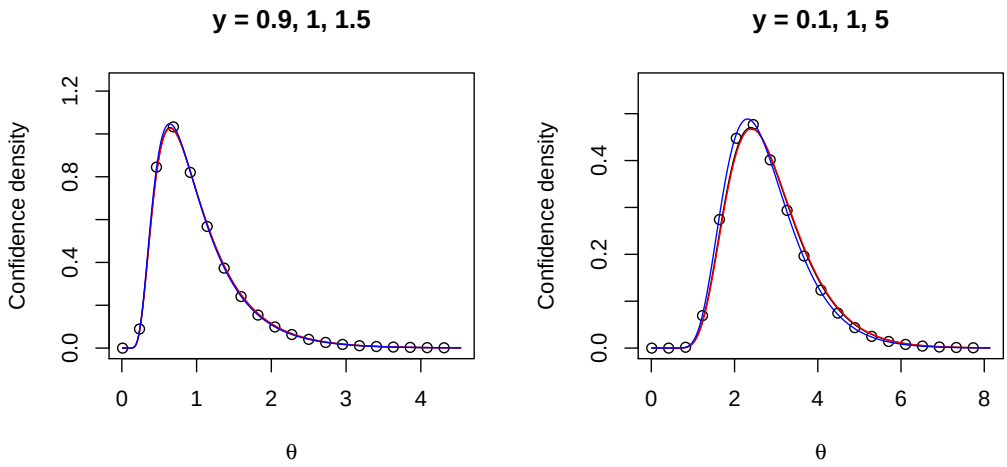


Figure 2.2: $c_{fi}(\theta)$ for $i = 1, 2, 3$ and $c_m(\theta)$ (circle) based on (a) $y = (0.9, 1, 1.5)$ and (b) $y = (0.1, 1, 5)$. In each panel, all three curves c_{fi} are actually drawn in solid line, but we can only see one curve because they track each other very closely.

Example 4. This example of a curved exponential model illustrates complex cases, where the MLE is not sufficient. Let $y_1, \ldots, y_n$ be iid sample from $N(\theta, \theta^2)$ for $\theta > 0$. Here $(\sum y_i^2, \sum y_i)$ is minimal sufficient, and the likelihood function is

$$L(\theta; y) = (2\pi\theta^2)^{-n/2} \exp\left(-\sum y_i^2/2\theta^2 + \sum y_i/\theta - n/2\right).$$

The MLE is given by

$$\widehat{\theta} = \widehat{\theta}(y) = \frac{-\sum y_i + \sqrt{(\sum y_i)^2 + 4\sum y_i^2}}{2n}$$

with a maximal ancillary $A(y) = \sum y_i / \left(\sum y_i^2\right)^{1/2}$. In terms of $(\widehat{\theta}, a) \equiv (\widehat{\theta}(y), A(y))$

with $b \equiv a^2 + a\sqrt{a^2+4n}$, the full likelihood is

$$L(\theta; y) = (2\pi\theta^2)^{-n/2} \exp\left(-\frac{(b+2n)\widehat{\theta}^2}{4\theta^2} + \frac{b\widehat{\theta}}{2\theta} - \frac{n}{2}\right).$$

First consider the confidence distribution based on y_1,

$$C_m(\theta; y_1) = P_\theta(Y_1 \geq y_1) = 1 - \Phi\left(\frac{y_1 - \theta}{\theta}\right).$$

We can see immediately that if we use y_1 as the statistic, the term inside the bracket converges to -1 as $\theta \to \infty$, and the confidence distribution goes to $1 - \Phi(-1) = 0.84$. Hence y_1 does not satisfy Condition 1. However, we can show that $t_1 \equiv |y_1|$ does lead to a valid confidence distribution:

$$\begin{aligned} C_m(\theta; t_1) &= P(|Y_1| \geq t_1) \\ &= 1 - \Phi\left(\frac{t_1 - \theta}{\theta}\right) + \Phi\left(\frac{-t_1 - \theta}{\theta}\right). \end{aligned}$$

After taking derivatives to get $c_{m1}(\theta; t_1)$ and $L(\theta; t_1)$, the implied prior based on t_1 is

$$c_0(\theta) \propto c_{m1}(\theta; t_1)/L(\theta; t_1) \propto \theta^{-1}.$$

The updating formula (2.15) gives the full confidence density

$$c_f(\theta; y) \propto c_0(\theta) L(\theta; y) \propto \frac{1}{\theta^{n+1}} \exp\left[\sum \frac{y_i^2}{2\theta^2} + \sum \frac{y_i}{\theta}\right]. \tag{2.22}$$

In Appendix 2.5.1 we show: (i) we get the same implied prior if we start with the MLE $\widehat{\theta}_1$ based on y_1 alone, or with the conditional $\widehat{\theta}_1|a_1$. So even though t_1 or $\widehat{\theta}_1$ are not sufficient, they still lead to a valid implied prior for the

full confidence; (ii) the conditional confidence density derived using Barndorff-Nielsen's formula (2.18) also gives the same implied prior; (iii) the confidence $c_m(\theta; \widehat{\theta}_1, \cdots, \widehat{\theta}_n)$ with $\widehat{\theta}_i = \widehat{\theta}(y_i)$ is a valid confidence density because $C_m(\theta \in \mathrm{CI}(\widehat{\theta}_1, \cdots, \widehat{\theta}_n)) = P_\theta(\theta \in \mathrm{CI}(\widehat{\theta}_1, \cdots, \widehat{\theta}_n))$. However, it is not epistemic because it does not use the full likelihood, so there is a potential loss of information.

To compare with exact theoretical results, we refer to Hinkley (1977), who derived the conditional density of $w = \theta^{-1} \left(\sum y_i^2\right)^{1/2}$ given the ancillary as

$$p(w|a) = w^{n-1} \exp\{-(w-a)^2/2\}/I_{n-1}(a) \tag{2.23}$$

where $I_{n-1}(a) = \int_0^\infty x^{n-1} \exp\{-(x-a)^2/2\} dx$. Hinkley used this result to get the conditional score-test. Let $T(y) = \left(\sum y_i^2\right)^{1/2}$, then we have

$$C_c(\theta; t|a) = P_\theta(T \geq t|a) = P(W \geq w|a) = 1 - F_a(w)$$

where $F_a(w) = \int_0^w p(u|a) du$. Thus, the confidence density becomes

$$c_c(\theta; t|a) = -\frac{\partial F_a(w)}{\partial \theta} = p(w = t/\theta|a)\ t/\theta^2 = p_\theta(t|a)\ t/\theta$$

so that the implied prior becomes $c_0(\theta; t|a) \propto c_c(\theta; t|a)/L(\theta; t|a) \propto 1/\theta$, which is again the same as the result from t_1. Thus, we have

$$c_c(\theta; t|a) = c_c(\theta; \widehat{\theta}|a) = c_f(\theta; y).$$

As numerical illustrations, we compare the exact conditional P-value $P_\theta(T > t|a)$ based on (2.23) for testing H_0: $\theta = 1$, the corresponding full confidence $C_f(\theta)$ at $\theta = 1$ and the P-value based on the score test. The latter was com-

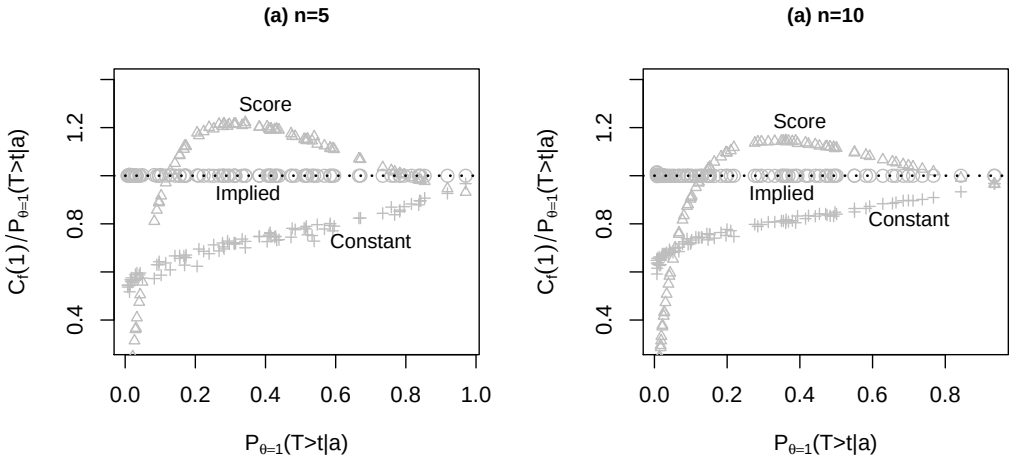


Figure 2.3: Approximations of the conditional P-value $P_\theta(T \geq t|a)$ to test H_0: $\theta = 1$ from $N(\theta, \theta^2)$. The x-axis is the exact theoretical value based on (2.23). The y-axis is ratio of $C_f(1) \equiv \int_0^1 c_f(\theta)d\theta$ to $P_\theta(T \geq t|a)$ at $\theta = 1$, where $c_f(\theta)$ is computed using the constant prior $c_0(\theta) = 1$ (+) and the implied prior $c_0(\theta) = 1/\theta$ ($\circ$). The corresponding P-value from the score test using Fisher's observed information is also shown ($\triangle$). (a) For $n = 5$ (b) For $n = 10$.

puted using the observed Fisher information, suggested by Hinkley (1977) to have good conditional properties. For Figure 2.3(a), we generate 100 datasets with $n = 5$ from $N(\theta, \theta^2)$ at $\theta = 1.2$. The full confidence $C_f(\theta \leq 1)$ is computed using the implied prior $c_0(\theta) \propto 1/\theta$, and a constant prior $c_0(\theta) \propto 1$. Panel (b) shows the result for $n = 10$. The full confidence with the implied prior $c_0(\theta) \propto 1/\theta$ agrees with the exact conditional P-value. The use of a non-implied constant prior $c_0(\theta) \propto 1$ generates over-optimistic P-values, particularly for small values. While Hinkley's (1977) score test appears correct on average, in these small samples, it has a poor conditional property. □

2.4 Discussion

We have described a concept of epistemic confidence for an observed confidence interval. Fisher tried to achieve the same purpose with the fiducial probability, but the use of the word 'probability' had generated much confusion and controversies, so the concept of fiducial probability has been practically abandoned. However, the confidence concept is mainstream, although it comes with a frequentist interpretation only, so it applies not to the observed interval but to the procedure. The confidence may not be a probability but an extended likelihood (Pawitan and Lee, 2021), whose ratio is meaningful in hypothesis testing and statistical inferences (Lee and Bjørnstad, 2013). The confidence is logically distinct from the classical likelihood. Our results show that we can turn a classical likelihood into a confidence density by multiplying it with an implied prior. Furthermore, we get epistemic confidence by establishing the absence of relevance subsets.

Schweder and Hjort (2016) and Schweder (2018) have been strong proponents of interpreting confidence as 'epistemic probability.' We are in general agreement with their sentiment, but it is unclear which version of probability this is. The only established and accepted epistemic probability is the Bayesian probability, but in their writing, the confidence concept is clearly non-Bayesian. Our use of the Dutch Book defines normatively the epistemic value of the confidence while staying within the non-Bayesian framework.

Conditional inference (Reid, 1995) has traditionally been the main area of statistics that tries to address the epistemic content of confidence intervals. However, it goes half-way to the end goal of epistemic confidence that Fisher would want. The general lack of unique maximal ancillary is a great stumbling

block, where it is then possible to come up with distinct relevant subsets with distinct conditional coverage probabilities. This raises an unanswerable question: What is then the 'proper' confidence for the observed interval? Our logical tool of the betting market overcomes this problem – in this case, the market cannot settle in an unambiguous price. But Corollary 3 still holds in the sense that you're still protected from the Dutch Book. We discuss this further with an example Appendix 2.5.3.

An issue arises in the subjective probability framework if there is a mismatch between the subjective and objective probabilities. Extra principles have been proposed to deal with it. In Lewis's (1980) 'Principal Principle',

$$P_s\{A|\text{Chance}(A) = x\} \equiv x,$$

where P_s denotes the subjective probability and 'Chance' the objective probability. So, the Principle simply declares that the subjective probability must be set to be equal to the objective probability, if the latter exists. Our Dutch Book argument can be used to justify the Principle, so the principle does not have to come out of the blue with no logical motivation. However, it is worth noting that epistemic confidence is not simply declared equal to probability. Instead, it is the consequence of a theorem that establishes no relevant subset in order to avoid the Dutch Book. In our setup, the frequentist probability applies to a market involving a large number of independent players. Moreover, the rational personal betting price is no longer 'subjective', for example, in the choice of the prior. Thus, the conceptual separation of the personal and the market prices allow both epistemic and frequentist meaning of confidence.

Finally, our use of money and bets to define epistemic confidence has some

echoes in the game-theoretic foundation of probability (Shafer and Vovk, 2019), an ambitious rebuilding of probability without measure theory. However, their key concept is a sequential game between two players. The word 'sequential' clearly implies that the game is not meant to involve a risk-free profit from a single transaction that we want in a Dutch Book. Our usage of probability is fully within the Kolmogorov axiomatic system, and we make a clear distinction between probability

2.5 Appendix

This appendix provides additional details and examples for (i) the curved exponential model in Example 4; (ii) a discrete case; (iii) a case where the maximal ancillary is not unique.

2.5.1 Curved exponential model: $N(\theta, \theta^2)$

We give more details of the $N(\theta, \theta^2)$ model, where different choices of initial statistics lead to the same implied prior. Denote the MLE and the ancillary based only on y_1 by $\widehat{\theta}_1 \equiv \widehat{\theta}(y_1)$ and a_1. The confidence distribution of $\widehat{\theta}_1$ is

$$\begin{aligned} C_m(\theta; \widehat{\theta}_1) &= P_\theta(\widehat{\theta}(Y_1) \geq \widehat{\theta}_1) = P_\theta\left(\frac{-Y_1 + \sqrt{5Y_1^2}}{2} \geq \widehat{\theta}_1\right) \\ &= P_\theta\left(Y_1 \geq \frac{2\widehat{\theta}_1}{\sqrt{5}-1}\right) + P_\theta\left(Y_1 \leq \frac{-2\widehat{\theta}_1}{\sqrt{5}+1}\right) \\ &= 1 - \Phi\left(\frac{1+\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta} - 1\right) + \Phi\left(\frac{1-\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta} - 1\right). \end{aligned}$$

Then $\lim_{\theta\to\infty} C_m(\theta;\widehat{\theta}_1) = 1$ and $\lim_{\theta\to 0} C_m(\theta;\widehat{\theta}_1) = 0$, so $C_m(\theta;\widehat{\theta}_1)$ is a proper distribution function. The corresponding confidence density is given by

$$c_{m1}(\theta;\widehat{\theta}_1) = \frac{1+\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta^2}\phi\left(\frac{1+\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta} - 1\right) - \frac{1-\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta^2}\phi\left(\frac{1-\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta} - 1\right).$$

The implied prior based on $\widehat{\theta}_1$ is

$$c_{0m1}(\theta;\widehat{\theta}_1) \propto c_{m1}(\theta;\widehat{\theta}_1)/L(\theta;\widehat{\theta}_1) \propto \theta^{-1}$$

where $L(\theta;\widehat{\theta}_1)$ is the likelihood function based on $\widehat{\theta}_1$.

The conditional confidence distribution based on $\widehat{\theta}_1|a_1$ is

$$C_c(\theta;\widehat{\theta}_1|a_1) = P_\theta(\widehat{\theta}(Y_1) \geq \widehat{\theta}_1|a_1) = \frac{1}{\Phi(a_1)}\Phi\left(\frac{-a_1-\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta} + a_1\right),$$

which is now a valid confidence distribution, with density

$$c_c(\theta;\widehat{\theta}_1|a_1) = \frac{\partial}{\partial\theta}C_c(\theta;\widehat{\theta}_1|a_1) = \frac{1}{\Phi(a_1)}\frac{a_1+\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta^2}\phi\left(\frac{-a_1-\sqrt{5}}{2}\frac{\widehat{\theta}_1}{\theta} + a_1\right).$$

The implied prior is

$$c_0(\theta;\widehat{\theta}_1|a_1) \propto c_c(\theta;\widehat{\theta}_1|a_1)/L(\theta;y_1) \propto \theta^{-1},$$

the same as the one derived based on $\widehat{\theta}_1$.

On the other hand, if we construct the full confidence densities by

$$c_{fi}(\theta;\widehat{\theta}(y_i), y_{(-i)}) \propto c_{mi}(\theta;\widehat{\theta}(y_i))L(\theta;y_{(-i)}),$$

then the resulting confidence density depends on the choice of y_i. In this case we should consider $c_{fi}(\theta; \widehat{\theta}_i, y_{(-i)})$ as an approximation to $c_f(\theta; y)$. Figure 2.4 plots n confidence densities $c_{fi}(\theta; \widehat{\theta}_i, y_{(-i)})$ (solid) and $c_f(\theta; y)$ (circle) with y_i from $N(1, 1)$. As shown in (b), when n becomes large, the difference becomes negligible and $c_{fi}(\theta; \widehat{\theta}_i, y_{(-i)})$ gets closer to $c_f(\theta; y)$ (circle).

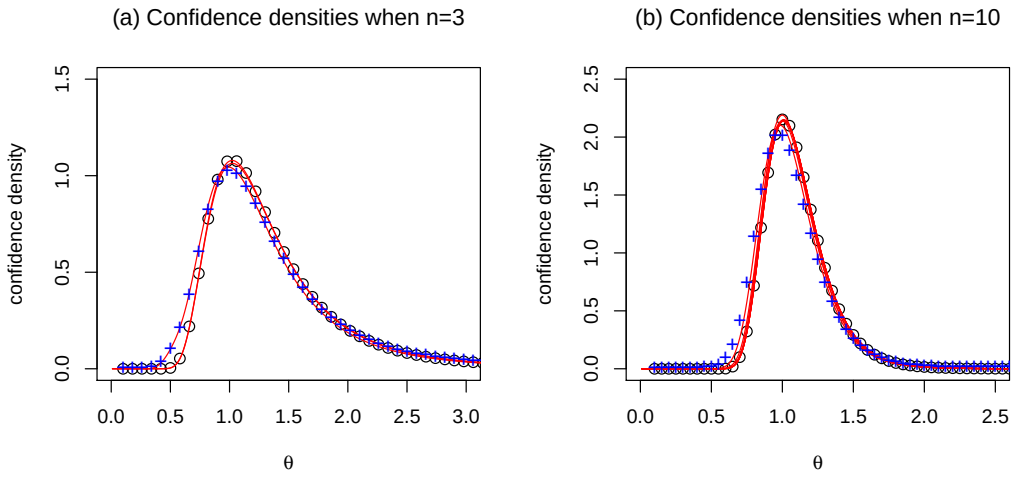


Figure 2.4: $c_{fi}(\theta; \widehat{\theta}_i)$ for $i = 1, ..., n$ (solid), $c_f(\theta; y)$ (circle) and $c_m(\theta; \widehat{\theta}_1, \cdots, \widehat{\theta}_n)$ (cross). (a) $n = 3$, (b) $n = 10$.

So

$$c_{fi}(\theta; \widehat{\theta}(y_i), y_{(-i)}) \propto \theta^{-1} L(\theta; \widehat{\theta}_i) L(\theta; y_{(-i)}) = \frac{L(\theta; \widehat{\theta}_i)}{L(\theta; y_i)} c_0(\theta) L(\theta; y).$$

There is a loss of information caused by using $c_{mi}(\theta)$, due to the sign of y_i as captured by $L(\theta; a_i)$. This is negligible even in small samples; see Figure 2.4. However, the marginal confidence

$$c_m(\theta; \widehat{\theta}_1, \cdots, \widehat{\theta}_n) \propto \theta^{-1} L(\theta; \widehat{\theta}_1, \cdots, \widehat{\theta}_n)$$

has a larger loss of information, as shown in both Figure 2.4(a) and (b).

It is also possible to compute the conditional confidence density by using Barndorff-Nielsen's formula as given in the main text, and to show that we end up with the same implied prior $c_0(\theta) = 1/\theta$. Firstly, the likelihood ratio is given by

$$\frac{L(\theta)}{L(\widehat{\theta})} = \frac{\widehat{\theta}^n}{\theta^n} \exp\left[-\frac{b+2n}{4}\left(\frac{\widehat{\theta}^2}{\theta^2} - 1\right) + \frac{b}{2}\left(\frac{\widehat{\theta}}{\theta} - 1\right)\right],$$

where $b \equiv a^2 + a\sqrt{a^2+4n}$, and the observed Fisher information

$$I(\widehat{\theta}) = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\Big|_{\theta=\widehat{\theta}} = \frac{b+4n}{2\widehat{\theta}^2}.$$

Then we have

$$p_\theta(\widehat{\theta}|a) \approx \frac{c}{\sqrt{2}} \frac{\sqrt{b+4n}}{\widehat{\theta}} \frac{\widehat{\theta}^n}{\theta^n} \exp\left[-\frac{b+2n}{4}\left(\frac{\widehat{\theta}^2}{\theta^2} - 1\right) + \frac{b}{2}\left(\frac{\widehat{\theta}}{\theta} - 1\right)\right].$$

Let $U \equiv \widehat{\theta}(Y)/\theta$ and let $u = \widehat{\theta}(y)/\theta$, then the conditional density of $u|a$ becomes

$$p_\theta(u|a) \approx \frac{c\sqrt{b+4n}}{\sqrt{2}} u^{n-1} \exp\left[-\frac{b+2n}{4}\left(u^2 - 1\right) + \frac{b}{2}\left(u - 1\right)\right],$$

which does not contain θ. Let $F_a(u) = \int p(u|a)du$, then

$$C_c(\theta; \widehat{\theta}|a) = P_\theta(\widehat{\theta}(Y) \geq \widehat{\theta}) = P_\theta(U \geq \widehat{\theta}/\theta) = 1 - F_a(\widehat{\theta}/\theta).$$

It gives the conditional confidence density

$$c_c(\theta;\widehat{\theta}|a) = -\frac{\partial F_a(\widehat{\theta}/\theta)}{\partial\theta} \approx \frac{c\sqrt{b+4n}}{\sqrt{2}}\frac{\widehat{\theta}^n}{\theta^{n+1}}\exp\left[-\frac{b+2n}{4}\left(\frac{\widehat{\theta}^2}{\theta^2}-1\right)F+\frac{b}{2}\left(\frac{\widehat{\theta}}{\theta}-1\right)\right],$$

and implied prior $c_0(\theta;\widehat{\theta}|a) \propto c_c(\theta;\widehat{\theta}|a)/L(\theta;y) \propto 1/\theta$. Thus, the conditional confidence density from Barndorff-Nielsen's formula becomes

$$c_c(\theta;\widehat{\theta}|a) = c_f(\theta;y) \propto \theta^{-1}L(\theta;y),$$

which, as we would expect, is the same as the full confidence.

2.5.2 Discrete case

A complication arises in the discrete case since the definition of the P-value is not unique, and the coverage probability function is guaranteed not to match any chosen confidence level. Given the observed statistic $T = t$, among several candidates, the mid P-value

$$P_\theta(T > t) + \frac{1}{2}P_\theta(T = t)$$

is often considered the most appropriate (Lancaster, 1961).

We shall discuss the specific case of the binomial and negative binomial models: $Y_1 \sim Bin(n,\theta)$ and $Y_2 \sim NB(y,\theta)$. The two models have an identical likelihood, proportional to $L(\theta) \propto \theta^y(1-\theta)^{n-y}$, but have different probability

mass functions, respectively

$$P_\theta(Y_1 = y) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \text{ and } P_\theta(Y_2 = n) = \binom{n-1}{y-1}\theta^y(1-\theta)^{n-y}$$

Thus, they have the common MLE $\widehat{\theta} = \widehat{\theta}_1 = \widehat{\theta}_2 = y/n$. However, the two MLEs have different supports

$$\widehat{\theta}_1 \in \left\{0, \frac{1}{n}, \cdots, \frac{n-1}{n}, 1\right\} \text{ and } \widehat{\theta}_2 \in \left\{1, \frac{y}{y+1}, \frac{y}{y+2}\cdots\right\},$$

and therefore $\widehat{\theta}_1$ and $\widehat{\theta}_2$ have different distribution, which lead to different P-values. Statistical models such as the binomial and negative binomial models describe how the unobserved future data will be generated. Thus, all the information about θ in the data and in the statistical model is in the extended likelihood. The use of the mid P-values

$$\begin{aligned}
C(\theta; y_1 = y) &= \frac{1}{2}P_\theta\left(\widehat{\theta}_1 = \frac{y}{n}\right) + P_\theta\left(\widehat{\theta}_1 > \frac{y}{n}\right) \\
&= \frac{1}{2}\Big(I_\theta(y, n-y+1) + I_\theta(y+1, n-y)\Big) \\
C(\theta; y_2 = n) &= \frac{1}{2}P_\theta\left(\widehat{\theta}_2 = \frac{y}{n}\right) + P_\theta\left(\widehat{\theta}_2 > \frac{y}{n}\right) \\
&= \frac{1}{2}\Big(I_\theta(y, n-y+1) + I_\theta(y, n-y)\Big)
\end{aligned}$$

lead to different confidence densities

$$\begin{aligned}
c(\theta; y_1) =& \frac{1}{2}\left(\frac{\theta^{y-1}(1-\theta)^{n-y}}{B(y, n-y+1)} + \frac{\theta^y(1-\theta)^{n-y-1}}{B(y+1, n-y)}\right) \\
c(\theta; y_2) =& \frac{1}{2}\left(\frac{\theta^{y-1}(1-\theta)^{n-y}}{B(y, n-y+1)} + \frac{\theta^{y-1}(1-\theta)^{n-y-1}}{B(y, n-y)}\right)
\end{aligned}$$

where $I_\theta(\cdot,\cdot)$ is the regularized incomplete beta function and $B(\cdot,\cdot)$ is the beta function.

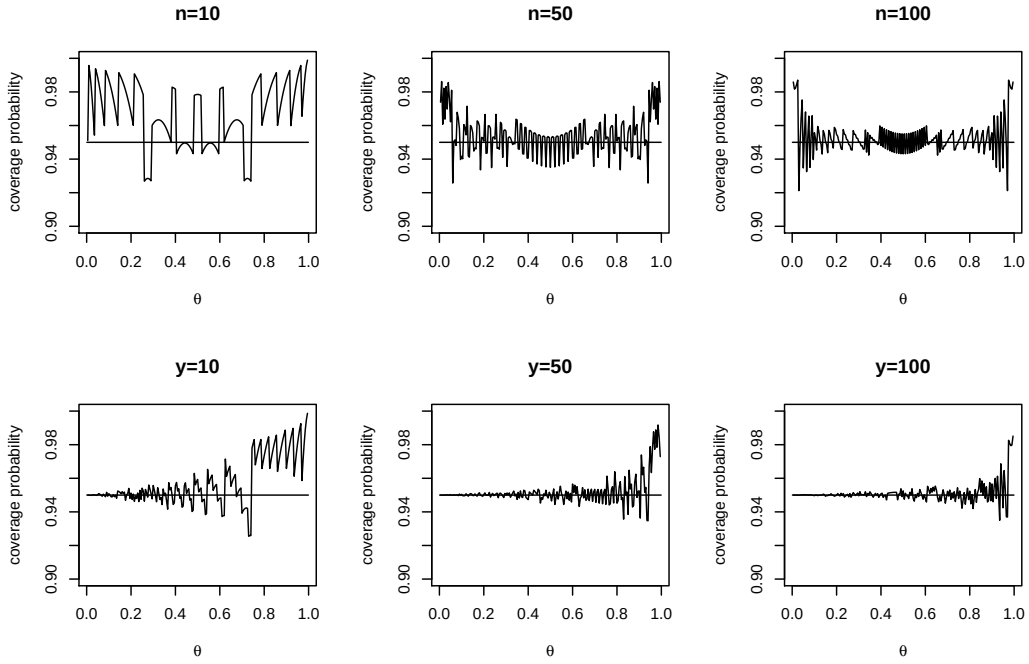


Figure 2.5: Coverage probabilities of the intervals from the confidence distribution based on the mid p-value for binomial models at $n = 10, 50, 100$ (top) and negative binomial models at $y = 10, 50, 100$ (bottom).

Figure 2.5 shows the coverage probabilities of the 95% two-sided confidence procedure based on the mid p-value of $\widehat{\theta}$ for binomial models and negative binomial models. We can see that the coverage probabilities fluctuate around 0.95 but they are not consistently biased in one direction. Moreover, as n or y becomes larger, the difference between the coverage probability and the confidence becomes smaller. In discrete case, it is not possible to access the exact objective coverage probability of the CI procedure. Here the confidence is a consistent estimate of the objective coverage probability. In negative binomial

models with $\theta = 1$, $y = n$ with probability 1, so that it behaves like binomial confidence procedure for $n = y$.

Besides information in the likelihood, the confidence uses information from the statistical model. Consider two different statistical models, M1: $N = X+1$ where $X \sim \text{Poisson}(\eta_1)$ and $Y_1|N = n \sim Bin(n, \theta)$ and M2: $Y = X + 1$ where $X \sim \text{Poisson}(\eta_2)$ and $Y_2|Y = y \sim NB(y, \theta)$. In M1, $Y_1|N = n$ and in M2, $Y_2|Y = y$ have common likelihood, but they are different models, so that they have no reason to have a common confidence.

2.5.3 When maximal ancillaries are not unique

When the maximal ancillary is not unique, the conditional coverage probability may depend on the choice of the ancillary. However, the lack of unique ancillary does not affect the validity of Corollary 3 in the main text. We illustrate here with an example from Evans (2013). The data $y = (y_1, y_2)$ are sampled from a joint distribution with probabilities under θ given in the following table:

(y_1, y_2)	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
$\theta = 1$	$1/6$	$1/6$	$2/6$	$2/6$
$\theta = 2$	$1/12$	$3/12$	$5/12$	$3/12$

Here both the data y and parameter θ are discrete. Strictly, our theory does not cover this case, but we shall use it because it can still illustrate clearly the issues with non-unique maximal ancillaries. The marginal probabilities are

$$P_\theta(Y_1 = 1) = 1/3; \quad P_\theta(Y_1 = 2) = 2/3$$
$$P_\theta(Y_2 = 1) = P_\theta(Y_2 = 2) = 1/2,$$

for $\theta = 1, 2$. So both Y_1 and Y_2 are ancillaries, i.e., their probabilities do not depend on θ. The conditional probabilities of (y_1, y_2) given $Y_1 = 1$ are

(y_1, y_2)	$(1,1)$	$(1,2)$
$\theta = 1$	$1/2$	$1/2$
$\theta = 2$	$1/4$	$3/4$

and, given $Y_2 = 1$ are

(y_1, y_2)	$(1,1)$	$(2,1)$
$\theta = 1$	$1/3$	$2/3$
$\theta = 2$	$1/6$	$5/6.$

Based on the unconditional model, on observing $(y_1, y_2) = (1, 1)$, we have the likelihood function $L(\theta = 1) = 1$ and $L(\theta = 2) = 1/2$, so the MLE $\widehat{\theta} = 1$. For $(y_1, y_2) = (2, 2)$ we have a different likelihood, but still $\widehat{\theta} = 1$. This means we cannot reconstruct the likelihood based on the MLE alone, hence MLE is not sufficient. But we can see immediately that we get the same likelihood function under the conditional model given $Y_1 = 1$ or given $Y_2 = 1$, so conditioning on each ancillary recovers the full likelihood and each ancillary is maximal.

Now consider using the MLE itself as a 'CI'. Conditional on the ancillaries, the probability that the MLE is correct is

$$P_1(\widehat{\theta} = \theta | Y_1 = 1) = P_1(Y_2 = 1 | Y_1 = 1) = 1/2$$
$$P_2(\widehat{\theta} = \theta | Y_1 = 1) = P_2(Y_2 = 2 | Y_1 = 1) = 3/4$$
$$P_1(\widehat{\theta} = \theta | Y_2 = 1) = P_1(Y_2 = 2 | Y_2 = 1) = 1/3$$
$$P_2(\widehat{\theta} = \theta | Y_2 = 1) = P_2(Y_2 = 1 | Y_2 = 1) = 5/6$$

These conditional 'coverage probabilities' are indeed distinct from each other. However, comparing the conditional coverage probabilities given Y_1 to that given Y_2, there is no consistent non-trivial bias in one direction across θ. So if you use Y_1 as the ancillary, you cannot construct further relevant subsets based on Y_2. This is the essence of our remark after Corollary 3 that the lack of unique maximal ancillary does not affect the validity of the corollary.

Unfortunately, in this example, the P-value is not defined because the parameter θ can be an unordered label. So it is not possible to compute any version of confidence function or any implied prior. In the continuous case, we define CI to satisfy $P_\theta(\theta \in \text{CI}) = \gamma$ for all θ. However, in discrete cases, it is often not possible for the coverage probabilities to be same for all θ, which violates the condition of Theorem 2.1. Fisher (1973) suggested that for problems such as this, the structure is not sufficient to allow an unambiguous probability-based inference, so only the likelihood is available.

Chapter 3

Point mass in confidence distributions

3.1 Introduction

Fisher (1930) introduced the fiducial distribution (FD) as an alternative to the Bayesian posterior distribution, but it is not certain whether the FD obeys the Kolmogorov's axioms (Edwards, 1977). Wilkinson (1977) interpreted that the FD is non-coherent and not the probability. Though the use of FD has been controversial, it becomes of recent interest, arising as the confidence distribution (CD). Efron (1998) noted,

> "Maybe Fisher's biggest blunder will become a big hit in the 21st century!"

Thus, it is important to understand the advantages of CD. Integrating out nuisance parameters is legitimate to obtain the marginal posterior. Analogous

to the marginalization paradox of Dawid et al. (1973), an integrated CD may be no longer a valid CD. Stein's (1959) problem is a benchmark of a drawback of the integrated CD, which yields the CI with bad behavior (Schweder and Hjort, 2016; Wilkinson, 1977). Wilkinson (1977) introduced the marginal CD and Bernardo (1979) introduced the reference posterior (RP) to resolve this paradoxical behavior.

The confidence is Neymanian interpretation of Fisher's FD (Schweder and Hjort, 2016). However, Fisher objected Neyman's coverage probability, because scientists will never repeat the same experiment. In this chapter, we demonstrate the complementary nature of confidence and FD, emphasizing their synergy rather than conflict. We hereby refer to them collectively as the CD, following the suggestion of Efron (1998). Pawitan and Lee (2021) showed that the CD is not a probability but an extended likelihood, which is necessary to avoid probability-related paradoxes (Pawitan and Lee, 2017). Integrating out the nuisance parameters is not a legitimate way to obtain the marginal extended likelihood. This implies that the integrated CD is not a proper CD in general, and it is no longer a paradox because the CD is not a probability. The extended likelihood principle (Bjørnstad, 1996) states that the extended likelihood contains all the information in the data. Pawitan et al. (2023) proposed an epistemic CD associated with the full data likelihood, which leaves no room for a relevant subset. They obtained a marginal CD with the full data likelihood by conditioning on maximal ancillary statistics.

In this chapter, we show that the CD does not have a fundamental deficiency of probabilistic inference, called the probability dilution. In satellite conjunction analysis, the probability of collision plays an important role in

risk assessment. However, a probability dilution occurs in which the use of lower-quality data appears to reduce the probability of collision. This results in a severe and persistent underestimate of risk exposure. Cunen et al. (2020) derived a marginal CD for satellite conjunction problem, but Martin et al. (2021) pointed out that the marginal CD has a drawback of not allowing two-sided CIs. Stein's problem is also a kind of probability dilution with increasing number of parameters k, whereas the satellite problem is a probability dilution with increasing variance σ^2. Hannig (2009) generalized the FD by the GFD, which is easily implementable in practical applications, but he claimed not to attempt to derive a paradox-free FD. We show that the integrated CD in Stein's problem can be viewed as a GFD. Thus, the GFD cannot avoid the probability dilution. The RP can improve the GFD with moderate k and σ, but it cannot avoid the probability dilution entirely. In this chapter, we study the role of point mass in the CD. The point mass in CD has been considered as paradoxical (Schweder and Hjort, 2016; Wilkinson, 1977), but we show that the point mass is necessary to maintain the stated coverage probability (confidence feature) of CI. In order to maintain the confidence feature, the CD cannot allow two-sided interval in some occasions. Thus, this property is not a drawback but an advantage of the CD. In consequence, the CD can overcome the probability dilution entirely, whereas the GFD and RP, which do not have a point mass, cannot maintain the confidence feature.

Recently, another deficiency in probabilistic inference was raised, called the false confidence (Balch et al., 2019), which is advocated to be avoidable by using the consonant belief (CB). An additional consonant feature (Balch, 2012) is claimed to be necessary for the CD to overcome the false confidence (Balch,

2020; Denoeux and Li, 2018; Martin et al., 2021). However, is the consonant feature indeed necessary for the CD? We show that the CD does not have a false confidence problem for the proposition of interest. We further introduce the null belief theorem, which precludes the use of an additional consonant feature. In satellite conjunction problem, risk assessment of collision is often based on binary hypothesis testing procedures (Hejduk et al., 2019). Due to the null belief, the CB cannot be used for the testing procedure and performs uniformly worse than the confidence-based decision making. GFD and RP also produce small collision probability under poor data to make a wrong decision when the satellite collides.

For the CD to become a big hit, it should have an advantage over the other existing methods. In Stein's problem and the satellite conjunction problem, the presence of a point mass allows the CD to avoid probability dilution and false confidence, which lead to wrong decision making. It is worth emphasizing that all these advantages are stemming from the fact that the CD is an extended likelihood, not probability.

3.2 Ambiguity in confidence of an observed CI

We first show that an observed CI can have an ambiguity in its confidence level (coverage probability) as a frequentist CI procedure. Stein's (1959) problem is important to show a probability dilution of the CI in high-dimensional cases. Suppose that $Y_1, ..., Y_k$ are independent with $Y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, ..., k$ and the parameter of interest is $\theta = ||\mu|| = \sqrt{\sum_{i=1}^{k} \mu_i^2}$. When $\theta = 0$, $\sum_{i=1}^{k} Y_i^2$ follows the Chi-square distribution and when $\theta > 0$ it follows the non-central

Chi-square distribution. Stein (1959) showed a probability dilution of the marginal posterior when $\theta \ll k$. Recently, the satellite conjunction problem is of interest, which can be reduced to the two-dimensional problem. Following Cunen et al. (2020) and Martin et al. (2021), let (Y_1, Y_2) be the measurements of the true difference between two satellites' positions (μ_1, μ_2) along each axis. Suppose that $Y_1 \sim N(\mu_1, \sigma^2)$ and $Y_2 \sim N(\mu_2, \sigma^2)$ are independent and σ is known. The parameter of interest is the Euclidean distance between the two satellites $\theta = \sqrt{\mu_1^2 + \mu_2^2}$. The measurement of θ is denoted by $D = \sqrt{Y_1^2 + Y_2^2}$. To distinguish the random variables Y_1, Y_2 and D from their observed values, the latter are written in lower cases y_1, y_2 and d. In satellite problem, the probability dilution is becoming apparent as $\sigma \to \infty$. In Stein's problem, the probability dilution becomes severe as $p \to \infty$. In this chapter, for simplicity of discussion, we mainly use the satellite problem ($k = 2$) for derivations, but our results can be applied to the Stein's problem.

Suppose that we want to make a frequentist CI based on the statistics D,

$$\frac{D^2}{\sigma^2} \sim \chi_2^2\left(\frac{\theta^2}{\sigma^2}\right), \tag{3.1}$$

where $\chi_{df}^2(\nu)$ denotes the non-central Chi-square distribution with the degrees of freedom df and non-centrality parameter ν. Suppose that $q_\alpha(\theta)$ is the $(1-\alpha)$ quantile function of D such that

$$P_\theta(D \leq q_\alpha(\theta)) = 1 - \alpha,$$

where $\alpha \in (0, 1)$. Since $q_\alpha(\theta)$ is a strictly increasing function of θ for any

$\alpha \in (0,1)$, we can consider an inverse function $q_\alpha^{-1}(d)$ such that

$$P_{\theta=q_\alpha^{-1}(d)}(D \le d) = P_\theta(q_\alpha^{-1}(D) \le \theta) = 1-\alpha.$$

However, the range of $q_\alpha(\theta)$ is $[q_\alpha(0), \infty)$ where $q_\alpha(0) > 0$ is the $(1-\alpha)$ quantile of the central Chi-square distribution. Thus, $q_\alpha^{-1}(d)$ is not defined for $d < q_\alpha(0)$. We let $q_\alpha^{-1}(d) = 0$ for such d. For $\alpha = 0$ and 1, let $q_0^{-1}(d) = \lim_{\alpha\to 0} q_\alpha^{-1}(d) = 0$ and $q_1^{-1}(d) = \lim_{\alpha\to 1} q_\alpha^{-1}(d) = \infty$.

Now we try to make a frequentist CI procedure with the α-confidence level as follows:

$$\mathrm{CI}_\alpha(D) = [\theta_L(D), \theta_U(D)), \tag{3.2}$$

where $\theta_L(D) = q^{-1}_{1-\alpha-\beta}(D)$ and $\theta_U(D) = q^{-1}_{1-\beta}(D)$ for some $0 \le \beta \le 1-\alpha$. For example, if $\alpha = 0.9$ and $\beta = 0.05$, $\theta_L(D) = q^{-1}_{0.05}(D)$ and $\theta_U(D) = q^{-1}_{0.95}(D)$ with the confidence level $\alpha = 0.95 - 0.05 = 0.9$. Then, the coverage probability of the CI procedure (3.2) is confidence level α, for all $\theta \in \Theta$

$$\begin{aligned} P_\theta(\theta \in \mathrm{CI}_\alpha(D)) &= P_\theta(\theta_L(D) \le \theta \le \theta_U(D)) \\ &= P_\theta(\theta_L(D) \le \theta) - P_\theta(\theta_U(D) \le \theta) \\ &= \alpha + \beta - \beta = \alpha. \end{aligned}$$

Note that $P_{\theta=0}(0 \in \mathrm{CI}_\alpha(D)) = P_{\theta=0}(\theta_L(D) = 0 < \theta_U(D)) = \alpha$. Thus, the CI procedure (3.2) with the level α always provides correct coverage probability. When $\beta = 0$ and $\beta = 1-\alpha$, the CI becomes the one-sided interval $[\theta_L(D), \infty)$ and $[0, \theta_U(D))$, respectively. When $\beta \in (0, 1-\alpha)$ we have two-sided CI. Given the observed data $D = d$, the two-sided CI procedure (3.2) with $0 < \beta < 1-\alpha$

leads to the following observed interval $\mathrm{CI}_\alpha(d) = [\theta_L(d), \theta_U(d))$:

(a) If $d > q_{1-\alpha-\beta}(0)$, the observed CI becomes two-sided interval $[\theta_L(d), \theta_U(d))$.

(b) If $q_{1-\beta}(0) < d \leq q_{1-\alpha-\beta}(0)$, the observed CI becomes one-sided interval $[0, \theta_U(d))$.

(c) If $d \leq q_{1-\beta}(0)$, the observed CI becomes an empty interval $[0, 0)$.

Thus, to maintain the coverage probability the two-sided CI procedure can yield an one-sided or an empty observed CI.

Let $\sigma = 1$ and $\beta = 0.05$. Figure 3.1 illustrates three CIs with level $\alpha = 0.95$, 0.90 and 0.60, respectively. When $\alpha = 0.95$ we have a one-sided CI procedure with $\theta_L(D) = 0$ and when $\alpha < 0.95$ we have a two-sided CI procedure to give an observed CI $[\theta_L(d), \theta_U(d))$. Three CIs have the common upper bound $\theta_U(d) = q^{-1}_{1-\beta}(d) = q^{-1}_{0.95}(d)$. In Figures, the horizontal and vertical axes represent d and θ, respectively. The dashed lines and the solid lines are $\theta_U(d)$ and $\theta_L(d)$, respectively. If $\alpha = 0.90$ (0.60), the two-sided CI procedure gives two-sided observed CI when $d > 2.448$ ($d > 1.449$). When $d \leq 0.320$, the three CI procedures give empty intervals. In figures, the horizontal arrows show the area $A = \{\ d : \theta_0 = 1 \in \mathrm{CI}(d)\ \}$ where θ_0 is the true value of θ. Thus, if $d \in A$, the observed CI contains the true parameter value $\theta_0 = 1$. Furthermore, $P(A) = P(\theta_0 \in \mathrm{CI}(D)) = \alpha$ implies that these three CI procedures have the correct coverage probabilities. The vertical arrows show the observed CIs at $d = 1$, 2, and 3. If $d = 2$, an observed CI $[0, 3.451)$ can be a realization of either 95% or 90% CI procedure. Moreover, if $d = 1$, an observed CI $[0, 2.287)$ can be a realization of 60%, 90% or 95% CI procedure. Thus, given an observed

CI, its coverage probability (confidence level) may not be uniquely determined. What is a meaning of the confidence for an observed CI?

3.3 GFD and probability dilution

Fisher (1930) used a sufficient statistic to construct the CD. Since y_1 and y_2 are sufficient statistics for (μ_1, μ_2), a joint CD for (μ_1, μ_2) can be obtained:

$$C_f(\mu_1, \mu_2; y_1, y_2) = P_{\mu_1,\mu_2}(Y_1 \geq y_1 \text{ and } Y_2 \geq y_2).$$

This leads to a joint confidence density,

$$\begin{aligned} c_f(\mu_1, \mu_2; y_1, y_2) &= \frac{\partial^2 C_f(\mu_1, \mu_2; y_1, y_2)}{\partial \mu_1 \partial \mu_2} \\ &= \frac{1}{\sigma^2}\phi\left(\frac{\mu_1 - y_1}{\sigma}\right)\phi\left(\frac{\mu_2 - y_2}{\sigma}\right) = L(\mu_1, \mu_2; y_1, y_2), \end{aligned}$$

where $\phi(\cdot)$ is the density function of $N(0,1)$ and $L(\mu_1, \mu_2; y_1, y_2)$ is the likelihood. Thus, it is also a joint posterior under uniform prior for (μ_1, μ_2). Then, the integrated CD (or marginal posterior) for θ is obtained

$$G(\theta; d) = \int_{\mu_1^2+\mu_2^2 \leq \theta^2} c_f(\mu_1, \mu_2; y_1, y_2)\ d(\mu_1, \mu_2) = \Gamma_2\left(\frac{\theta^2}{\sigma^2}; \frac{d^2}{\sigma^2}\right), \tag{3.3}$$

which gives the density

$$g(\theta; d) = \frac{\partial}{\partial \theta} G(\theta; d) = \frac{2\theta}{\sigma^2}\ \gamma_2\left(\frac{\theta^2}{\sigma^2}; \frac{d^2}{\sigma^2}\right), \tag{3.4}$$

where $\gamma_2(\theta; \cdot) = d\Gamma_2(\theta; \cdot)/d\theta$.

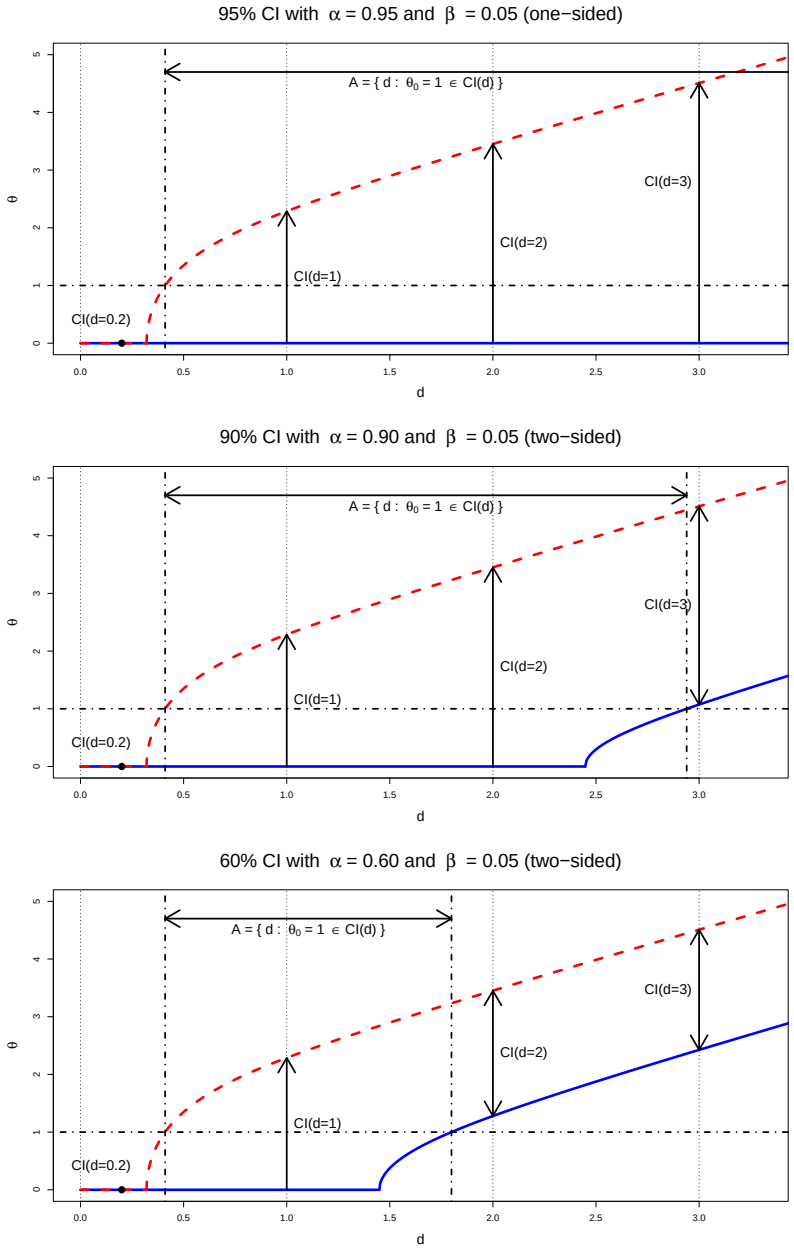


Figure 3.1: Confidence intervals with (a) $\alpha = 0.95$, (b) $\alpha = 0.90$, (c) $\alpha = 0.60$. For all three CIs, $\beta = 0.05$.

Hannig (2009) introduced the GFD as a generalization of the CD. We first show that the integrated CD is also a GFD of Hannig (2009). Consider a data generating mechanism

$$(Y_1, Y_2) = (\theta \cos\theta + \sigma U_1, \theta \sin\theta + \sigma U_2),$$

where U_1 and U_2 are independent random variables from $N(0,1)$. We can define the set-valued function as

$$\begin{aligned} Q(y_1, y_2, U_1^*, U_2^*) &= \{\ \theta : (y_1, y_2) = (\theta \cos\theta + \sigma U_1^*, \theta \sin\theta + \sigma U_2^*)\ \} \\ &= \sigma \sqrt{\left(\frac{y_1}{\sigma} - U_1^*\right)^2 + \left(\frac{y_2}{\sigma} - U_2^*\right)^2} \sim \sigma \sqrt{\chi_2^2\left(\frac{d^2}{\sigma^2}\right)}, \end{aligned}$$

then following Hannig (2009), $G(\theta; d)$ satisfies the definition of GFD.

In the satellite conjunction problem, we see that integrated CD, marginal posterior under uniform prior, and this GFD are equivalent. We denote the distribution (3.3) by GFD for notational convenience, but we clearly notice that GFD is not unique; for example, the marginal CD can also be expressed as a GFD. Stein (1959) noted its probability dilution as $k \to \infty$. In consequence, Wilkinson (1977) and Pedersen (1978) showed that the GFD cannot maintain the correct coverage probability. All these problems remain to hold, we show, in satellite conjunction problem with $k = 2$. To resolve the probability dilution, Wilkinson (1977) proposed the use of the marginal CD and Bernardo (1979) proposed the RP in the next section.

3.4 CD and related methods

Let (θ, ψ) and (D, T) be the polar coordinate representations of (μ_1, μ_2) and (Y_1, Y_2),

$$(\mu_1, \mu_2) = (\theta \cos\psi, \theta \sin\psi) \quad \text{and} \quad (Y_1, Y_2) = (D \cos T, D \sin T),$$

where $\psi = \tan^{-1}(\mu_2/\mu_1)$ and $T = \tan^{-1}(Y_2/Y_1)$. Here, the distributions of T and $T|D$ still depend on both θ and ψ; hence, D alone is not a sufficient statistic for θ under the full data (y_1, y_2). If we have a maximal ancillary statistics we may derive the CD with full data likelihood based on the distribution of $D|A$. But, in Stein's problem and satellite conjunction problem, the maximal ancillary statistics are not known. However, the current definition of the CD (Schweder and Hjort, 2016) only requires that

$$C(\theta_0; D) \sim \text{Uniform}[0, 1] \tag{3.5}$$

at the true value θ_0 of θ, which guarantees that the CD to maintain the correct coverage probability. Using (3.1), Cunen et al. (2020) derived the marginal CD for θ based on the statistics D,

$$C(\theta; d) = P_\theta(D \geq d) = 1 - \Gamma_2\left(\frac{d^2}{\sigma^2}; \frac{\theta^2}{\sigma^2}\right), \tag{3.6}$$

where $\Gamma_2(\cdot; \nu)$ denotes the non-central Chi-square distribution function, and they showed that this CD does not have probability dilution. This is equivalent to Wilkinson's (1977) marginal CD for Stein's problem with $k > 2$. This marginal CD satisfies the current definition (3.5) of the CD, so in this chapter

we call it the CD. However, it has been considered as paradoxical due to the point mass. Martin et al. (2021) insisted that the CD is at risk of false confidence. In particular, as is well known, CDs only guarantees reliable inferences on one-sided CIs. Otherwise, including two-sided CIs and their complements, are still subject to the false confidence phenomenon. We can view this as the current status of the CD. In this chapter, we show that the above mentioned properties are not drawbacks, but rather they are indeed advantages of the CD.

With a slight abuse of notation, we may define

$$C(A) = C(A; d) = C(\theta \in A) = \int_A c(\theta; d) d\theta.$$

Then, the CD has a point mass at $\theta = 0$, since

$$C(\{0\}) = 1 - \Gamma_2\left(\frac{d^2}{\sigma^2}; 0\right) > 0.$$

Let Θ be the parameter space of θ and Ω_D be the sample space of D. If we assume $\Theta = (0, \infty)$, the point mass at zero becomes an unassigned probability and the CD is not a probability to have $C(\Theta) < 1$, as Wilkinson (1977) noted. This makes the point mass look paradoxical (Schweder and Hjort, 2016). However, if we assume $\Theta = [0, \infty)$, then $C(\Theta) = 1$ and this CD satisfies the confidence property,

$$C(\theta_0; D) = C([0, \theta_0]; D) = 1 - \Gamma_2\left(\frac{D^2}{\sigma^2}; \frac{\theta_0^2}{\sigma^2}\right) \sim \text{Uniform}[0, 1],$$

to give correct coverage probability for any true value $\theta_0 \in \Theta$.

Let $M(D) = C(\{0\}; D)$ denote the point mass at $\theta = 0$ and $M(d)$ denote its realized value of the point mass. As $\sigma \to 0$ or $\theta \to \infty$, the point mass $M(D)$ vanishes

$$M(D) \xrightarrow{p} 1 - \Gamma_2(\infty; 0) = 0.$$

As $\sigma \to \infty$ or $\theta \to 0$,

$$M(D) \xrightarrow{d} \text{Uniform}[0, 1].$$

Here the confidence density can be expressed as

$$c(\theta; d) = M(d) \cdot (\theta) + c_+(\theta; d),$$

where (θ) denotes the Dirac delta function to give a point mass at $\theta = 0$ and $c_+(\theta; d) = \partial P_\theta(D \geq d)/\partial\theta$.

3.4.1 CD and confidence level of an observed CI

A point mass at a boundary prevents the probability dilution to maintain the confidence feature. We first investigate a necessary and sufficient condition for a point mass in the CD. Suppose that $\theta \in \Theta$ is the parameter of continuous scalar statistic $D \in \Omega_D$ and the $1 - \alpha$ quantile $q_\alpha(\theta)$ is a strictly increasing function of θ for any for any $\alpha \in (0, 1)$. Then, we have the following theorem with a proof in Appendix 3.8.1.

Theorem 3.1. *Let $\partial\Omega_D$ and $\partial\Theta$ denote the boundary of Ω_D and Θ, respectively. Then, $C(\theta; d)$ has no point mass if and only if*

$$q_\alpha(\theta) \to \partial\Omega_D \quad \text{as } \theta \to \partial\Theta, \quad \forall\alpha \in (0, 1). \tag{3.7}$$

Pawitan et al. (2023) considered a curved exponential model. Let $y = 1$ be an observation from $Y \sim N(\theta, \theta^2)$ for $\theta \geq 0$, then one may consider a confidence distribution,

$$C(\theta; y) = P_\theta(Y \geq y) = 1 - \Phi\left(\frac{y-\theta}{\theta}\right),$$

where $\Phi(\cdot)$ denotes the cumulative function of $N(0,1)$. However, this leads to

$$\lim_{\theta\to\infty} C(\theta; y) = 1 - \Phi(-1) \approx 0.84 < 1.$$

Here $C(\{0\}; y = 1) = 0$. Thus, there is no point mass at $\theta = 0$. According to Wilkinson (1977), this CD has an unassigned probability $0.16 = 1 - 0.84$. This problem occurs because the quantile function $q_\alpha(\theta)$ is not increasing function of θ. Now let $d = |y|$ be an observation of $D = |Y|$ with $\Omega_D = \Theta$. Then the corresponding CD is defined as

$$C(\theta; d) = P_\theta(D \geq d) = 1 - \Phi\left(\frac{d-\theta}{\theta}\right) + \Phi\left(\frac{-d-\theta}{\theta}\right),$$

which becomes a proper distribution function without a point mass

$$\lim_{\theta\to 0} C(\theta; d) = 1 - \Phi(\infty) + \Phi(-\infty) = 0$$
$$\lim_{\theta\to\infty} C(\theta; d) = 1 - \Phi(-1) + \Phi(-1) = 1.$$

When there is no point mass, under appropriate conditions, Pawitan et al. (2023) showed that

$$C(\theta \in CI(d)) = \int_{CI(d)} c(\theta; d)d\theta = P_\theta(\theta \in CI(D)),$$

where the LHS is the confidence of the observed CI and the RHS is the coverage probability of the CI procedure. Thus, the confidence of the observed CI corresponds to the coverage probability of the unique CI procedure.

In satellite conjunction problem and Stein's problem, lower bounds of Ω_D and Θ are zero but $q_\alpha(0) \neq 0$. Thus, Theorem 3.1 implies that the corresponding CD has a point mass at zero. Now we extend the relationship between coverage probability and confidence with the presence of a point mass. The confidence of the observed $\text{CI}_\alpha(d)$ is as follows, corresponding to each case of Section 3.2.

(a) When the observed CI is two-sided, $\theta_L(d) > 0$,

$$C(\text{CI}_\alpha(d)) = C(\theta < \theta_U(d); d) - C(\theta < \theta_L(d); d) = (1-\beta)-(1-\alpha-\beta) = \alpha.$$

(b) When the observed CI becomes one-sided, $\theta_L(d) = 0$ and $\theta_U(d) > 0$),

$$\begin{aligned} C(\text{CI}_\alpha(d)) &= C(\theta < \theta_U(d); d) = 1 - \beta \\ &= \max\{\ \alpha : \text{CI}_\alpha(d) = [0, \theta_U(d)) \text{ for some } 0 \leq \beta \leq 1 - \alpha\ \}, \end{aligned}$$

which is the maximum coverage probability (confidence level) among CI procedures having the observed CI $[0, \theta_U(d))$. For example, in Figure 3.1, both $\alpha = 0.95$ and $\alpha = 0.9$ lead to the same observed CI $[0, 3.451)$

for $d = 2$. Here, the CD gives a confidence

$$C([0, 3.451); d = 2) = 1 - \beta = 0.95.$$

(c) When the observed CI becomes an empty set, $\theta_L(d) = \theta_U(d) = 0$,

$$M(d) = C(\{0\}; d) = 1 - \Gamma_2\left(\frac{d^2}{\sigma^2}; 0\right) = \max\{\ \alpha : \mathrm{CI}_\alpha(d) = \emptyset\ \},$$

which is the maximum confidence level among CI procedures, having an empty interval. The point mass leads to a nice interpretation. In Figure 3.1, all the three procedures lead to $\mathrm{CI}_\alpha(d = 0.2) = \emptyset$. Here, the point mass of CD is

$$M(d = 0.2) = C(\{0\}; d) = 0.980,$$

which implies that the CI procedure produces an empty observed CI if $\alpha < 0.98$. Thus, given $d = 0.2$, we can allow the CI only with $\alpha \geq 0.98$. Here the confidence interval $[0, 0] = \{0\}$ has the confidence level 0.98.

Given an observed CI, its CD gives the maximum attainable coverage probability (confidence level) among CI procedures,

$$\begin{aligned} C(\theta \in CI(d)) &= \max P_\theta(\theta \in CI(D)) \\ &= \max\{\ \alpha : \mathrm{CI}_\alpha(d) = CI(d) \text{ for some } 0 \leq \beta \leq 1 - \alpha\ \}. \end{aligned}$$

3.4.2 CD versus GFD

Martin et al. (2021) claimed that a drawback of CD is that it cannot have the two-sided CIs. We see that the CD allows a two-sided CI procedure,

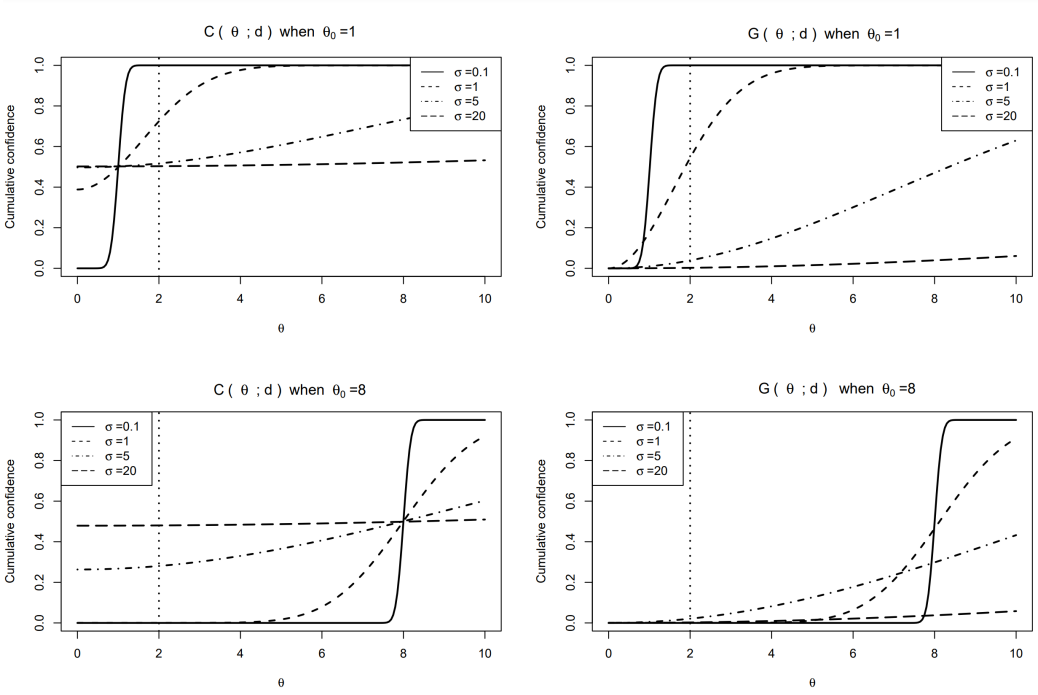


Figure 3.2: Average of $C(\theta; d)$ and $G(\theta, d)$ over 10,000 repeats.

but it can provide a one-sided observed interval to maintain the confidence feature (coverage probability). We shall show that any procedure, which can always allow two-sided interval, cannot avoid the probability dilution. Figure 3.2 shows the averages of cumulative functions, based on CD and GFD from 10,000 repetitions, where θ_0 is 1 or 8 and σ varies in (0.1, 1, 5, 20). Both provide the cumulative distribution for θ. Compared with the CD, the GFD has apparent probability dilution. The CD and GFD become identical when $\sigma \to 0$. However, they can be quite different when σ is large. Since the CD has a point mass at zero, $C(\{0\}) > 0$, whereas the GFD does not: $G(\{0\}) = 0$, the GFD can always provide a two-sided interval, but it leads to probability dilution, losing the confidence feature as we shall see.

3.4.3 CD versus RP

Bernardo (1979) proposed the RP to resolve the probability dilution of the GFD (also the marginal posterior under uniform prior) especially when $\theta \ll k$. Figure 3.3 shows the coverage probabilities of the one-sided and two-sided 80% CIs for satellite conjunction problem ($k = 2$) and Stein problem ($k = 100$), computed from 10,000 repetitions. Probability dilution of GFD is evident, especially when $\theta \ll k$. RP improves GFD. Both RP and GFD are the probability on $\Theta = (0, 1)$, without a point mass at zero. This causes a probability dilution that both RP and GFD cannot maintain confidence feature at near zero. Note that the CD-based two-sided CIs automatically becomes the one-sided interval to maintain confidence feature at zero when the observation d is small. The figure shows that only the CD maintains the confidence feature for all $\theta \in \Theta$.

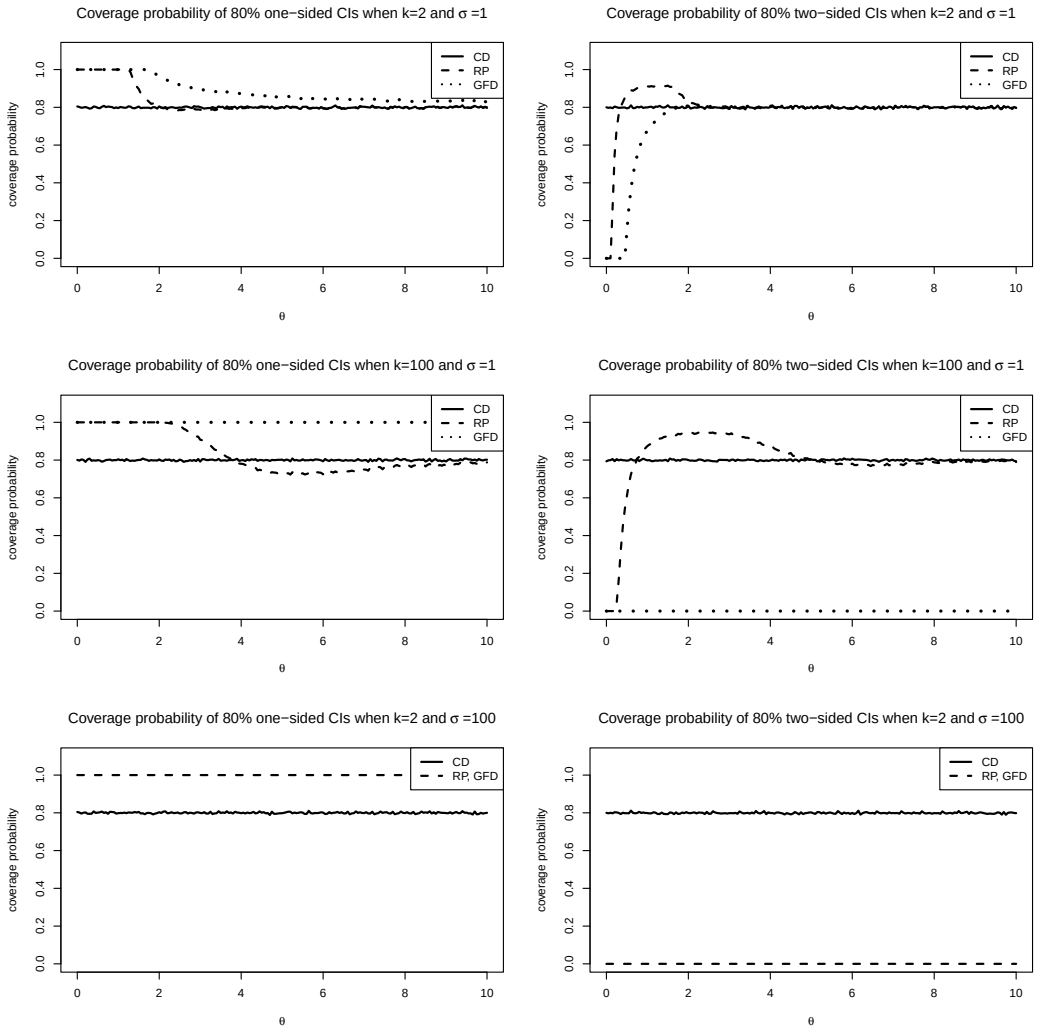


Figure 3.3: Coverage probabilities of 80% CIs based on CD, GFD, and RP when $k = 2$ and $k = 100$.

3.5 On false confidence

Pawitan and Lee (2021) showed that the CD is an extended likelihood, not a probability. However, Martin et al. (2021) noted another fundamental deficiency in probabilistic inference, namely the false confidence property and claimed that the CD also cannot avoid this deficiency. They introduced the false confidence theorem below.

False confidence theorem (Balch et al., 2019) *For any* $\theta_0 \in \Theta$, $\alpha \in (0,1)$, *and* $p \in (0,1)$, *there exists a subset* $A \subset \Theta$ *such that* $\theta_0 \notin A$ *and*

$$P_{\theta_0}\{C(A;D) \geq 1-\alpha\} \geq p. \tag{3.8}$$

This theorem avoids the existence of any false proposition. Existence of a false proposition with a high confidence is annoying if it is of interest. However, we may not need a protection from a meaningless false proposition. For example, suppose that θ_0 is a true value of θ. If we use the CD, GFD, and RP, a false proposition $A = \{\theta : \theta \neq 1\}$ satisfies (3.8). However, since $C(\{1\}) = 0$ such a false proposition A is meaningless when the true θ_0 is unknown. However, $\theta_0 = 0$ is an interesting proposition, since D follows the central Chi-square distribution if the proposition holds. For the CD (not GFD and RP), the false confidence theorem does not hold for $A \subseteq \{\theta : \theta \neq 0\}$, since a point mass occurs at zero. For any false proposition ,

$$C(A;D) \leq C(\theta \neq 0;D) = 1 - M(D) \sim \text{Uniform}[0,1],$$

which means that

$$P_{\theta_0=0}\{C(A;D) \geq 1-\alpha\} \leq P_{\theta_0=0}\{M(D) \leq \alpha\} = \alpha.$$

Thus, the false confidence theorem does not hold when $\theta_0 = 0$. Let R be the sum of the radii of two satellites and let H_0 (collision; $\theta \leq R$) be the true, and let H_1 (non-collision; $\theta > R$) be the false proposition A. Then, the level of false confidence becomes

$$P_{\theta_0}\{C(H_1;D) \geq 1-\alpha\} = P_{\theta_0}\{C(H_0;D) \leq \alpha\} \leq P_{\theta_0}\{C(\theta_0;D) \leq \alpha\} = \alpha.$$

Hence, if H_0 is true, the level of false confidence cannot grow arbitrarily large. Thus, assertion H_0 avoids false confidence because it includes $\theta_0 = 0$. This satisfies the Martin-Liu validity criterion (Martin and Liu, 2015) at least for H_1, the false proposition of interest. Thus, the CD does not allow a high false confidence for a false proposition of inferential interest.

To avoid the false confidence theorem, Martin et al. (2021) proposed the use of CB. Let Θ be the parameter space. Then the CB for a subset $A \subset \Theta$ is defined by

$$\text{Bel}(A;d) = 1 - \sup_{\theta \in A^c} \text{pls}(\theta;d), \tag{3.9}$$

where $\text{Bel}(\cdot;d)$ is the consonant belief function and $\text{pls}(\theta;d) = 1-|2C(\theta;d)-1|$ is the plausibility contour. This CB can avoid the false confidence of any false proposition. But this unnecessary through protection of CBs against any false protection make them vulnerable in other prospects of statistical inferences. In this chapter, we introduce the 'null belief theorem', which cannot be avoided

by any CBs, from the opposite point of view of the false confidence theorem.

Theorem 3.2 (Null belief theorem)**.** *Consider a CB Bel*$(\cdot; d)$ *characterized by either a CD or a probability. Then, for any true* $\theta_0 \in \Theta$ *and any* $p \in (0, 1)$*, there exists an interval* I *with positive length such that*

$$\theta_0 \in I \subset \Theta \quad \textit{and} \quad P_{\theta_0}\{\textit{Bel}(I; D) = 0\} \geq p.$$

Proof. First, take a small interval near the true θ_0. Let $\hat{\theta}(d)$ be the median of the CD such that $C\left(\hat{\theta}(d); d\right) = 0.5$. Then, for any $p \in (0, 1)$, there exists $\epsilon > 0$ such that

$$P_{\theta_0}\{\hat{\theta}(D) \in (\theta_0 - \epsilon, \theta_0 + \epsilon)\} \leq 1 - p.$$

Let $I = (\theta_0 - \epsilon, \theta_0 + \epsilon)$ be an interval that contains the true value θ_0 but is such that

$$P_{\theta_0}\{\text{Bel}(I; D) = 0\} = P_{\theta_0}\{\hat{\theta}(D) \notin I\} \geq p,$$

then the theorem is proved. □

Even if an interval I contains the true value θ_0, the probability of $\text{Bel}(I; D) = 0$ is greater than $p > 0$. Such a CB seems to be counter-intuitive. On the other hand, the CD and the probability including the RP and GFD, avoid the null belief theorem. The belief function can be useful for the trinary decision problem with presence of an additional plausibility function (Dempster, 1968), but the null belief theorem implies that the belief function alone could not be applied to give a valid hypothesis testing procedure, as we shall show later.

3.5.1 False confidence and probability dilution

Balch et al. (2019) claimed that the probability dilution is a symptom of the false confidence. They think that the CD is a probability like GFD and RP. We emphasize that the probability dilution should be distinguished with false confidence. Let $H_0 : \theta \leq R$ be an assertion of collision and H_1 : not H_0 be an assertion of non-collision. Here, $G(H_0) = G(\theta \leq R) = G([0, R])$ is the probability of collision based on the GFD. Probability dilution means that

$$G(H_0) \to 0 \quad \text{as } \sigma \to \infty.$$

This is counter-intuitive because lower-quality data paradoxically appear to dilute the risk of impending collision (Balch et al., 2019; Hejduk and Snow, 2019). This causes a severe and persistent underestimate of risk exposure. Based on the CD, Cunen et al. (2020) investigated the probability of collision $C(H_0)$ at the true $\theta_0 = 1.99 \approx 2 = R$, and $C(H_0) = C([0, 2]) \approx C([0, 1.99]) \sim$ Uniform$[0, 1]$ for any σ. Thus, in their simulation, the average probability of collision remained close to 0.5 for all σ. In general, the point mass is less than or equal to $C(\theta_0; d)$,

$$M(d) = C(0; d) \leq C(\theta_0; d) \sim \text{Uniform}[0, 1],$$

but when $\sigma \to \infty$ or $\theta_0 \to 0$, the point mass at $\theta = 0$ converges in distribution to Uniform$[0, 1]$,

$$M(D) = 1 - \Gamma_2\left(\frac{D^2}{\sigma^2}; 0\right) \xrightarrow{d} \text{Uniform}[0, 1].$$

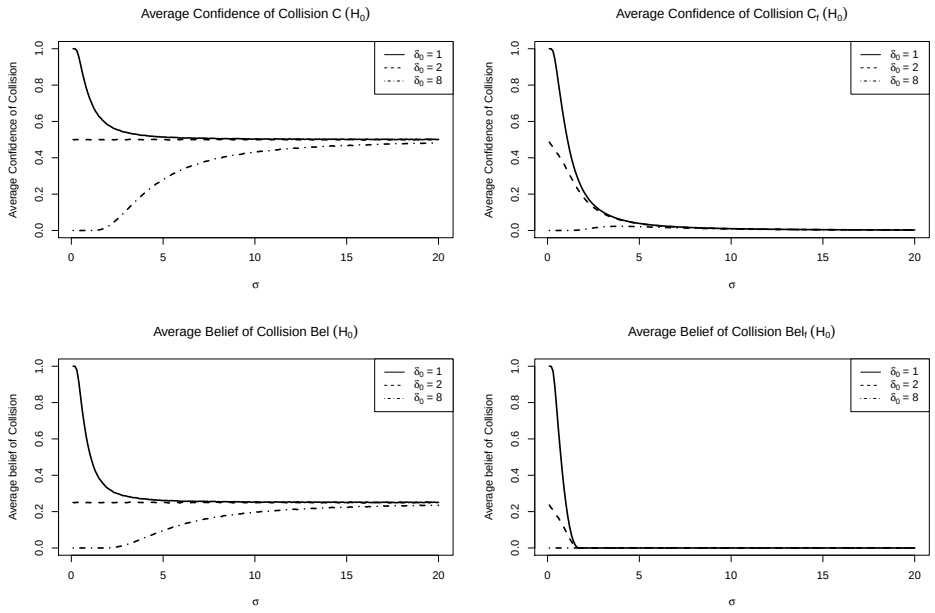


Figure 3.4: Average confidences and beliefs regarding collision over 100,000 repetitions.

Thus, as $\sigma \to \infty$, the average of $M(D)$ converges to 0.5. This prevents the probability dilution.

Figure 3.4 shows the average confidences and beliefs of collision as the uncertainty σ varies from 0 to 20. As σ increases, $C(H_0)$ decreases to 0.5 when $\theta_0 = 1$, and $C(H_0)$ increases to 0.5 when $\theta_0 = 8$. We see that these phenomena are caused by a point mass at zero: $C(\{0\}) > 0$. $\text{Bel}(H_0)$ is the CB (3.9) based on the CD, $C(\theta; d)$. $\text{Bel}(H_0)$ converges to 0.223 as $\sigma \to \infty$. We see that $\text{Bel}(H_0) < C(H_0)$, i.e., the additional consonant feature in CB leads to the loss of the confidence feature. The GFD has no point mass, so $G(H_0)$ goes to zero as $\sigma \to \infty$. $\text{Bel}_G(\cdot)$ is the CB function (3.9) based on the GFD. $\text{Bel}_f([0, \theta])$ has no point mass, so it also suffers from a probability dilution.

Thus, it is the confidence feature, not the consonant feature, that prevents a severe and persistent underestimate of risk for satellite collision.

3.6 Hypothesis testing

Hypothesis testing procedures are often used in risk assessment for satellite conjunction problem. Depending on certain fundamental questions, the null hypothesis H_0 can be either $\theta \leq R$ (collision) or $\theta > R$ (non-collision). The probability (confidence) of collision is the most frequently used test statistic in satellite conjunction problem (Hejduk and Snow, 2019). For illustration, we suppose that H_0 is the assertion of collision. Then, from the property of CD,

$$C(\theta \leq \theta_0; D) \sim \text{Uniform}[0, 1],$$

the confidence $C(H_0; d)$ directly becomes the observed p-value for testing H_0, i.e.,

$$\max_{\theta \in H_0} P_\theta \left(C(H_0; D) \leq \alpha \right) = \alpha.$$

Thus, the CD yields α-level hypothesis testing procedure for any σ. However, since probabilities such as the GFD and RP have no point mass, as $\sigma \to \infty$

$$G(H_0; D) \to 0 \quad \text{and} \quad R(H_0; D) \to 0.$$

Thus, if the data are of a poor quality (σ is large), the GFD and RP cannot accept the null hypothesis even if $d < R$. For example, suppose that we observe $d = 1 < R = 2$. When $\sigma = 1$, the CD, GFD and RP give $C(H_0) = 0.918$, $G(H_0) = 0.731$ and $R(H_0) = 0.891$, respectively. Thus, all of them would

not reject H_0. Here the RP becomes close to the CD. However, in satellite conjunction problem, σ is often much greater than R (Balch et al., 2019). When $\sigma = 100$, the CD yields $C(H_0) = 1.000$, hence the CD would not reject H_0. However, $G(H_0) = 0.000$ and $R(H_0) = 0.016$ to reject H_0 though the observed value implies the collision ($d < R$). Therefore, if the collision probability is given by the CD, there is no reason for engineers to ignore an impending collision risk due to the negligible probability of collision. However, if it is based on the GFD or RP, engineers may ignore the impending danger because of the dilution of collision probability $G(H_0)$ and $R(H_0)$. In Stein's problem it is often of interest to test

$$H_0 : \ \theta = 0 \quad \text{vs.} \quad H_1 : \ \theta \neq 0.$$

Due to the point mass at $\theta = 0$, the CD gives

$$P_{\theta \in H_0}(C(H_0; D) < \alpha) = P_{\theta=0}(C(0; D) < \alpha) = \alpha.$$

Thus, if we use $C(H_0)$ as a p-value, we can directly achieve a valid hypothesis testing procedure with

$$P_{H_0}(\text{Reject } H_0) = P_{\theta=0}(C(H_0) < \alpha) = P_{\theta=0}(M(D) < \alpha) = \alpha.$$

On the other hand, the GFD and RP have no point mass. Thus, $G(H_0)$ and $R(H_0)$ do not lead to a valid hypothesis testing, because $G(H_0) = R(H_0) = 0$ for any observation d.

Now, suppose that we use the CB in satellite conjunction problem, $\text{Bel}(H_0) =$

$\text{Bel}([0, R])$. Note here that the CB often becomes zero, due to the null belief theorem. When true $\theta_0 = R$ (collision), we have

$$P_{\theta_0=R}\{\text{Bel}(H_0) = 0\} = P_{\theta_0=R}\{C(R; D) \leq 0.5\} = 0.5.$$

Suppose that the CB is used for testing by rejecting H_0 if $\{\text{Bel}(H_0) \leq \alpha\}$. Then,

$$P_{\theta_0=R}(\text{Reject } H_0) = P_{\theta_0=R}\{\text{Bel}(H_0) \leq \alpha\} \geq P_{\theta_0=R}\{\text{Bel}(H_0) = 0\} = 0.5.$$

Thus, though H_0 (collision) is true, the CB rejects H_0 with probability 0.5. It implies that the CB cannot achieve the significance level under 0.5. Similarly, $\text{Bel}_G(H_0)$ cannot achieve the significance level under 0.847. Thus, the CBs cannot be applied to hypothesis testing for risk assessment.

When $\sigma = \infty$, the data (y_1, y_2) are meaningless as an estimate of (μ_1, μ_2). However, poor data cannot justify the small collision probability $G(H_0) \approx 0$ under impending collision situations. The low collision probability ($G(H_0)$ or $R(H_0)$) does not mean that the two satellites are far apart; it is only a statement of the general unlikelihood of such an alignment if all one knows is that the two satellites happen to be in the same general area (Hejduk et al., 2019). However, it is undesirable for engineers to ignore impending danger because they believe a negligible collision probability caused by poor data. Since $C(H_0) = 1 - C(H_1) \xrightarrow{d} \text{Uniform}[0, 1]$ as $\sigma \to \infty$, the CD always acknowledges a non-negligible collision probability even with poor data. In this respect, the CD is useful for lowering the risk in satellite conjunction problem.

3.7 Concluding remarks

Neyman's confidence provides an objective frequentist interpretation to the CI procedure, whereas the CD offers an epistemic interpretation for an observed CI. Under appropriate conditions, they are equivalent, hence these two concepts are complementary. The confidence of an observed CI can be understood as the coverage probability of repeated experiments from a frequentist perspective. After observing the data, Neyman's observed CI attains a meaningful epistemic interpretation, especially when its coverage probability is uncertain. We demonstrate that consonant feature itself cannot prevent probability dilution. It is the confidence feature that is key to avoiding severe and persistent underestimation of risk exposure. The presence of a point mass in the CD allows the maintenance of the confidence feature. However, probabilities such as GFD and RP lose the confidence feature near the origin. The CD in (3.6) is based on the marginal distribution of a non-sufficient statistic D, which may not fully exploit all the information in the data. Since the CD is an extended likelihood, the integrated CD is no longer a CD. If a maximal ancillary exists, the CD for θ can be derived using the conditional distribution without information loss, allowing no relevant subset (Pawitan et al., 2023). The pursuit of a CD with a full data likelihood presents an interesting avenue for future research. Many properties of the CD still require further investigation and understanding.

3.8 Appendix

3.8.1 Proof of Theorem 3.1

Let D_U and D_L be the upper and lower bounds of Ω_D, and let θ_U and θ_L be the upper and lower bounds of Θ. Since the quantile function is continuous, we write $q_\alpha(\theta_L) = \lim_{\theta \to \theta_L} q_\alpha(\theta)$ and $q_\alpha(\theta_U) = \lim_{\theta \to \theta_U} q_\alpha(\theta)$. Note here that we allow the bounds to be $\pm\infty$.

($\Rightarrow$) Suppose that there exists $0 \leq \alpha \leq 1$ such that

$$q_\alpha(\theta_L) \neq D_L \quad \text{or} \quad q_\alpha(\theta_U) \neq D_U.$$

If $q_\alpha(\theta_L) \neq D_L$, there exists d^* such that $D_L < d^* < q_\alpha(\theta_L)$. Since $C(\theta; d^*)$ does not have a point mass,

$$C(\theta_L; d^*) = P_{\theta_L}(D \geq d^*) = 0.$$

However, by definition of the quantile,

$$C(\theta_L; d^*) > C(\theta_L; q_\alpha(\theta_L)) = P_{\theta_L}(D \geq q_\alpha(\theta_L)) = \alpha > 0.$$

This leads to contradiction. If $q_\alpha(\theta_U) \neq D_U$, there exists d^* such that $q_\alpha(\theta_U) < d^* < D_U$. Similarly,

$$C(\theta_U; d^*) = P_{\theta_U}(D \geq d^*) = 1,$$

but we have

$$C(\theta_U; d^*) < C(\theta_U; q_\alpha(\theta_U)) = P_{\theta_U}(D \geq q_\alpha(\theta_U)) = \alpha < 1,$$

which leads to contradiction. Thus, $q_\alpha(\theta) \to \partial\Omega_D$ as $\theta \to \partial\Theta$ for all $\alpha \in (0, 1)$.
($\Leftarrow$) Let d be an arbitrary value in Ω_D, then (3.7) leads to

$$C(\theta_L; d) = P_{\theta_L}(D \geq d) \geq P_{\theta_L}(D \geq q_\alpha(\theta_L)) = \alpha.$$

Taking $\alpha = 0$ leads to $C(\theta_L; d) = 0$. Similarly we can obtain $C(\theta_U; d) = 1$. Thus, $C(\theta; d)$ does not have a point mass for any $d \in \Omega_D$. □

Chapter 4

Foundations of h-likelihood inference for random unknowns

4.1 Introduction

Fisher (1922) introduced the classical likelihood for statistical models with fixed unknowns (parameters), whose maximum likelihood estimators (MLEs) are asymptotically the best, achieving the Cramer-Rao lower bound. Furthermore, its information matrix and associated delta-method provide a necessary estimator with a standard error estimator for any function of parameters of interest. This makes the maximum likelihood procedure is popular and widely used in practice. It necessitates an extension of the Fisher likelihood to general models with additional random unknowns (unobservables; Henderson et al., 1959). In the statistical literature unobservables appear with various names such as random effects, latent processes, factor, missing data, unobserved future observations, potential outcomes etc. Thus, it is of interest to have a

proper extension of the classical likelihood to give asymptotically the best predictors for random unknowns, in addition to MLEs for fixed unknowns. However, despite many attempts, it has yet remained unsuccessful. Bayarri et al. (1988) demonstrated difficulties by showing that maximization of all the existing extended likelihoods cannot give sensible estimators for both fixed and random parameters. Lee and Nelder (1996) proposed the hierarchical (h-)likelihood, defined on a particular scale of random parameters. Their aim was that maximum h-likelihood estimators (MHLEs) give MLEs for fixed parameters and at the same time asymptotically best unbiased predictors (BUPs) for random parameters. Their h-likelihood is an extension of Henderson's (1959) joint density for normal linear mixed models to hierarchical generalized linear models (HGLMs). However, in general MHLEs cannot provide MLEs. Even in linear mixed models, MHLEs cannot provide MLEs for the variance components. Thus, the Laplace approximation (LA) has been advocated to obtain an approximate MLEs (Lee and Nelder, 2001). Little and Rubin (2019) described the current status of the h-likelihood approach as,

> "Unlike maximization of the classical likelihood of Fisher (1922), maximization of an extended likelihood does not generally give consistent estimates of the parameters (Breslow and Lin, 1995). .... Lee and Nelder (2001) and Lee et al. (2006) proposed maximizing a 'modification' .... which is the correct ML approach."

However, the modification such as the LA can give severely biased estimation (Shun and McCullagh, 1995), especially in binary data. Meng (2009) noted another difficulty to have an asymptotic theory for prediction of missing data, as we shall discuss later.

In this chapter, we derive the new h-likelihood whose MHLEs are MLEs for fixed parameters and at the same time asymptotically best unbiased predictors (BUPs) for random parameters, achieving the generalized Cramer-Rao lower bound. Firth (2006) noted the ambiguity in forming the h-likelihood of Lee and Nelder (1996) and Meng (2009) noted a difficulty of a consistent prediction of unobservables in missing data problems. This explains why the expectation and maximization (EM) algorithm (Dempster et al., 1977) does not pay much attention to the prediction of unobservables. In this chapter, we show how the reformulated h-likelihood overcomes the difficulties raised by Firth (2006) and Meng (2009). We show that maximum h-likelihood estimators provide asymptotically the best estimation and prediction for fixed parameters and random parameters, respectively. Then, we further show how the h-likelihood theories can be applied when either an estimator for fixed parameter or a predictor of random parameter is not consistent. We also discuss how to obtain the h-likelihood when it is not explicitly expressed.

4.2 Hierarchical likelihood

Suppose that we have a statistical model $f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{u})$, composed with fixed unknowns $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$, random unknowns $\mathbf{u} = (u_1, ..., u_m)^T$ and observed data $\mathbf{y} = (y_1, ..., y_N)^T$. For the likelihood of both fixed and random parameters $(\boldsymbol{\theta}, \mathbf{u})$, the joint density of $(\mathbf{y}, \mathbf{u})$ has been considered,

$$L_e(\theta, \mathbf{u}; \mathbf{y}) \equiv f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{u}). \tag{4.1}$$

This agrees with the suggestion made by Henderson et al. (1959), Kaminsky and Rhodin (1985), Butler (1986), Berger and Wolpert (1988), Bjørnstad (1996), and Lee and Nelder (1996). The joint density of $\mathbf{y}$ and $\mathbf{u}$ is often from a completely specified hierarchical model,

$$f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{u}) = f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u}) f_{\boldsymbol{\theta}}(\mathbf{u}), \tag{4.2}$$

describing the data generation scheme. In normal linear mixed models, Henderson et al. (1959) proposed the use of estimators, maximizing this joint density with respect to fixed and random parameters. Joint maximization algorithms have been extended by a number of authors (Breslow and Clayton, 1993; Gilmour et al., 1985; Harville and Mee, 1984; Schall, 1991; Wolfinger, 1993) via different justifications. However, care is necessary in defining the joint density due to the Jacobian term associated with random parameters. This makes a prediction of random unknowns different from an estimation of fixed unknowns. It is Lee and Nelder (1996) to argue that a specific scale of random parameters should be used to form the joint density for MHLEs. In this chapter, we define the extended likelihood for statistical inferences by

$$L_e(\boldsymbol{\theta}, u; y) = L(\boldsymbol{\theta}; y) L_p(\mathbf{u}; \mathbf{y}), \tag{4.3}$$

where $L(\boldsymbol{\theta}; \mathbf{y}) \equiv f_{\boldsymbol{\theta}}(\mathbf{y}) = \int f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{u}) d\mathbf{u}$ is the classical likelihood (Fisher, 1922) and $L_p(\mathbf{u}; \mathbf{y}) \equiv f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u})$ is the predictive likelihood (Lee et al., 2017).

Example 4.1 (One-way random effect model). Consider a one-way ran-

dom effect model with a linear predictor

$$\eta_{ij} = \mu_{ij} = \mathrm{E}(y_{ij}|v_i) = \mu_0 + v_i, \quad \text{for} \quad i = 1, ..., n;\ j = 1, ..., m,$$

where $\mathbf{v} = (v_1, ..., v_n)^T \sim N(0, \lambda^2 \mathbf{I}_n)$, and $\mathbf{y}|\mathbf{v} \sim N(\mu_0 + \mathbf{v}, \sigma^2 \mathbf{I}_N)$, $N = mn$ is the sample size, $\mathbf{I}_N$ is an identity matrix of size N. For the simplicity of arguments, let the variance components $\sigma^2 = \lambda^2 = 1$. Then, the extended log-likelihood of $(\mu_0, \mathbf{v})$ is equivalent to Henderson et al.'s (1959) joint density,

$$\begin{aligned}\ell_e(\mu_0, \mathbf{v}) &= \log L_e(\mu_0, \mathbf{v}; \mathbf{y}) = \log f_{\mu_0}(\mathbf{y}, \mathbf{v}) \\ &= -\frac{n(m+1)}{2} \log 2\pi - \frac{1}{2} \sum_{i,j} (y_{ij} - \mu_0 - v_i)^2 - \frac{1}{2} \sum_i v_i^2,\end{aligned}$$

whose maximization over μ_0 and $\mathbf{v}$ gives the MLE $\widehat{\mu}_0 = \bar{y} = \sum_{i,j} y_{ij}/N$. However, if we re-parameterize $\mathbf{v}$ in terms of the log-normal distribution $u_i = \log v_i$, then we can set up another extended likelihood of μ_0 and $\mathbf{u} = (u_1, ..., u_n)^T$, given by

$$\begin{aligned}\ell_e(\mu_0, \mathbf{u}) &= \log L_e(\mu_0, \mathbf{u}; \mathbf{y}) = \log f_{\mu_0}(\mathbf{y}, \mathbf{u}) \\ &= -\frac{n(m+1)}{2} \log 2\pi - \frac{1}{2} \sum_{i,j} (y_{ij} - \mu_0 - v_i)^2 - \frac{1}{2} \sum_i v_i^2 - \sum_i v_i.\end{aligned}$$

The last term on the right-hand side is the Jacobian term. The two models in terms of $\mathbf{v}$ and $\mathbf{u}$ are of course equivalent, but now the joint maximization of $\ell_e(\mu_0, \mathbf{u})$ gives $\widehat{\mu}_0 = \bar{y} + 1$. Therefore, without a general principle to deal with this problem, the joint maximization concept would indeed be useless. □

Example 4.2 (Multiplicative Poisson random effect model). Consider

a multiplicative Poisson random effect model,

$$\mu_{ij} = \mathrm{E}(y_{ij}|u_i) = u_i \cdot \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})$$

to give a linear predictor

$$\eta_{ij} = \log \mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i,$$

where $y_{ij}|v_i \sim \mathrm{Poi}(\mu_{ij})$ and $v_i = \log u_i$ is the scale of random effects, which is additive to fixed effects in the linear predictor. Lee et al. (2017) called the scale of $\mathbf{v}$ weak canonical, which makes the support of $\mathbf{v}$ to be the whole real line, and called the resulting extended likelihood $L_e(\boldsymbol{\theta}, \mathbf{v}; \mathbf{y}) = f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v})$ the h-likelihood. In Example 4.1 and 2, $\mathbf{v}$ (not $\mathbf{u}$) is weak canonical. However, in general, MHLEs cannot give MLEs for fixed parameters. □

4.2.1 Reformulation of H-Likelihood

In this chapter, we want to have the h-likelihood, whose MHLEs give MLEs for fixed parameters. Bjørnstad (1996) proved the extended likelihood principle that all the information about $(\boldsymbol{\theta}, \mathbf{u})$ in the data is in the extended likelihood

$$L_e(\boldsymbol{\theta}, \mathbf{u}; \mathbf{y}) = L(\boldsymbol{\theta}; \mathbf{y}) L_p(\mathbf{u}; \mathbf{y}).$$

According to the classical likelihood principle of Birnbaum (1962), $L(\boldsymbol{\theta}; \mathbf{y})$ contains all the information about $\boldsymbol{\theta}$ in the data $\mathbf{y}$. Thus, $L_p(\mathbf{u}; \mathbf{y})$ cannot have additional information about $\boldsymbol{\theta}$. Since $L(\boldsymbol{\theta}; \mathbf{y})$ does not involve $\mathbf{u}$, intuitively, $L_p(\mathbf{u}; \mathbf{y})$ captures all the information about $\mathbf{u}$ in the data $\mathbf{y}$. Together with the

classical likelihood principle, this means $L_p(\mathbf{u};\mathbf{y})$ must contain all the information about $\mathbf{u}$ in the data. This tells us that inference of random parameters should be based on the conditional model $L_p(\mathbf{u};\mathbf{y}) = f_{\boldsymbol{\theta}}(\mathbf{u}|\mathbf{y})$.

Let $\ell_e(\boldsymbol{\theta}, \mathbf{v}; \mathbf{y})$ be Lee and Nelder's (1996) original h-likelihood. Whereas the MLEs are invariant with respect to any transformation of $\boldsymbol{\theta}$, Lee and Nelder (2005) showed that MHLEs for $\mathbf{v}$ can be invariant only for the linear transformation of $\mathbf{v}$ due to the Jacobian term. Given the data, let $\mathbf{v}^*$ be a linear transformation of $\mathbf{v}$,

$$\mathbf{v}^* = \exp\{c(\boldsymbol{\theta};\mathbf{y})\} \cdot \mathbf{v}$$

to give a Jacobian term $|\partial \mathbf{v}^*/\partial \mathbf{v}| = \exp\{c(\boldsymbol{\theta};\mathbf{y})\}$. Then, $\ell_e(\boldsymbol{\theta}, \mathbf{v}^*)$ and $\ell_e(\boldsymbol{\theta}, \mathbf{v})$ give the identical inference for $\mathbf{v}$. Define the h-likelihood as an extended likelihood on a scale $\mathbf{v}^*$,

$$h(\boldsymbol{\theta}, \mathbf{v}) = \ell(\boldsymbol{\theta};\mathbf{y}) + \ell_p(\mathbf{v}^*;\mathbf{y}) = \ell_e(\boldsymbol{\theta}, \mathbf{v}^*;\mathbf{y}) = \ell_e(\boldsymbol{\theta}, \mathbf{v};\mathbf{y}) + c(\boldsymbol{\theta};\mathbf{y}). \tag{4.4}$$

Here $c(\boldsymbol{\theta};\mathbf{y}) = \log|\partial \mathbf{v}^*/\partial \mathbf{v}|$ is a function of $\boldsymbol{\theta}$ and $\mathbf{y}$. Now we want that MHLEs for $\boldsymbol{\theta}$ are MLEs. Note that such a function $c(\boldsymbol{\theta};\mathbf{y})$ always exists. For example, let $c(\boldsymbol{\theta};\mathbf{y}) = -\ell_p(\widetilde{\mathbf{v}};\mathbf{y})$ where $\ell_p(\mathbf{v};\mathbf{y}) = \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})$ and

$$\widetilde{\mathbf{v}} = \arg\max_{\mathbf{v}} h(\boldsymbol{\theta}, \mathbf{v};\mathbf{y}) = \arg\max_{\mathbf{v}} \ell_p(\mathbf{v};\mathbf{y}),$$

then we have

$$h(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) = \ell(\boldsymbol{\theta};\mathbf{y}) = \log f_{\boldsymbol{\theta}}(\mathbf{y}).$$

Lee et al. (2017) called the scale $\mathbf{v}$ canonical if $c(\boldsymbol{\theta};\mathbf{y}) = -\ell_p(\widetilde{\mathbf{v}};\mathbf{y})$ does not

depend on $\boldsymbol{\theta}$, but such a scale rarely exists. The new h-likelihood can give MLEs for any scale of $\mathbf{v}$.

The classical likelihood $L(\boldsymbol{\theta}; \mathbf{y})$ is identical to the statistical model $f_\theta(\mathbf{y})$ for the data generation. However, the hierarchical model,

$$f_{\boldsymbol{\theta}}(\mathbf{v}^*) f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}^*),$$

would be invalid for the data generation, since $\mathbf{v}^*$ depends on the data $\mathbf{y}$. Thus, for a hierarchical model of the data generation, we use

$$f_{\boldsymbol{\theta}}(\mathbf{v}) f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}),$$

but for inferences we use $h(\boldsymbol{\theta}, \mathbf{v}) = \log L_e(\boldsymbol{\theta}, \mathbf{v}^*; \mathbf{y}) \neq \log L_e(\boldsymbol{\theta}, \mathbf{v}; \mathbf{y})$.

Example 4.1 (continued). Return to the one-way random-effect model with an extended likelihood,

$$\ell_e(\boldsymbol{\theta}, \mathbf{v}) = -\frac{1}{2}\left[\frac{\sum_{i,j}(y_{ij} - \mu_0 - v_i)^2}{\sigma^2} + \frac{\sum_i v_i^2}{\lambda^2} + N\log\sigma^2 + n\log\lambda^2\right],$$

and $\boldsymbol{\theta} = (\mu_0, \sigma^2, \lambda^2)^T$. Here, the joint maximization cannot give MLEs for the variance components σ^2 and λ^2. With the new h-likelihood,

$$h(\boldsymbol{\theta}, \mathbf{v}) = \ell_e(\boldsymbol{\theta}, \mathbf{v}) - \frac{n}{2}\log\left(\frac{m\lambda^2 + \sigma^2}{\lambda^2\sigma^2}\right),$$

the MHLEs become MLE of $\boldsymbol{\theta}$ and BUP of $\mathbf{v}$, i.e., E($\mathbf{v}|\mathbf{y}$). Lee and Lee (2023) showed that the new h-likelihood (4.4) is well-defined in general linear mixed models with temporal-spatial random effects, leading to an efficient fitting

algorithm for deep neural network (DNN) with normal random effects. □

Example 4.2 (continued). Consider a Poisson random effect model with a linear predictor

$$\eta_{ij} = \log(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \log u_i,$$

where $u_i \sim \text{Gamma}(\alpha, \alpha)$ with $\text{E}(u_i) = 1$ and $\text{Var}(u_i) = 1/\alpha$. Here $\mathbf{v}$-scale is weak canonical, leading to the h-likelihood,

$$\begin{aligned} h(\boldsymbol{\theta}, \mathbf{v}) = &\sum_{i,j} \left\{ y_{ij}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i) - \exp(\mathbf{x}_{ij}^T + v_i) \right\} \\ &+ \sum_i \left\{ \alpha(v_i - e^{v_i}) + \alpha \log \alpha - \log \Gamma(\alpha) + c_i(\alpha; \mathbf{y}) \right\}, \end{aligned}$$

where $c_i(\alpha; \mathbf{y}) = (y_{i+} + \alpha) + \log \Gamma(y_{i+} + \alpha) - (y_{i+} + \alpha) \log(y_{i+} + \alpha)$. Besides the MLEs for the whole fixed parameters $\boldsymbol{\theta}$, including variance component α, this h-likelihood yields the BUP for u_i's,

$$\widetilde{u}_i = \frac{y_{i+} + \alpha}{\mu_{i+} + \alpha} = \text{E}(u_i|\mathbf{y}),$$

and

$$\left[-\frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial u_i^2} \right]^{-1}_{u_i = \widetilde{u}_i} = \frac{\widetilde{u}_i^2}{y_{i+} + \alpha} = \frac{y_{i+} + \alpha}{(\mu_{i+} + \alpha)^2} = \text{Var}(u_i|\mathbf{y}).$$

Lee et al. (2023a) showed that this new h-likelihood gives a fast end-to-end learning algorithm for Poisson-gamma DNN. Ha and Lee (2003) showed that the frailty models for survival analysis can be fitted by Poisson HGLMs. Based on the new profiled h-likelihood, Lee et al. (2023b) proposed an online learning algorithm for DNN gamma frailty models. □

4.2.2 Bartlizable scale of random effects

We derived a new formulation of the h-likelihood (4.4) in Section 4.2.1. In this section, we focus on the scale of random parameters to form the h-likelihood, which gives asymptotically the BUPs. For the notational convenience, we define h-score as the first derivative,

$$S(\boldsymbol{\theta}, \mathbf{v}) = \frac{\partial h(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})},$$

h-information as the negative Hessian matrix,

$$I(\boldsymbol{\theta}, \mathbf{v}) = \frac{-\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})\partial(\boldsymbol{\theta}, \mathbf{v})^T},$$

expected h-information as $\mathcal{I}_{\boldsymbol{\theta}} = \mathrm{E}\{I(\boldsymbol{\theta}, \mathbf{v})\}$ and observed h-information as $\widehat{I} = I(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})$ with MHLEs $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{v}}$. Meng (2009) showed that Lee and Nelder's (1996) h-likelihood with the weak canonical scale $\ell_e(\boldsymbol{\theta}, \mathbf{v}; \mathbf{y})$ is 'Bartlizable' if and only if

$$\mathrm{E}\left[\frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{v})}{\partial \mathbf{v}}\right] = \int \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{v})}{\partial \mathbf{v}} d\mathbf{v} = 0 \quad \text{and}$$
$$\mathrm{E}\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(\mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T} + \left(\frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{v})}{\partial \mathbf{v}}\right)\left(\frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{v})}{\partial \mathbf{v}}\right)^T\right] = \int \frac{\partial^2 f_{\theta}(\mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T} d\mathbf{v} = 0.$$

Meng (2009) further derived an easily verifiable sufficient condition for Bartlizability that $f_{\boldsymbol{\theta}}(\mathbf{v}) = 0$ and $\partial f_{\boldsymbol{\theta}}(\mathbf{v})/\partial \mathbf{v}$ at the boundary $\partial\Omega_{\mathbf{v}}$ of the support $\Omega_{\mathbf{v}}$ of $\mathbf{v}$. Under the weak canonical scale, it is straightforward to have $f_{\boldsymbol{\theta}}(\mathbf{v}) = \partial f_{\boldsymbol{\theta}}(\mathbf{v})/\partial \mathbf{v} = 0$ at $\mathbf{v} = \pm\infty \in \partial\Omega_{\mathbf{v}}$. Thus, Lee and Nelder's (1996) h-likelihood $\ell_e(\boldsymbol{\theta}, \mathbf{v})$ is Bartlizable (Meng, 2009). Firth (2006) noted that the

weak canonical scale is ambiguous if there is no fixed effect in the linear predictor. In this chapter, we call $h(\boldsymbol{\theta}, \mathbf{v})$ of the form (4.4) the h-likelihood, if the scale $\mathbf{v}$ is Bartlizable. This resolves the ambiguity in defining the h-likelihood raised by Firth (2006). Note that the Bartlizable scale is more general than the weak-canonical scale as we shall discuss. Now we modify Meng's (2009) Bartlizability for the new h-likelihood and shows a sufficient condition.

Definition 4.1. *The h-likelihood $h(\boldsymbol{\theta}, \mathbf{v})$ is Bartlizable if the following first and second Bartlett identities hold:*

$$E[S(\boldsymbol{\theta}, \mathbf{v})] = 0 \quad \textit{and} \quad E\left[S(\boldsymbol{\theta}, \mathbf{v})S(\boldsymbol{\theta}, \mathbf{v})^T - I(\boldsymbol{\theta}, \mathbf{v})\right] = 0.$$

Then, following Lemma relaxes Meng's (2009) sufficient condition for Bartlizability. Proofs are derived in Appendix 4.6.

Lemma 4.1. (*i*) *For any continuous random effects $\mathbf{u} = (u_1, ..., u_n)^T$, there exists a Bartlizable transformation $v_i = g_v(u_i)$.*
(*ii*) *The scale $\mathbf{v}$ is Bartlizable if*

$$f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y}) = 0 \quad \textit{and} \quad \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{v}|\boldsymbol{y})}{\partial \mathbf{v}} = \mathbf{0} \quad \textit{for all } \mathbf{v} \in \partial\Omega_{\mathbf{v}}. \tag{4.5}$$

(*iii*) *If $f_{\boldsymbol{\theta}}(\mathbf{v})$ is differentiable for any $\mathbf{v} = (v_1, ..., v_n)^T \in \mathbb{R}^n$, the scale $\mathbf{v}$ is Bartlizable.*

Figure 4.1 shows the relationships among the h-likelihood $h(\boldsymbol{\theta}, \mathbf{v})$ (blue), the Fisher likelihood $h(\boldsymbol{\theta}, \widetilde{\mathbf{v}})$ (red), and the profiled predictive likelihood $h(\widetilde{\boldsymbol{\theta}}, \mathbf{v})$ (orange) with $\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{v})$. The Fisher likelihood is the upper bound of projection of h-likelihood to the parameter space $\boldsymbol{\Theta}$, whereas the profiled

predictive likelihood Lee and Nelder (2002) is the upper bound of projection of h-likelihood to the support of random effects $\Omega_{\mathbf{v}}$. See-saw algorithm leads to the global maxima of both fixed and random parameters. Here, $h(\widetilde{\boldsymbol{\theta}}, \mathbf{v})$ is associated with the E-step and $h(\boldsymbol{\theta}, \widetilde{\mathbf{v}})$ is associated with the M-step. Thus, the h-likelihood procedure is computationally straightforward, providing MHLEs for both fixed and random parameters and their standard error estimators.

Example 4.1 (continued). For the one-way random effect model, consider the h-likelihood on the **v**-scale,

$$h\left(\boldsymbol{\theta}, \mathbf{v}\right) = -\frac{\sum_{i,j}(y_{ij} - \mu_0 - v_i)^2}{2\sigma^2} - \frac{\sum_i v_i^2}{2\lambda^2} - \frac{1}{2}\left[N \log \sigma^2 + n \log \lambda^2 + n \log\left(\frac{m\lambda^2 + \sigma^2}{\lambda^2 \sigma^2}\right)\right].$$

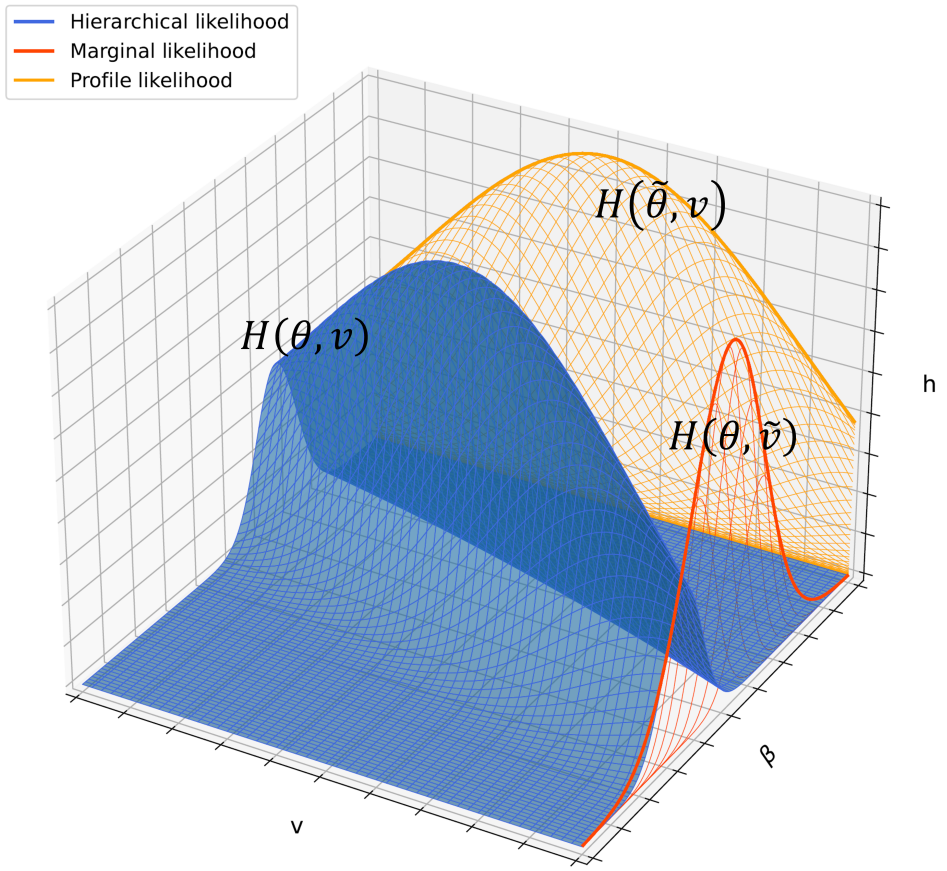
Lemma 4.1 implies that **v** is Bartlizable. This h-likelihood gives MLE for $\boldsymbol{\theta}$ and BUP for v_i,

$$\widetilde{v}_i^{(1)} = \frac{m\lambda^2(\bar{y}_i - \mu_0)}{\sigma^2 + m\lambda^2} = \mathrm{E}(v_i|\mathbf{y}),$$

where $\bar{y}_i = \sum_j y_{ij}/m$. Henderson et al. (1959) called it the best linear unbiased predictor (BLUP) for v_i, because it is linear in **y**. Here, **u** is also Bartlizable to give a h-likelihood,

$$h\left(\boldsymbol{\theta}, \mathbf{u}\right) = -\frac{\sum_{i,j}(y_{ij} - \mu_0 - \log u_i)^2}{2\sigma^2} - \frac{\sum_i (\log u_i)^2}{2\lambda^2} - \frac{1}{2}\left[N \log \sigma^2 + n \log \lambda^2 + n \log\left(\frac{m\lambda^2 + \sigma^2}{\lambda^2 \sigma^2}\right)\right] - \sum_i \log u_i,$$

which gives MLEs for $\boldsymbol{\theta}$. Remind that $\ell_e(\boldsymbol{\theta}, \mathbf{u})$ cannot give MLEs. The MHLEs



Find $\tilde{v}$ by solving $\frac{\partial h(\theta,v)}{\partial v} = 0$

$$\begin{aligned} H(\theta, \tilde{v}) &= L_p(\tilde{v} \mid y; \theta) L_m(\theta) \\ &\propto L_m(\theta) \\ H(\tilde{\theta}, v) &= L_p(v \mid y; \tilde{\theta}) L_m(\tilde{\theta}) \\ &\propto L_p(v \mid y; \tilde{\theta}) \end{aligned}$$

Find $\tilde{\theta}$ by solving $\frac{\partial h(\theta,v)}{\partial \theta} = 0$

Figure 4.1: The h-likelihood (blue), the marginal likelihood (red), and the profile likelihood (orange) are computed with $N = 2000$ samples from $y_i \sim N(x_i\beta + v, 1)$ and $v \sim N(0, 1)$. True values of β and v are set to be 1 and -1, respectively. x_i's are generated from Uniform$(0, 1)$.

for v_i is

$$\widetilde{v}_i^{(2)} = \frac{m\lambda^2(\bar{y}_i - \mu_0) - \sigma^2\lambda^2}{\sigma^2 + m\lambda^2}$$

which is not the BUP of $\mathbf{v}$. However, it gives the BUP of $w_i = 1/u_i^2$,

$$\widetilde{w}_i^{(2)} = \exp\left[\frac{-2m\lambda^2(\bar{y}_i - \mu_0) + 2\sigma^2\lambda^2}{\sigma^2 + m\lambda^2}\right] = \mathrm{E}(w_i|\mathbf{y}).$$

Thus, for a prediction of $1/u_i^2$, the h-likelihood $h(\boldsymbol{\theta}, \mathbf{u})$ would be desirable. □

Example 4.2 (continued). Consider a Poisson-gamma HGLM with a linear predictor

$$\eta_{ij} = \log(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \log u_i,$$

where $u_i \sim \mathrm{Gamma}(\alpha, \alpha)$ with $\mathrm{E}(u_i) = 1$ and $\mathrm{Var}(u_i) = 1/\alpha$. If we define a h-likelihood with $\mathbf{u}$-scale, for $\alpha \leq 1$,

$$\begin{aligned}\mathrm{E}\left[\frac{\partial h(\boldsymbol{\theta}, \mathbf{u})}{\partial u_i}\right] &= \mathrm{E}\left[\mathrm{E}\left[\frac{y_{i+} + \alpha - 1}{u_i} - (\mu_{i+} + \alpha)\middle|\mathbf{u}\right]\right] \\ &= (\alpha - 1)\mathrm{E}(u_i^{-1}) - \alpha = \infty,\end{aligned}$$

where $y_{i+} = \sum_j y_{ij}$, $\mu_{i+} = \sum_j \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})$, and $\boldsymbol{\theta} = (\beta_0, ..., \beta_p, \alpha)^T$. Thus, $\mathbf{u}$-scale is not Bartlizable. It produces the MHLE, $\widetilde{u}_i = 0$ when $y_{i+} < 1 - \alpha$, which is not a BUP. Therefore, it is important to use Bartlizable scale for the h-likelihood to have BUP. □

4.3 Main Results

We first study asymptotic properties of MHLEs when $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = o_p(1)$ and $\widehat{\mathbf{v}} - \mathbf{v} = o_p(1)$. In HGLMs, we generally have $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = o_p(1)$ and $\widehat{\mathbf{v}} - \mathbf{v} = o_p(1)$ to allow the consistency for MHLEs.

Example 4.1 (continued). In the one-way random effect models,

$$\widetilde{v}_i = v_i(\boldsymbol{\theta}, \mathbf{y}) = \arg\max_{v_i} h(\boldsymbol{\theta}, \mathbf{u}) = \frac{\lambda^2}{\sigma^2 + n\lambda^2} \sum_{j=1}^{n} (y_{ij} - \mu_0) = \mathrm{E}(v_i|y).$$

Here, $\widehat{v}_i = v_i(\widehat{\boldsymbol{\theta}}, \mathbf{y})$ leads to

$$\lim_{N\to\infty} \widehat{v}_i = \lim_{N\to\infty} \widetilde{v}_i = \lim_{N\to\infty} \frac{n\lambda^2}{\sigma^2 + n\lambda^2} (\bar{y}_{i\cdot} - \mu_0) = \lim_{N\to\infty} \frac{n\lambda^2}{\sigma^2 + n\lambda^2} \left(v_i + \bar{\epsilon}_{i\cdot}\right) = v_i,$$

where $\bar{\epsilon}_i = \sum_{j=1}^{n} \epsilon_{ij}/n$, $\epsilon_{ij} = y_{ij} - \mu_0 - v_i \sim N(0, \sigma^2)$, $\bar{y}_i = \sum_{j=1}^{n} y_{ij}/n$, and $N \to \infty$ with $n \to \infty$ and $m \to \infty$. Thus, we have $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = o_p(1)$ and $\widehat{\mathbf{v}} - \mathbf{v} = o_p(1)$ in the one-way random effect model. □

Besides the MLEs for $\boldsymbol{\theta}$, we show that the MHL procedures give the asymptotic BUPs for $\mathbf{v}$. Suppose that $\boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})$ is an arbitrary function of $(\boldsymbol{\theta}, \mathbf{v})$ and $\widehat{\boldsymbol{\zeta}}(\mathbf{y})$ is an unbiased estimator of $\boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})$ such that

$$\mathrm{E}\left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})\right] = 0.$$

Then, it is immediate to have a generalized Cramer-Rao lower bound.

Theorem 4.1. *Under the regularity conditions in Appendix,*

$$Var\left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})\right] \geq \boldsymbol{\zeta}'_{\boldsymbol{\theta}}\, \mathcal{I}_{\boldsymbol{\theta}}^{-1}\, \boldsymbol{\zeta}'^{\,T}_{\boldsymbol{\theta}} \tag{4.6}$$

where the matrix inequality $A \geq B$ means that $A - B$ is positive semi-definite and

$$\boldsymbol{\zeta}'_{\boldsymbol{\theta}} = E\left\{ \frac{\partial \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \right\}.$$

Remark: Given $\boldsymbol{\theta}$, the lower bound (4.6) becomes the Bayesian Cramer-Rao bound (Van Trees, 1968), since

$$\frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^T} = \frac{\partial^2 \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^T} = \frac{\partial^2 \ell_p(\mathbf{v}; \mathbf{y})}{\partial \mathbf{v} \partial \mathbf{v}^T}.$$

Thus, the lower bound (4.6) extends the Cramer-Rao bound for $\boldsymbol{\zeta}(\boldsymbol{\theta})$ and the Bayesian Cramer-Rao bound for $\boldsymbol{\zeta}(\mathbf{v})$. The Bayesian Cramer-Rao bound can be obtained by the BUP $\mathrm{E}(\boldsymbol{\zeta}(\mathbf{v})|\mathbf{y})$, even in finite samples. However, since the MHLE is the mode, not the conditional expectation, it achieves the bound asymptotically. As we shall show, in one-way random effect models, the MHLE $\widetilde{\mathbf{v}} = \arg\max_{\mathbf{v}} h(\boldsymbol{\theta}, \mathbf{v}) = \mathrm{E}(\mathbf{v}|\mathbf{y})$ is the BUP of $\mathbf{v}$, achieving the lower bound (4.6) if $\boldsymbol{\theta}$ is known. However, in finite samples, $\boldsymbol{\zeta}(\widetilde{\mathbf{v}})$ may not be the BUP for $\boldsymbol{\zeta}(\mathbf{v})$ in general. The following theorem shows that $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{v}} = \widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}}; \mathbf{y})$ can achieve the lower bound (4.6) asymptotically.

Theorem 4.2. *Under the regularity conditions in Appendix,*

$$\begin{bmatrix} \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \\ \widehat{\mathbf{v}} - \mathbf{v} \end{bmatrix} \stackrel{d}{\to} N\left(\mathbf{0},\ \mathcal{I}_{\boldsymbol{\theta}}^{-1}\right).$$

and the variance can be consistently estimated by the inverse of observed h-information $\widehat{I}^{-1} = I(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})^{-1}$.

Remark: Theorem 4.1 and 4.2 show the asymptotic efficiency of the MHLEs, which achieves the unfulfilled aim of Lee and Nelder's (1996) h-likelihood.

Example 4.1 (continued). In the one-way random effect model, we have two different MHLEs,

$$\widetilde{v}_i^{(1)} = \frac{m\lambda^2(\bar{y}_i - \mu_0)}{\sigma^2 + m\lambda^2} = \mathrm{E}(v_i|\mathbf{y}), \quad \text{and} \quad \widetilde{v}_i^{(2)} = \frac{m\lambda^2(\bar{y}_i - \mu_0) - \sigma^2\lambda^2}{\sigma^2 + m\lambda^2},$$

from the **v**-scale and **u**-scale, respectively. Here, **v**-scale leads to BUP for μ_{ij},

$$\widetilde{\mu}_{ij}^{(1)} = \mu_0 + \widetilde{v}_i^{(1)} = \mu_0 + \mathrm{E}(v_i|\mathbf{y}) = \mathrm{E}(\mu_{ij}|\mathbf{y}).$$

On the other hand, **u**-scale leads to

$$\widetilde{\mu}_{ij}^{(2)} = \mu_0 + \widetilde{v}_i^{(2)} = \mu_0 + \mathrm{E}(v_i|\mathbf{y}) - \frac{\sigma^2\lambda^2}{\sigma^2 + m\lambda^2} = \mathrm{E}(\mu_{ij}|\mathbf{y}) + O(1/m).$$

For simplicity of arguments, let $\sigma^2 = \lambda^2 = 1$ and $n = 2$. Let us investigate the asymptotic efficiency of the MHLEs. The use of v-scale leads to the MLE $\widehat{\mu}_0 = \bar{y} = (\bar{y}_1 + \bar{y}_2)/2$ and the BUP $\widehat{v}_1^{(1)} = m(\bar{y}_1 - \widehat{\mu}_0)/(m+1) = m(\bar{y}_1 - \bar{y}_2)/(2m+2)$ and $\widehat{v}_2^{(1)} = m(\bar{y}_2 - \widehat{\mu}_0)/(m+1) = m(\bar{y}_2 - \bar{y}_1)/(2m+2)$. Note that $\mathrm{E}(\mu_0 - \widehat{\mu}_0) = \mathrm{E}(v_1 - \widehat{v}_1^{(1)}) = \mathrm{E}(v_2 - \widehat{v}_2^{(1)}) = 0$. Here the variance and covariance are given by

$$\begin{aligned}
&\mathrm{Var}(\mu_0 - \widehat{\mu}_0) = \mathrm{Var}(\bar{y}) = \mathrm{Var}(\mathrm{E}(\bar{y}|v_1, v_2)) + \mathrm{E}(\mathrm{Var}(\bar{y}|v_1, v_2)) = \frac{m+1}{m}, \\
&\mathrm{Var}(v_i - \widehat{v}_i^{(1)}) = \mathrm{Var}(\mathrm{E}(v_i - \widehat{v}_i^{(1)}|\mathbf{y})) + \mathrm{E}(\mathrm{Var}(v_i - \widehat{v}_i^{(1)}|\mathbf{y})) = \frac{m+2}{2m+2}, \\
&\mathrm{Cov}(\mu_0 - \widehat{\mu}_0, v_i - \widehat{v}_i^{(1)}) = -\mathrm{E}\left[\bar{y}\{\mathrm{E}(v_i|\mathbf{y}) - \widehat{v}_i^{(1)}\}\right] = -\frac{1}{2}, \\
&\mathrm{Cov}(v_1 - \widehat{v}_1^{(1)}, v_2 - \widehat{v}_2^{(1)}) = \frac{m}{2m+2},
\end{aligned}$$

for $i = 1, 2$. Since the expected h-information gives

$$\left[-\frac{\partial^2 h(\mu_0, v_1, v_2)}{\partial(\mu_0, v_1, v_2)\partial(\mu_0, v_1, v_2)^T} \right]^{-1} = \begin{pmatrix} \frac{m+1}{2m} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{m+2}{2m+2} & \frac{m}{2m+2} \\ -\frac{1}{2} & \frac{m}{2m+2} & \frac{m+2}{2m+2} \end{pmatrix},$$

we can see that Theorem 4.2 holds exactly. □

Even though $\widetilde{\mathbf{v}} = \mathrm{E}(\mathbf{v}|\mathbf{y})$ is the BUP, $g(\widetilde{\mathbf{v}}) = g(\mathrm{E}(\mathbf{v}|\mathbf{y})) \neq \mathrm{E}(g(\mathbf{v})|\mathbf{y})$ is no longer the BUP of a nonlinear function $g(\mathbf{v})$. Consider a HGLM with the linear predictor,

$$\eta_{ij} = g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i.$$

We want to know when there exists a predictor $\widetilde{v}_i$ that gives the BUP of μ_{ij} for all j in finite samples, i.e.,

$$\widetilde{\mu}_{ij} = g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \widetilde{v}_i) = \mathrm{E}(\mu_{ij}|\mathbf{y}). \tag{4.7}$$

We have already seen that such $\widetilde{v}_i$ exists for the linear mixed models and Poisson-gamma HGLMs. However, the following theorem shows that such $\widetilde{v}_i$ may not always exist in finite samples.

Theorem 4.3. *Suppose that $\boldsymbol{\beta} \neq \mathbf{0}$ and $\boldsymbol{x}_{ij}^T \boldsymbol{\beta}$ can freely move on $\mathbb{R}$. Given $\boldsymbol{\theta}$, there exists $\widetilde{v}_i$ satisfying* (4.7) *if and only if there exists a constant $c_i \in \Omega_v$ such that*

$$E\left(g_{(k)}^{-1}(v_i)|\mathbf{y}\right) = g_{(k)}^{-1}(c_i) \quad \textit{for all } k = 0, 1, 2, ... \tag{4.8}$$

where $g_{(k)}^{-1}(\cdot)$ is the k-th order derivative of inverse link $g^{-1}(\cdot)$.

Remark: When $g(\cdot)$ is identity, as in a linear mixed model, $g_{(k)}^{-1}(v_i) =$

0 for all $k \geq 1$. Thus, $c_i = \mathrm{E}(v_i|\mathbf{y})$ satisfies the condition. When $g(\cdot)$ is logarithm, as in a Poisson-gamma HGLM, $g_{(k)}^{-1}(v_i) = \exp(v_i)$ for all k. Thus, $c_i = \log \mathrm{E}(\exp(v_i)|\mathbf{y})$ satisfies the condition. Now consider a binomial HGLM with $p_{ij} = P(Y_{ij} = 1|\mathbf{x}_{ij}, v_i)$ and logit link,

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i.$$

Suppose that there exists a predictor $\widetilde{v}_i^*$ satisfying (4.7). Then, by Theorem 4.3, there exists a constant c_i satisfying (4.8). Note here that $g_{(0)}^{-1}(v_i) = \{1 + \exp(-v_i)\}^{-1}$ and

$$g_{(1)}^{-1}(v_i) = \exp(-v_i)\{1+\exp(-v_i)\}^{-2} = g_{(0)}^{-1}(v_i) - \{g_{(0)}^{-1}(v_i)\}^2.$$

Let $T = g_{(0)}^{-1}(v_i)$ and $t = g_{(0)}^{-1}(c_i)$. If $\mathrm{E}(T) = t$ holds, then $\mathrm{E}(T - T^2) < E(T) - E(T)^2 = t - t^2$ by the Jensen's inequality. This implies that the equation (4.8) cannot be satisfied simultaneously for both $k = 0$ and $k = 1$. Thus, in finite samples, the BUP for μ_{ij} cannot be obtainable. However, the MHLE gives an asymptotic BUP for μ_{ij}.

4.4 H-likelihood theory for irregular cases

In missing data problems, Meng (2009) pointed out that predictors of missing data cannot be summarizable, $\widehat{\mathbf{v}} - \mathbf{v} = O_p(1)$, even though $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = o_p(1)$. Thus, an asymptotic consistency of the predictor $\widehat{\mathbf{v}}$ may not be possible. This explains why the EM algorithm (Dempster et al., 1977) does not pay much attention to prediction of unobservable missing data. We also investigate the

case when $\widehat{\mathbf{v}} - \mathbf{v} = o_p(1)$ and $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(1)$. We study how to overcome difficulty raised by insummarizability.

4.4.1 Missing data problem when $\widehat{\mathbf{v}} - \mathbf{v} = O_p(1)$

Suppose that $\mathbf{y} = (y_1, ..., y_n)^T$ are i.i.d. sample from $f_{\boldsymbol{\theta}}(\mathbf{y})$ with fixed unknown parameter $\boldsymbol{\theta}$ and let $u = y_{n+1}$ be a future observation. Then, prediction of a future observation u can be viewed as the missing data problem. Here, $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = o_p(1)$. But for any scale $v = g(u)$, Meng (2009) noted that

$$\widehat{v} - v = g\left(\widehat{u}\right) - g\left(u\right) = g'\left(\widetilde{u}\right)\left(\widehat{u} - u\right) + R_n,$$

where $\widehat{u} = g^{-1}(\widehat{v})$,

$$R_n = O_p(1) \text{ and } g'(\widetilde{u})(\widehat{u} - u) = O_p(1).$$

Thus, $\widehat{v} - v = O_p(1)$ because of the *non-negligible* remainder term R_n, i.e., the consistency and asymptotic normality for MHLE $\widehat{v}$ look non-sensible.

It is of interest to investigate how the MHLE $\widehat{v}$ can be used for the prediction of v. As the MLE $\widehat{\boldsymbol{\theta}}$ consistently estimates $\boldsymbol{\theta}$, similarly the MHLE $\widehat{v}$ predicts v by consistently estimating $\widetilde{v} = \widetilde{v}(\boldsymbol{\theta}, \mathbf{y})$, which is a function of the data and fixed unknown parameter $\boldsymbol{\theta}$. This clarifies the summarizability problem raised by Meng (2009); while $\widehat{v} - v = O_p(1)$, but

$$\widehat{v} - \widetilde{v} = o_p(1)$$

as derived in Theorem 4.4 below. Let $\varepsilon = v - \widetilde{v}$, then

$$v - \widehat{v} = \widetilde{v} - \widehat{v} + \varepsilon,$$

and $\varepsilon = O_p(1)$ in missing data problems. In view of predicting future (or missing) data, we estimate ε as null, i.e., $\widehat{v}$ is estimating $\widetilde{v}$ to predict v. Then, we have

$$\mathrm{Var}\,(\widehat{v} - v) = \mathrm{Var}(\widehat{v} - \widetilde{v}) + \mathrm{Var}(\varepsilon|\mathbf{y}).$$

The first term is the variance due to estimating $\widetilde{v}$ by $\widehat{v}$ and the second term is the variance due to the unidentifiable error term ε. But the second term can be determined by the model assumption, which does not decrease with larger sample size. To obtain a standard error for prediction of v, we need to estimate

$$\mathrm{Var}(\varepsilon|\mathbf{y}) = \mathrm{Var}(v - \widetilde{v}|\mathbf{y}) = \mathrm{Var}(v|\mathbf{y}).$$

From Theorem 4.2, we have

$$\frac{\partial \widetilde{v}^T}{\partial \theta} = -I_{\theta v} I_{vv}^{-1}$$

and the variance estimator of $\widehat{\theta}$ is $\widehat{I}^{\theta\theta}$, where $I_{\theta v} = -\partial^2 h/\partial\theta\partial v^T|_{v=\widetilde{v}}$ and $I_{vv} = -\partial^2 h/\partial v\partial v^T|_{v=\widetilde{v}}$. Then, by using the delta method, we have the asymptotic normality of $\widehat{v}$ as follows.

Theorem 4.4. *Under regularity conditions in Appendix, we have*

$$\sqrt{n}\,(\widehat{v} - \widetilde{v}) \xrightarrow{d} N(0, V)\,,$$

where $V = \lim_{n\to\infty} n\widehat{I}_{vv}^{-1}\widehat{I}_{v\theta}\widehat{I}^{\theta\theta}\widehat{I}_{\theta v}\widehat{I}_{vv}^{-1}$ *and* $\widehat{I}_{\theta v}$, $\widehat{I}_{vv}$ *are evaluated at* $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. *The variance of* $\widehat{v} - \widetilde{v}$ *can be estimated as*

$$\widehat{Var}(\widehat{v} - \widetilde{v}) = \widehat{I}_{vv}^{-1}\widehat{I}_{v\theta}\widehat{I}^{\theta\theta}\widehat{I}_{\theta v}\widehat{I}_{vv}^{-1}. \tag{4.9}$$

Remark: If $\mathrm{E}(\varepsilon) = 0$, $\widehat{v}$ is an asymptotically unbiased predictor of v. However, the assumption $\mathrm{E}(\varepsilon) = 0$ is coming from the model assumption which may not be checkable by observed data. Now, to discuss the estimation of the variance due to the model error ε, suppose that there exists a normalizing transformation $z = k(v) = k\{g(u)\} = k \circ g(u) = r(u)$ with $r(\cdot) = k \circ g(\cdot)$ such that $L_p(z|y;\theta)$ is from the normal density with mean $\widetilde{v} = \arg\max_z L_p(v|\mathbf{y};\theta)$ and covariance matrix I_{vv}^{-1}, where $I_{vv} = -\partial^2 h(\theta, z)/\partial z \partial z^T|_{z=\widetilde{v}}$. Then, it gives the h-likelihood

$$h\left(\theta, z\right) = \ell_m(\theta) + \frac{1}{2}\log\left|\frac{1}{2\pi}I_{vv}\right| - \frac{1}{2}\left(z - \widetilde{v}\right)^T I_{vv}\left(z - \widetilde{v}\right).$$

Here, $\widetilde{v} = \mathrm{E}(z|\mathbf{y}) = r(\widetilde{u})$ provided by the normality of the predictive likelihood $L_p(v|\mathbf{y};\boldsymbol{\theta})$. This leads to $\mathrm{E}(\varepsilon) = \mathrm{E}(z - \widetilde{z}) = 0$,

$$\mathrm{Var}\left(\widehat{z} - z\right) = \mathrm{Var}\left(\widehat{z} - \widetilde{z}\right) + \mathrm{E}\left\{\mathrm{Var}\left(\widetilde{z} - v|\mathbf{y}\right)\right\}$$

and $\widehat{\mathrm{Var}}(\widetilde{z} - v|\mathbf{y}) = \widehat{I}_{vv}^{-1}$, where $\widehat{z} = r(\widehat{u})$ and $\widetilde{z} = r(\widetilde{u}) = \mathrm{E}(z|y)$ with $\widetilde{u} = \widetilde{u}(\boldsymbol{\theta}, \mathbf{y})$. This gives

$$\widehat{\mathrm{Var}}\left(\widehat{z} - z\right) = \widehat{\mathrm{Var}}\left(\widehat{z} - \widetilde{z}\right) + \widehat{\mathrm{Var}}\left(\widetilde{z} - v|\mathbf{y},\right) = \widehat{I}_{vv}^{-1}\widehat{I}_{z\theta}\widehat{I}^{\theta\theta}\widehat{I}_{\theta z}\widehat{I}_{vv}^{-1} + \widehat{I}_{vv}^{-1} = \widehat{I}^{zz}.$$

Therefore, if a normalizing transformation exists, the h-likelihood gives not

only MHLEs but also their variance estimators. Moreover, if u itself satisfies normal approximation well, then, we can have a reasonable variance estimator from the Hessian matrix of h-likelihood

$$\begin{aligned}\widehat{\mathrm{Var}}\,(\widehat{u}-u) &= \widehat{\mathrm{Var}}\,(\widehat{u}-\widetilde{u})+\widehat{\mathrm{Var}}\,(\widetilde{u}-u|y)\\ &= \widehat{I}_{uu}^{-1}\widehat{I}_{u\theta}\widehat{I}^{\theta\theta}\widehat{I}_{\theta u}\widehat{I}_{uu}^{-1}+\widehat{I}_{uu}^{-1}=\widehat{I}^{uu}.\end{aligned}$$

where $\widehat{I}_{uu}^{-1}$ is coming from the model assumption.

Example 4.3 (Censored exponential model). Little and Rubin (2019) considered a censored exponential model, where $\mathbf{y}=(y_1,...,y_n)^T$ have an independent exponential distribution with mean θ and the missing mechanism $\delta = I(Y \le c)$ with a known constant c, so that the missing mechanism is not ignorable. Suppose that only the first n_{obs} data are observed and the rest $n_{mis}=n-n_{obs}$ data are missed. Let $\mathbf{u}=(u_1,...,u_{n_{mis}})^T$ with $u_i=y_{n_{obs}+i}$ for $i=1,...,n_{mis}$, then an extended (complete-data) likelihood can be defined as

$$\ell_e(\theta,\mathbf{u})=\log f_\theta(\mathbf{y}|\mathbf{u})+\log f_\theta(\mathbf{u})=-n\log\theta-\frac{n_{obs}\bar{y}_{obs}}{\theta}-\frac{n_{mis}\bar{u}}{\theta},$$

where $\bar{y}_{obs}=\sum_{i=1}^{n_{obs}} y_i/n_{obs}$ and $\bar{u}=\sum_{i=1}^{n_{mis}} u_i/n_{mis}$ are the sample means based on observed data and missing data, respectively. Little and Rubin (2019) noted that maximization of $\ell_e(\theta,\mathbf{u})$ provides the nonsensical modes

$$\widehat{\theta}=\frac{n_{obs}\bar{y}_{obs}+n_{mis}c}{n_{obs}+n_{mis}} \quad \text{and} \quad \widehat{u}_i=c.$$

However, it is worth emphasizing that this nonsensical result is due to the use

of non-Bartlizable u-scale, since

$$\mathrm{E}\left[\frac{\partial \log f_\theta(u_i)}{\partial u_i}\right] = \int_c^\infty \frac{1}{\theta}\exp\left\{-\frac{(u_i-c)}{\theta}\right\} du_i \neq 0.$$

Let $\mathbf{v} = (v_1, ..., v_{n_{mis}})^T$ with $v_i = \log(u_i - c) \in \mathbb{R}$, then v-scale is Bartlizable from Lemma 4.1, leading to the h-likelihood

$$h(\theta, \mathbf{v}) = \ell_e(\theta, \mathbf{v}) + n_{\text{mis}} = -n\log\theta - \frac{n_{obs}\bar{y}_{obs}}{\theta} - \frac{n_{mis}\bar{u}}{\theta} + n_{mis}\bar{v} + n_{mis}.$$

Here, $\widetilde{v}_i(\theta) = \log\theta$ gives

$$h(\theta, \widetilde{\mathbf{v}}) = -n_{obs}\log\theta - \frac{n_{obs}\bar{y}_{obs}}{\theta} - \frac{n_{mis}c}{\theta} = \ell(\theta),$$

which gives the MLE for fixed parameter θ and the MHLE for random parameters u,

$$\widehat{\theta} = \bar{y}_{obs} + c \cdot n_{mis}/n_{obs} \quad \text{and} \quad \widehat{u}_i = \widehat{\theta} + c > c,$$

respectively. Here, as $n_{obs} \to \infty$,

$$\mathrm{Var}(\widehat{\theta} - \theta) = \mathrm{Var}(\widehat{\theta}) = \mathrm{Var}(\bar{y}_{obs}) = \theta^2/n_{obs} \to 0$$

to give $\widehat{\theta} - \theta = o_p(1)$, but

$$\mathrm{Var}(\widehat{u}_i - u_i) = \theta^2 + \mathrm{Var}(\widehat{u}_i) \to \theta^2$$

to give $\widehat{u}_i - u_i = O_p(1)$. Note here that inverse of the observed h-information

with respect to θ and $\mathbf{u}$ is given by

$$I(\widehat{\theta}, \widehat{\mathbf{u}})^{-1} = \widehat{\theta}^2 \begin{pmatrix} n_{obs}^{-1} & n_{obs}^{-1} & \cdots & n_{obs}^{-1} \\ n_{obs}^{-1} & 1 + n_{obs}^{-1} & \cdots & n_{obs}^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ n_{obs}^{-1} & n_{obs}^{-1} & \cdots & 1 + n_{obs}^{-1} \end{pmatrix},$$

which leads to

$$\widehat{\mathrm{Var}}(\widehat{\theta}) = \widehat{\theta}^2/n_{obs} \to \mathrm{Var}(\widehat{\theta}) = \theta^2/n_{obs}$$

and

$$\widehat{\mathrm{Var}}(\widehat{u}_i - u_i) = \widehat{\theta}^2 + \widehat{\theta}^2/n_{obs} \to \mathrm{Var}(\widehat{u}_i - u_i) = \theta^2 + \theta^2/n_{obs}.$$

Here,

$$\mathrm{Var}(\widehat{u}_i - u_i) = \mathrm{Var}(\widehat{u}_i - \widetilde{u}_i) + \mathrm{Var}(\widetilde{u}_i - u_i),$$

where

$$\widehat{\mathrm{Var}}(\widehat{u}_i - \widetilde{u}_i) = \widehat{I}_{uu}^{-1}\widehat{I}_{u\theta}\widehat{I}^{\theta\theta}\widehat{I}_{\theta u}\widehat{I}_{uu}^{-1} = \widehat{\theta}^2/n_{obs} \to \mathrm{Var}(\widehat{u}_i - \widetilde{u}_i) = \theta^2/n_{obs}$$

and

$$\widehat{\mathrm{Var}}(\widetilde{u}_i - u_i) = \widehat{I}_{uu}^{-1} = \widehat{\theta}^2 \to \mathrm{Var}(\widetilde{u}_i - u_i) = \theta^2.$$

Thus, $\widehat{u}_i$ is consistently estimating $\widetilde{u}_i$, which is an unbiased estimator of u_i ($\mathrm{E}(\tilde{u}_i - u_i) = 0$) and the h-information matrix gives a consistent variance estimation for $\mathrm{Var}(\widehat{u}_i - u_i)$. □

4.4.2 Exponential-exponential HGLM when $\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta} = O_p(1)$

Let $u \sim \text{Exp}(1/\theta)$ and $y_i|u \sim \text{Exp}(u)$ with the density functions,

$$f_\theta(u) = \frac{1}{\theta}\exp\left(-\frac{u}{\theta}\right) \quad \text{and} \quad f(y_i|u) = u\exp(-uy_i),$$

for $i = 1, ..., n$. Then, the classical log-likelihood is

$$\ell(\theta) = -\log\theta - (n+1)\log(\theta^{-1} + n\bar{y}) + \log\Gamma(n+1),$$

where $\Gamma(\cdot)$ is the gamma function and $\bar{y} = (y_1 + \cdots + y_n)/n$. The MLE of θ is given by $\widehat{\theta} = 1/\bar{y}$ with the expectation

$$\text{E}(\widehat{\theta}) = \text{E}(1/\bar{y}) = \text{E}(\text{E}(1/\bar{y}|u)) = \frac{n\theta}{n-1} \to \theta$$

and the variance

$$\begin{aligned}\text{Var}(\widehat{\theta}) &= \text{Var}(1/\bar{y}) = \text{E}(\text{Var}(1/\bar{y}|u)) + \text{Var}(\text{E}(1/\bar{y}|u)) \\ &= \frac{n^3\theta^2}{(n-1)^2(n-2)} \to \theta^2 < \infty\end{aligned}$$

as $n \to \infty$. Thus, the MLE is asymptotically unbiased but $\widehat{\theta}-\theta = O_p(1)$. Here, the Hessian of classical likelihood gives an asymptotically consistent variance estimation $\widehat{\text{Var}}(\widehat{\boldsymbol{\theta}}) = \bar{y}^{-2}(n+1)/n$.

In this example, random parameter predictor is consistent. Lemma 4.1 implies that $v = \log u \in \mathbb{R}$ is Bartlizable. Thus, let us first define the h-

likelihood using v-scale,

$$\begin{aligned} h_1(\theta, v) &= \ell_e(\theta, v) + c_1(\theta; \mathbf{y}) \\ &= -\log\theta - e^v(\theta^{-1} + n\bar{y}) + (n+1)\{v + 1 - \log(n+1)\} + \log\Gamma(n+1), \end{aligned}$$

which gives the MLE $\widehat{\theta} = 1/\bar{y}$ and the MHLE $\widehat{v}_1 = -\log\bar{y}$. Here, the expected h-information is

$$\mathcal{I}_1(\theta, v) = \mathrm{E}\{I_1(\theta, v)\} = \mathrm{E}\left\{-\frac{\partial^2 h_1(\theta, v)}{\partial(\theta, v)\partial(\theta, v)^T}\right\} = \begin{pmatrix} \theta^{-2} & -\theta^{-1} \\ -\theta^{-1} & n+1 \end{pmatrix}$$

and the observed h-information leads to

$$\widehat{\mathrm{Var}}\begin{pmatrix} \theta - \widehat{\theta} \\ v - \widehat{v}_1 \end{pmatrix} = I_1(\widehat{\theta}_1, \widehat{v}_1)^{-1} = \begin{pmatrix} \bar{y}^2 & -\bar{y} \\ -\bar{y} & n+1 \end{pmatrix}^{-1} = \frac{1}{n}\begin{pmatrix} (n+1)\bar{y}^{-2} & \bar{y}^{-1} \\ \bar{y}^{-1} & 1 \end{pmatrix}.$$

Note here that $\mathrm{Var}(\theta - \widehat{\theta}) = \mathrm{Var}(\widehat{\theta})$ since θ is fixed, whereas $\mathrm{Var}(v - \widehat{v}_1) \neq \mathrm{Var}(\widehat{v}_1)$ since v is random. The h-information and the Fisher information of classical likelihood give the same variance estimation for MLE of θ. For the random parameters, as $n \to \infty$, we have $\mathrm{E}(v - \widehat{v}_1) = \log n - \psi(n) \to 0$ and its variance estimation $\widehat{\mathrm{Var}}(v - \widehat{v}_1) = 1/n \to \mathrm{Var}(v - \widehat{v}_1) = \psi^{(1)}(n)$ from the h-information, where $\psi(\cdot)$ and $\psi^{(1)}(\cdot)$ are digamma and trigamma functions, respectively. This asymptotically achieves the lower bound of Theorem 4.1,

$$\mathrm{Var}(v - \widehat{v}_1) = \psi^{(1)}(n) = \frac{1}{n} + O(n^{-2}) \geq \frac{1}{n} = (0, 1)\ \mathcal{I}_\theta^{-1}\ (0, 1)^T.$$

Thus, even though $\theta - \widehat{\theta} = O_p(1)$, we have

$$v - \widehat{v}_1 = o_p(1).$$

For given θ, the predictor $\widetilde{v}_1 = \log(n+1) - \log(\theta^{-1} + n\bar{y})$ leads to

$$E(u|\mathbf{y}) = (n+1)/(\theta^{-1} + n\bar{y}) = \widetilde{u}_1 = \exp(\widetilde{v}_1)$$

and

$$\mathrm{Var}(u|\mathbf{y}) = \mathrm{Var}(u - \widetilde{u}_1|\mathbf{y}) = \frac{n+1}{(\theta^{-1} + n\bar{y})^2} = \left(\frac{\partial^2 h_1(\theta, u)}{\partial u^2} \right)^{-1}_{u=\widetilde{u}_1}.$$

Thus, the h-likelihood $h_1(\theta, v)$ gives the BUP of u and the observed h-information gives $\mathrm{Var}(u - \widetilde{u}_1|\mathbf{y})$. Furthermore, the expected h-information gives

$$\begin{aligned} \mathrm{Var}(u - \widetilde{u}_1) &= \mathrm{E}\{\mathrm{Var}(u - \widetilde{u}_1|\mathbf{y})\} + \mathrm{Var}\{\mathrm{E}(u - \widetilde{u}_1|\mathbf{y})\} \\ &= \mathrm{E}\{\mathrm{Var}(u|\mathbf{y})\} = \mathrm{E}\left\{ \frac{n+1}{(\theta^{-1} + n\bar{y})^2} \right\} = \frac{2\theta^2}{n+2} = O(n^{-1}). \end{aligned}$$

This means that $\widetilde{u}_1$ is a consistent predictor of unobserved random variable u, which can be viewed as an unbiased estimator of $\theta = E(u)$. Thus, even though we cannot estimate θ consistently, $\widehat{\theta} - \theta = O_p(1)$, we can predict its unbiased estimator u consistently, $\widehat{u}_1 - u = o_p(1)$. For the conditional mean $\mu = \mathrm{E}(y_i|u)$, we have $\widetilde{\mu}_1 = 1/\widetilde{u} = (\theta^{-1} + n\bar{y})/(n+1)$ to give $\widehat{\mu}_1 = \bar{y}$.

Now we consider another scale of random effects. Note that Bartlizable scale is more general than the weak canonical scale. Though u-scale is not weak canonical, it is Bartlizable when $n \geq 2$ from Lemma 4.1. Thus, the

h-likelihood with u-scale can be defined,

$$h_2(\theta, v) = \ell_e(\theta, u) + c_2(\theta; y)$$
$$= -\log\theta - e^v(\theta^{-1} + n\bar{y}) - \log(\theta^{-1} + n\bar{y}) + n\{v + 1 - \log n\} + \log\Gamma(n+1),$$

which gives the MLE $\widehat{\theta} = \bar{y}$ and the MHLE $\widehat{v}_2 = -\log\bar{y} + \log\{n/(n+1)\}$. Here the observed h-information leads to

$$\widehat{\mathrm{Var}}\begin{pmatrix}\theta - \widehat{\theta}\\ v - \widehat{v}_2\end{pmatrix} = I_2(\widehat{\theta}, \widehat{v}_2)^{-1} = \frac{1}{n}\begin{pmatrix}(n+1)\bar{y}^{-2} & \bar{y}^{-1}\\ \bar{y}^{-1} & 1 + (n+1)^{-1}\end{pmatrix}$$
$$= \widehat{\mathrm{Var}}\begin{pmatrix}\theta - \widehat{\theta}\\ v - \widehat{v}_1\end{pmatrix} \cdot \{1 + O(n^{-1})\}.$$

Thus, both $h_1(\theta, v)$ and $h_2(\theta, v)$ yield asymptotically correct inferences. For given θ, the predictor $\widetilde{v}_2 = \log n - \log(\theta^{-1} + n\bar{y})$ leads to $\widetilde{u}_2 = \exp(\widetilde{v}_2) = n/(\theta^{-1} + n\bar{y})$ and

$$\widetilde{\mu}_2 = 1/\widetilde{u}_2 = (\theta^{-1} + n\bar{y})/n = E(\mu|y).$$

Then, we have

$$\mathrm{E}(\mu|\mathbf{y}) = \frac{\theta^{-1} + n\bar{y}}{n} = \widetilde{\mu}_2 = \exp(-\widetilde{v}_2)$$

and

$$\mathrm{Var}(\mu|\mathbf{y}) = \mathrm{Var}(\mu - \widetilde{\mu}_2|\mathbf{y}) = \frac{(\theta^{-1} + n\bar{y})^2}{n^2(n-1)} = \left(\frac{\partial^2 h_2(\theta, \mu)}{\partial\mu^2}\right)^{-1}_{\mu=\widetilde{\mu}_2}\left\{1 + O(n^{-1})\right\},$$

where $(\partial^2 h_2(\theta, \mu)/\partial\mu^2)^{-1}_{\mu=\widetilde{\mu}_2} = (\theta^{-1} + n\bar{y})^2/n^3$. Thus, the h-likelihood with u-

scale yields the BUP of conditional mean $\mu = 1/u$. It can be shown that

$$\mathrm{E}[(\mu - \widetilde{\mu}_1)^2 - (\mu - \widetilde{\mu}_2)^2 | \mathbf{y}] = \frac{(\theta^{-1} + n\bar{y})^2}{n^2(n+1)^2} > 0 \tag{4.10}$$

Thus, in finite samples, we expect that $h_2(\theta, v)$ provides a better prediction for μ than $h_1(\theta, v)$. This shows that in finite samples it is interesting to find a Bartlizable scale which gives the BUP for the random parameter of interest.

This example becomes Bayarri et al.'s (1988) example when $n = 1$ with a parameterization $\xi = 1/\theta$. Here, care is necessary since $\widehat{v} - v = O_p(1)$ and $\widehat{\theta} - \theta = O_p(1)$. When $n = 1$, u-scale is not Bartlizable since

$$\mathrm{E}\left[\frac{\partial h_2(\theta, u)}{\partial u}\middle| y\right] = \mathrm{E}\left(u^{-1}|y\right) - (\theta^{-1} + y) = 0$$

but

$$\mathrm{E}\left[\mathrm{E}\left[\frac{\partial^2 h_2(\theta, u)}{\partial u^2} + \left\{\frac{\partial h_2(\theta, u)}{\partial u}\right\}^2 \middle| y\right]\right] = \mathrm{E}\left[(\theta^{-1} + y)^2 - 2\right] \neq 0.$$

Here we can find the MLE $\widehat{\theta} = 1/y$ and the MHLE $\widehat{u} = 1/2y$, but their expectations and variances are infinity. However, given θ and y, the inequality (4.10) shows that u-scale may still have an advantage for prediction of μ. When $n \geq 2$, we have the classical log-likelihood,

$$\ell(\xi) = \log \xi - (n+1) \log(\xi + n\bar{y}) + \log \Gamma(n+1).$$

Then, the MLE of the parameter ξ is given by $\widehat{\xi} = \bar{y}$ with the expectation

$$\mathrm{E}(\widehat{\xi}) = \mathrm{E}(\bar{y}) = \mathrm{E}(\mathrm{E}(\bar{y}|u)) = \infty$$

and the variance

$$\mathrm{Var}(\widehat{\xi}) = \mathrm{Var}(\bar{y}) = \mathrm{E}(\mathrm{Var}(\bar{y}|u)) + \mathrm{Var}(\mathrm{E}(\bar{y}|u)) = \infty.$$

Here, the meaning of the MLE for ξ could be controversial like that of Cauchy distribution. Thus, we prefer a parameterization θ. When $\widehat{\theta} - \theta = O_p(1)$, the scale of fixed parameter is also important for inferences like the scale of random parameters.

4.5 When the h-likelihood is not explicit

In the h-likelihood (4.4), $c(\boldsymbol{\theta}; \mathbf{y})$ may not have an explicit form. Han and Lee (2022) proposed a Monte-Carlo method for approximating the classical likelihood $L(\theta; \mathbf{y})$ by

$$L_B(\theta; \mathbf{y}) = \frac{1}{B} \sum_{b=1}^{B} \frac{L_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{q(\mathbf{v}^{(b)})},$$

where $\mathbf{v}^{(b)}$ are independent samples from a probability density function $q(\mathbf{v}^{(b)})$, having the same support with $\mathbf{v}$. When $c(\boldsymbol{\theta}; \mathbf{y})$ is not explicitly known, we use an approximated h-likelihood,

$$h_B(\boldsymbol{\theta}, \mathbf{v}) = \ell_e(\boldsymbol{\theta}, \mathbf{v}) - \ell_e(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) + \log L_B(\theta; \mathbf{y}), \tag{4.11}$$

and approximated h-information,

$$I_B(\boldsymbol{\theta}, \mathbf{v}) = -\frac{\partial^2 \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})\partial(\boldsymbol{\theta}, \mathbf{v})^T} + \frac{\partial^2 \ell_e(\boldsymbol{\theta}, \widetilde{\mathbf{v}})}{\partial(\boldsymbol{\theta}, \mathbf{v})\partial(\boldsymbol{\theta}, \mathbf{v})^T} - \begin{pmatrix} I_{11}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $w_b = \frac{L_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})/q(\mathbf{v}^{(b)})}{\sum_{b=1}^{B} L_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})/q(\mathbf{v}^{(b)})}$ and

$$\begin{aligned} I_{11}(\boldsymbol{\theta}) = & \left[\sum_{b=1}^{B} \left\{ w_b \cdot \frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{\partial \boldsymbol{\theta}} \right\}\right] \left[\sum_{b=1}^{B} \left\{ w_b \cdot \frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{\partial \boldsymbol{\theta}} \right\}\right]^T \\ & - \left[\sum_{b=1}^{B} w_b \left\{ \left(\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{\partial \boldsymbol{\theta}} \right)^T + \left(\frac{\partial^2 \ell_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) \right\}\right]. \end{aligned}$$

This leads to the following theorem.

Theorem 4.5. *Under the regularity conditions in Appendix,*

$$\begin{bmatrix} \widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta} \\ \widehat{\mathbf{v}}_B - \mathbf{v} \end{bmatrix} \xrightarrow{d} N\left(\mathbf{0},\ \mathcal{I}_{\boldsymbol{\theta}}^{-1}\right),$$

hence the approximate MHLEs are asymptotically the best.

4.6 Appendix

Proof of Lemma 4.1

(i) It is enough to investigate the derivatives with respect to $\mathbf{v}$. Since $f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y}) = 0$ at the boundary, we have

$$\mathrm{E}\left\{ \frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})}{\partial \mathbf{v}} \Big| \mathbf{y} \right\} = \int_{\Omega_{\mathbf{v}}} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})}{\partial \mathbf{v}} d\mathbf{v} = \mathbf{0},$$

which leads to the first Bartlett identity

$$\mathrm{E}\left[\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial \mathbf{v}}\right] = \mathrm{E}\left[\frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})}{\partial \mathbf{v}}\right] = \mathrm{E}\left[\mathrm{E}\left\{\frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})}{\partial \mathbf{v}}\Big|\mathbf{y}\right\}\right] = \mathbf{0}.$$

Using the similar argument, we have

$$\mathrm{E}\left\{\frac{\partial^2 \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})}{\partial v_i \partial v_j}\Big|\mathbf{y}\right\} = \int_{\Omega_\mathbf{v}} \frac{\partial}{\partial v_i}\left(\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})}{\partial v_j}\right) d\mathbf{v} = 0,$$

for any $i, j = 1, ..., q$. This leads to the second Bartlett identity.

(ii) Consider $v(u) = \mathrm{logit} F_{\boldsymbol{\theta}}(u) = \log F_{\boldsymbol{\theta}}(u) - \log(1 - F_{\boldsymbol{\theta}}(u))$ where $F_{\boldsymbol{\theta}}(u)$ denotes the cumulative distribution function of u, then $v(\cdot)$ is an increasing function from Ω_u to $(-\infty, \infty)$. The density function of v is given by

$$f_{\boldsymbol{\theta}}(v) = f_{\boldsymbol{\theta}}(u)\left|\frac{du}{dv}\right| = f_{\boldsymbol{\theta}}(u)\left|\frac{f_{\boldsymbol{\theta}}(u)}{F_{\boldsymbol{\theta}}(u)} + \frac{f_{\boldsymbol{\theta}}(u)}{1 - F_{\boldsymbol{\theta}}(u)}\right|^{-1} = F_{\boldsymbol{\theta}}(u)(1 - F_{\boldsymbol{\theta}}(u))$$

and its derivative is

$$f'_{\boldsymbol{\theta}}(v) = \frac{du}{dv}\frac{d}{du}\left\{F_{\boldsymbol{\theta}}(u)(1 - F_{\boldsymbol{\theta}}(u))\right\} = F_{\boldsymbol{\theta}}(u)(1 - F_{\boldsymbol{\theta}}(u))(1 - 2F_{\boldsymbol{\theta}}(u)).$$

Since $F_{\boldsymbol{\theta}}(u)$ is cumulative distribution function that increases from 0 to 1 and $v(u)$ is an increasing function of u, we have $f_{\boldsymbol{\theta}}(v) = f'_{\boldsymbol{\theta}}(v) = 0$ at the boundary $v = \pm\infty$. This satisfies a sufficient condition for the first and second Bartlett identities (Meng, 2009), hence we can always find at least one Bartlizable transformation.

(iii) Now suppose that the probability density function $f_{\boldsymbol{\theta}}(v)$ is differentiable with respect to u for any u in the support $\Omega_v = (-\infty, \infty)$. Since $f_{\boldsymbol{\theta}}(v) > 0$

and $\int_{-\infty}^{\infty} f(v)dv = 1$, we have $f(v) = 0$ at the boundary $v = \pm\infty$. Then we also have $f'(v) = 0$ at the boundary $v = \pm\infty$ since $\int_{-\infty}^{t} f'(v)dv < \infty$ for any $t \in \mathbb{R}$. Thus, v achieves the sufficient condition for the Bartlett identities.

Proof of Theorem 4.1

We assume the following regularity conditions. Throughout the Appendix, it is also assumed that the continuity, differentiability and integrability hold whenever needed.

(R1) $f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v})$ and its first and second derivatives are absolutely integrable with respect to $\mathbf{y}$ and $\mathbf{v}$.

(R2) $c(\boldsymbol{\theta}, \mathbf{y})$ does not depend on $\mathbf{y}$ or $(\partial^2 c(\boldsymbol{\theta}, \mathbf{y}))/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$ is positive semi-definite.

(R3) $\widehat{\boldsymbol{\zeta}}(\mathbf{y})$ is an unbiased estimator of $\boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})$ such that

$$\mathrm{E}\left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})\right] = 0. \tag{4.12}$$

with

$$\lim_{\mathbf{v}\to\partial\Omega_{\mathbf{v}}} \int_{\Omega_{\mathbf{y}}} \left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})\right] f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v})d\mathbf{y} = 0 \tag{4.13}$$

Note first that

$$\begin{aligned}
&\frac{\partial}{\partial(\boldsymbol{\theta}, \mathbf{v})}\left[\int_{\Omega_{\mathbf{y}}} \left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})\right] f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v})d\mathbf{y}\right] \\
&= -\int_{\Omega_{\mathbf{y}}} \frac{\partial\boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v})d\mathbf{y} + \int_{\Omega_{\mathbf{y}}} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})\right] d\mathbf{y}
\end{aligned}$$

Integrating with respect to $\mathbf{v}$ leads to

$$\text{LHS} = \int_{\Omega_\mathbf{v}} \frac{\partial}{\partial(\boldsymbol{\theta}, \mathbf{v})} \left[\int_{\Omega_\mathbf{y}} \left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v}) \right] f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v}) d\mathbf{y} \right] d\mathbf{v} = 0$$

from the assumption (4.12) and (4.13). Thus, we have

$$\begin{aligned}
\text{E}\left[\frac{\partial \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \right] &= \iint \frac{\partial \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v}) d\mathbf{y} d\mathbf{v} \\
&= \iint \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v}) \right] d\mathbf{y} d\mathbf{v} \\
&= \iint \frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v}) \right] f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v}) d\mathbf{y} d\mathbf{v}.
\end{aligned}$$

Note that

$$\begin{aligned}
\text{Var}\left[\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \right] &= \text{E}\left[\left(\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \right)^T \left(\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \right) \right] \\
&= \text{E}\left[-\frac{\partial^2 \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v}) \partial(\boldsymbol{\theta}, \mathbf{v})^T} \right] \\
&\geq \text{E}\left[-\frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v}) \partial(\boldsymbol{\theta}, \mathbf{v})^T} \right]
\end{aligned}$$

and

$$\begin{aligned}
\text{Cov}\left[\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})}, \widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v}) \right] &= \text{E}\left[\left\{ \widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v}) \right\} \left(\frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \right) \right] \\
&= \iint \frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v}) \right] f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v}) d\mathbf{y} d\mathbf{v} \\
&= \text{E}\left[\frac{\partial \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})} \right].
\end{aligned}$$

By the multivariate Cauchy-Schwartz inequality, we have

$$\mathrm{Var}\left[\widehat{\boldsymbol{\zeta}}(\mathbf{y}) - \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})\right] \geq \mathrm{E}\left[\frac{\partial \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})}\right] \mathrm{E}\left[-\frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})\partial(\boldsymbol{\theta}, \mathbf{v})^T}\mathrm{E}\right]^{-1} \left[\frac{\partial \boldsymbol{\zeta}(\boldsymbol{\theta}, \mathbf{v})}{\partial(\boldsymbol{\theta}, \mathbf{v})}\right]^T$$

□

Proof of Theorem 4.2

Asymptotic normality can be easily proved by the central limit theorem and the Bartlett identities. Note that the maximum h-likelihood estimator (MHLE) $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{v})$ adapts the asymptotic normality of the MLE,

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N\left(0, I_m(\boldsymbol{\theta})^{-1}\right),$$

where $I_m(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]$ is the expected Fisher information. Thus, the variance of $\widehat{\boldsymbol{\theta}}$ can be estimated by using the observed information,

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) = \left[-\frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1} = \left[-\frac{\partial^2 h(\boldsymbol{\theta}, \widetilde{\mathbf{v}}(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}.$$

By the definition of $\widetilde{\mathbf{v}}$, we have $\frac{\partial h(\boldsymbol{\theta},\mathbf{v})}{\partial \mathbf{v}}\big|_{\mathbf{v}=\widetilde{\mathbf{v}}} = 0$. Then the chain rule leads to

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}} &= \frac{\partial h(\boldsymbol{\theta}, \widetilde{\mathbf{v}}(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}} \\ &= \left(\frac{\partial h(\boldsymbol{\theta},\mathbf{v})}{\partial\boldsymbol{\theta}}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\right) + \left(\frac{\partial \widetilde{\mathbf{v}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial h(\boldsymbol{\theta},\mathbf{v})}{\partial\mathbf{v}}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\right) = \frac{\partial h(\boldsymbol{\theta},\mathbf{v})}{\partial\boldsymbol{\theta}}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\end{aligned}$$

and the Hessian matrix of $\ell(\boldsymbol{\theta};\mathbf{y})$,

$$\frac{\partial^2\ell(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}=\frac{\partial^2 h(\boldsymbol{\theta},\widetilde{\mathbf{v}}(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}=\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\right)+\left(\frac{\partial\widetilde{\mathbf{v}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\mathbf{v}\partial\boldsymbol{\theta}^T}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\right).$$

From the fact that

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial h(\boldsymbol{\theta},\mathbf{v})}{\partial\mathbf{v}^T}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\right)=\left(\frac{\partial\widetilde{\mathbf{v}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\mathbf{v}\partial\mathbf{v}^T}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\right)+\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\boldsymbol{\theta}\partial\mathbf{v}^T}\Big|_{\mathbf{v}=\widetilde{\mathbf{v}}}\right)=0,$$

we have an estimator

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}})=\left[-\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)+\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\boldsymbol{\theta}\partial\mathbf{v}^T}\right)\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\mathbf{v}\partial\mathbf{v}^T}\right)^{-1}\left(\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\mathbf{v}\partial\boldsymbol{\theta}^T}\right)\right]^{-1}_{\mathbf{v}=\widetilde{\mathbf{v}}}$$

which is the submatrix of the inverse of Hessian matrix of $h(\boldsymbol{\theta},\mathbf{v})$ with respect to the whole fixed and random parameters $(\boldsymbol{\theta},\mathbf{v})$. From the convergence of MLE, $\widehat{\boldsymbol{\theta}}\to\boldsymbol{\theta}$ as $n\to\infty$, it can be shown that $\widehat{\mathbf{v}}=\widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}};\mathbf{y})\to\widetilde{\mathbf{v}}(\boldsymbol{\theta};\mathbf{y})$. Furthermore, $\widetilde{\mathbf{v}}(\boldsymbol{\theta};\mathbf{y})\to\mathrm{E}(\mathbf{v}|\mathbf{y})$ since $\mathrm{Var}(\mathbf{v}|\mathbf{y})\to 0$ as each $n_i\to\infty$. Thus, the MHL predictor $\widehat{\mathbf{v}}=\widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}})$ converges to the best unbiased predictor,

$$\widehat{\mathbf{v}}\to\mathrm{E}_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})\quad\text{as } n_i\to\infty \text{ for } i=1,...,q.$$

The variance of the MHL predictor $\widehat{\mathbf{v}}$ can be expressed as

$$\mathrm{Var}(\widehat{\mathbf{v}}-\mathbf{v})=\mathrm{Var}\left(\widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}})-\widetilde{\mathbf{v}}(\boldsymbol{\theta})+\widetilde{\mathbf{v}}(\boldsymbol{\theta})-\mathbf{v}\right).$$

By the delta method, $\widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}}) \approx \widetilde{\mathbf{v}}(\boldsymbol{\theta}) + (\partial\widetilde{\mathbf{v}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})^T (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ implies that

$$\mathrm{Var}\left(\widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}}) - \widetilde{\mathbf{v}}(\boldsymbol{\theta})\right) \approx \left(\frac{\partial\widetilde{\mathbf{v}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) \mathrm{Var}(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial\widetilde{\mathbf{v}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T,$$

hence $\mathrm{Var}(\widehat{\mathbf{v}} - \mathbf{v})$ can be estimated by

$$\widehat{\mathrm{Var}}(\widehat{\mathbf{v}} - \mathbf{v}) = \mathbf{D}^{-1}\mathbf{B}^T(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T)^{-1}\mathbf{B}\mathbf{D}^{-1} + \mathbf{D}^{-1}$$

where

$$\mathbf{A} = \frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}, \quad \mathbf{B} = \frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial\boldsymbol{\theta}\partial\mathbf{v}^T}, \quad \text{and} \quad \mathbf{D} = \frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial\mathbf{v}\partial\mathbf{v}^T}.$$

Note here that $\widehat{\mathrm{Var}}(\widehat{\mathbf{v}} - \mathbf{v})$ is the submatrix of the inverse of Hessian matrix of $h(\boldsymbol{\theta}, \mathbf{v})$ with respect to $(\boldsymbol{\theta}, \mathbf{v})$. Similarly, the covariance component becomes

$$\mathrm{Cov}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}), (\widehat{\mathbf{v}} - \mathbf{v})\right] \approx \mathrm{Cov}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \left\{(\widetilde{\mathbf{v}} - \mathbf{v}) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\left(\frac{\partial\widetilde{\mathbf{v}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\right\}\right],$$

which can be estimated by

$$\begin{aligned}\widehat{\mathrm{Cov}}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}), (\widehat{\mathbf{v}} - \mathbf{v})\right] &= \left(\frac{\partial\widetilde{\mathbf{v}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T \widehat{\mathrm{E}}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] \\ &= -\mathbf{D}^{-1}\mathbf{B}^T(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T)^{-1}.\end{aligned}$$

□

Proof of Theorem 4.3

($\Leftarrow$) By the Taylor expansion and (4.8),

$$\begin{aligned}
\mathrm{E}\left[g^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+v_i)|\mathbf{y}\right] &= \mathrm{E}\left[\sum_{k=1}^{\infty}\frac{g_{(k)}^{-1}(v_i)}{k!}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^k\middle|\mathbf{y}\right] \\
&= \sum_{k=1}^{\infty}\frac{\mathrm{E}\left(g_{(k)}^{-1}(v_i)|\mathbf{y}\right)}{k!}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^k \\
&= \sum_{k=1}^{\infty}\frac{g_{(k)}^{-1}(c_i)}{k!}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^k = g^{-1}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}+c_i\right).
\end{aligned}$$

Thus, taking $\widehat{v}_i = c_i$ proves the existence of $\widehat{\mathbf{v}}$ which gives the BUP for each μ_{ij}.

($\Rightarrow$) Since the equation (4.7) hold for any value of $\mathbf{x}_{ij}^T\boldsymbol{\beta} \in \mathbb{R}$, substituting $\mathbf{x}_{ij}^T\boldsymbol{\beta} = t$ and $c = \widehat{v}_i$ implies that

$$\mathrm{E}\left(g^{-1}(v_i+t)|\mathbf{y}\right) = g^{-1}(c+t) \quad \forall t \in \mathbb{R}.$$

If the equation $\mathrm{E}\left(g_{(k-1)}^{-1}(v_i+t)|\mathbf{y}\right) = g_{(k-1)}^{-1}(c+t)$ holds for all $t \in \mathbb{R}$, then

$$\begin{aligned}
\mathrm{E}\left(g_{(k)}^{-1}(v_i+t)|\mathbf{y}\right) &= \mathrm{E}\left[\lim_{\epsilon\to 0}\frac{g_{(k-1)}^{-1}(v_i+t+\epsilon)-g_{(k-1)}^{-1}(v_i+t)}{\epsilon}\middle|\mathbf{y}\right] \\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[\mathrm{E}\left(g_{(k-1)}^{-1}(v_i+t+\epsilon)|\mathbf{y}\right)-\mathrm{E}\left(g_{(k-1)}^{-1}(v_i+t)|\mathbf{y}\right)\right] \\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[g_{(k-1)}^{-1}(c+t+\epsilon)-g_{(k-1)}^{-1}(c+t)\right] \\
&= g_{(k)}^{-1}(c+t).
\end{aligned}$$

By induction, it is proved that for any $k = 0, 1, 2, ...,$

$$\mathrm{E}\left(g_{(k)}^{-1}(v_i + t)|\mathbf{y}\right) = g_{(k)}^{-1}(c + t) \quad \forall t \in \mathbb{R}.$$

Substituting $t = 0$ ends the proof.

Proof of Theorem 4.4

We consider the following regularity conditions.

(R1) Let $\theta_0 = \mathrm{argmax}_\theta \mathrm{E}_\theta\{\ell_m(\theta)\}$ be the true value of θ. Here, the number of fixed parameters does not depend on n_{obs}. Then, the MLE $\widehat{\theta} = \arg\max_\theta \ell_m(\theta)$ satisfies the asymptotic normality with mean θ_0 and variance $\mathcal{I}_0^{-1} = \mathcal{I}^{-1}(\theta_0)$, where

$$\mathcal{I}(\theta) = \lim_{n_{\mathrm{obs}} \to \infty} \frac{1}{n_{\mathrm{obs}}} \left(-\frac{\partial^2 \ell_m(\theta)}{\partial\theta\partial\theta^T}\right)\Big|_{\theta=\theta_0}$$

is the expected Fisher information.

(R2) The support of missing values

$$\Omega_u = \left\{ u \in \mathbb{R}^{n_{\mathrm{mis}}} : \prod_{i=n_{\mathrm{obs}}+1}^{n} f_\theta(u_i, \delta_i = 0|\boldsymbol{x}_i) > 0 \right\} \subset \mathbb{R}^{n_{\mathrm{mis}}}$$

does not depend on fixed parameter θ.

Proof of Theorem 4.5

Lemma 4.2. *Suppose that* $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})$ *is a MHLE from the h-likelihood and* $(\widehat{\boldsymbol{\theta}}_B, \widehat{\mathbf{v}}_B)$ *is an approximate MHLE from the approximated h-likelihood* (4.11)*. Then, as*

$B \to \infty$,

$$(\widehat{\boldsymbol{\theta}}_B, \widehat{\mathbf{v}}_B) \xrightarrow{p} (\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}}) \quad \textit{and} \quad I_B(\widehat{\boldsymbol{\theta}}_B, \widehat{\mathbf{v}}_B) \xrightarrow{p} I(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}}).$$

Proof. We assume the following regularity conditions in Han and Lee (2023):

(R1) There exists a compact set $\boldsymbol{\Theta}$ which contains the true value $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$.

(R2) $H(\boldsymbol{\theta}, \mathbf{v})$ is a smooth function which is log-concave and has continuous second derivatives with respect to $(\boldsymbol{\theta}, \mathbf{v})$.

(R3) There exists a function $M(\mathbf{y}, \mathbf{v})$ such that

$$\mathrm{E}\left[M(\mathbf{y}, \mathbf{v})|\mathbf{y}\right] = \int_{\Omega_\mathbf{v}} M(\mathbf{y}, \mathbf{v}) \widehat{L}_p(\mathbf{v}|\mathbf{y}) d\mathbf{v} < \infty$$

and $M(\mathbf{y}, \mathbf{v}) > H(\boldsymbol{\theta}, \mathbf{v}) / \widehat{L}_p(\mathbf{v}|\mathbf{y})$ for all $(\boldsymbol{\theta}, \mathbf{v})$.

Han and Lee (2023) showed that

$$\frac{1}{B} \sum_{b=1}^{B} \frac{L_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{q(\mathbf{v}^{(b)})} \xrightarrow{p} L(\boldsymbol{\theta}; \mathbf{y}) \quad \text{and} \quad I_{11}(\widehat{\boldsymbol{\theta}}_B) \xrightarrow{p} - \left\{ \frac{\partial^2 \log f_{\boldsymbol{\theta}}(\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}},$$

Since $\log f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}|\mathbf{y}) = \log f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}|\mathbf{y}) + \log f_{\boldsymbol{\theta}}(\mathbf{y}) - \log f_{\boldsymbol{\theta}}(\mathbf{y}) = \ell_e(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) + \ell_m(\boldsymbol{\theta})$, we have

$$h_B(\boldsymbol{\theta}, \mathbf{v}) \xrightarrow{p} \ell_e(\boldsymbol{\theta}, \mathbf{v}) - \ell_e(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) + \ell_m(\boldsymbol{\theta}) = \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y}) - \log f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}|\mathbf{y}) + \log f_{\boldsymbol{\theta}}(\mathbf{y}) = h(\boldsymbol{\theta}, \mathbf{v}).$$

This leads to

$$\widehat{\boldsymbol{\theta}}_B = \arg\max_{\boldsymbol{\theta}} h_B(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) = \arg\max_{\boldsymbol{\theta}} \left\{ \frac{1}{B} \sum_{b=1}^{B} \frac{L_e(\boldsymbol{\theta}, \mathbf{v}^{(b)})}{q(\mathbf{v}^{(b)})} \right\} \xrightarrow{p} \arg\max_{\boldsymbol{\theta}} L_m(\boldsymbol{\theta}) = \widehat{\boldsymbol{\theta}}$$

and $\widehat{\mathbf{v}}_B = \widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}}_B) \xrightarrow{p} \widetilde{\mathbf{v}}(\widehat{\boldsymbol{\theta}}) = \widehat{\mathbf{v}}$. Furthermore, convergence of $I_{11}(\widehat{\boldsymbol{\theta}}_B)$ leads to

$$\begin{aligned}
&\left[-\frac{\partial^2 \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \frac{\partial^2 \ell_e(\boldsymbol{\theta}, \widetilde{\mathbf{v}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - I_{11}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_B, \mathbf{v}=\widehat{\mathbf{v}}_B} \\
&\xrightarrow{p} \left[-\frac{\partial^2 \{\ell_e(\boldsymbol{\theta}, \mathbf{v}) - \ell_e(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) + \ell_m(\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \mathbf{v}=\widehat{\mathbf{v}}} \\
&= -\left\{\frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right\}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \mathbf{v}=\widehat{\mathbf{v}}}.
\end{aligned}$$

Since $\partial h(\boldsymbol{\theta}, \mathbf{v})/\partial \mathbf{v} = \partial \ell_e(\boldsymbol{\theta}, \mathbf{v})/\partial \mathbf{v}$, we have $I_B(\widehat{\boldsymbol{\theta}}_B, \widehat{\mathbf{v}}_B) \xrightarrow{p} I(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})$. □

Note here that

$$\frac{\partial h(\boldsymbol{\theta}, \mathbf{v})}{\partial \mathbf{v}} = \frac{\partial \ell_e(\boldsymbol{\theta}, \mathbf{v})}{\partial \mathbf{v}} = \frac{\partial h_B(\boldsymbol{\theta}, \mathbf{v})}{\partial \mathbf{v}}.$$

Thus, it is enough to consider the derivatives with respect to $\boldsymbol{\theta}$ only. Han and Lee (2023) showed that $\widetilde{L}_B(\boldsymbol{\theta}) \xrightarrow{p} \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{y})$ and substituting $\mathbf{v} = \widetilde{\mathbf{v}}$ leads to $h_B(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) = \log \widetilde{L}_B(\boldsymbol{\theta})$. Thus, we have

$$\widehat{\boldsymbol{\theta}}_B = \arg\max_{\boldsymbol{\theta}} h_B(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) = \arg\max_{\boldsymbol{\theta}} \widetilde{L}_B(\boldsymbol{\theta}) \xrightarrow{p} \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{y}) = \widehat{\boldsymbol{\theta}}.$$

For the second derivatives, by using the theorem in Han and Lee (2023), we can show that

$$I_{11}(\widehat{\boldsymbol{\theta}}_B) \xrightarrow{p} \left(\frac{\partial^2 \ell_e(\boldsymbol{\theta}, \widetilde{\mathbf{v}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}.$$

This leads to

$$\widetilde{I}_B(\widehat{\boldsymbol{\theta}}_B, \widehat{\mathbf{v}}_B) \xrightarrow{p} I(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}}),$$

and $\widehat{I}$ is a consistent estimator of $\mathcal{I}_{\boldsymbol{\theta}}$ from the Theorem 4.2. □

Chapter 5

DNN with temporal-spatial random effects via h-likelihood

5.1 Introduction

Deep neural network (DNN) models have served as the method of learning highly nonlinear relationship between the input and output variables with strong prediction performance (Goodfellow et al., 2016; LeCun et al., 2015). However, most DNN models implicitly assume independence of the data and ignore underlying correlation structures, despite large-scale data in the real world often being clustered by multiple categorical features. Recently, there have been emerging attempts to enhance the prediction for clustered data by introducing the random effects into the DNN models (Mandel et al., 2021; Simchoni and Rosset, 2021, 2023; Tran et al., 2020).

Simchoni and Rosset (2021, 2023) proposed linear mixed model neural network (LMMNN) models with single independent random effects and extended

LMMNN models to multiple random effects allowing temporal-spatial correlation structure. However, their conventional integrated likelihood approach is computationally intractable because it does not allow decomposition like an ordinary loss function in DNNs. They proposed the use of block-diagonal approximation to the covariance matrix to obtain approximate maximum likelihood estimators (MLEs) for their LMMNN models. However, their approximate likelihood can give a severe bias in parameter estimation for models with correlated random effects. Also, this difficulty prevents them from obtaining restricted maximum likelihood estimators (REMLEs) for LMMNN models. Variational approach can be an alternative. However, this cannot provide exact MLEs either but only approximate MLEs.

Lee and Nelder (1996) proposed the use of h-likelihood as an extension of classical likelihood for statistical models with random effects. In LMMs, the h-likelihood is Henderson's joint likelihood (Henderson et al., 1959) of which the joint maximization gives the MLEs for fixed effects and the best linear unbiased predictors (BLUPs) for random effects. However, it does not give MLEs for variance components by a simple joint maximization. This causes the computational difficulty of Simchoni and Rosset (2021, 2023). In this chapter, we introduce the new h-likelihood for LMMNN models with various temporal-spatial random effects from the multiple categorical features. The proposed negative h-likelihood serves as a loss function, which allows the exact MLEs for all fixed parameters and BLUPs for random effects. The proposed negative h-likelihood for LMMNN models allows the highly non-linear functions of input variables and multiple random effects with complex covariance structures, which is the key to overcoming the computational difficulties in LMMNN

models.

In Section 5.2, we briefly review the integrated likelihood approach to LMMs. In Section 5.3, the h-likelihood for LMMs with multiple random effects is proposed. It is worth emphasizing that its simple joint maximization can give the MLEs for the whole fixed parameters and BLUPs for random effects, and bypasses the heavy computation difficulties to obtain the exact MLEs. In Section 5.4, we propose the use of negative h-likelihood as a loss function of LMMNN models and introduce a useful adjustment for random effect predictions. This allows online learning algorithm. To compare with the existing methods, we provide simulation studies in Section 5.5 and real data analyses in Section 5.6, followed by concluding remarks in Section 5.7. All the proofs and technical details are in Appendix.

5.2 Integrated likelihood approach for LMMs

Let $\mathbf{y}$ be a vector of N responses, $\mathbf{X}$ and $\mathbf{Z}$ be $N \times p$ and $N \times q$ model matrices for fixed effects $\boldsymbol{\beta} \in \mathbb{R}^p$ and random effects $\mathbf{v} \in \mathbb{R}^q$, respectively. We start with a standard LMM,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

where $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$ is a vector of N random noises, $\mathbf{v} \sim N(\mathbf{0}, \mathbf{D})$ is a vector of q random effects, $\mathbf{I}_N$ is $N \times N$ identity matrix and $\mathbf{D} = \mathbf{D}(\boldsymbol{\lambda})$ is $q \times q$ covariance matrix parameterized by $\boldsymbol{\lambda}$. Let $\boldsymbol{\psi} = (\sigma_e^2, \boldsymbol{\lambda})$ be the vector of dispersion parameters and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi})$ be the vector of whole fixed parameters. To obtain the estimates for $\boldsymbol{\beta}$ and $\mathbf{v}$, Henderson et al. (1959) proposed to

maximize the Henderson's joint likelihood,

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}, \mathbf{v}) &= \log f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v}) = \log f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}) + \log f_{\boldsymbol{\theta}}(\mathbf{v}) \\ &= -\frac{1}{2\sigma_e^2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}||^2 - \frac{N}{2}\log(2\pi\sigma_e^2) - \frac{1}{2}\mathbf{v}^T\mathbf{D}^{-1}\mathbf{v} - \frac{1}{2}\log|2\pi\mathbf{D}|,\end{aligned} \tag{5.1}$$

where $||\cdot||^2$ denotes the L2-norm and $|\cdot|$ denotes the determinant. For given variance components $\boldsymbol{\psi} = (\sigma_e^2, \boldsymbol{\lambda})$, optimization of the joint likelihood (5.1) gives MLEs for $\boldsymbol{\beta}$ and the BLUPs for $\mathbf{v}$,

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{Z}\widehat{\mathbf{v}}), \\ \widehat{\mathbf{v}} &= \widehat{\mathrm{E}}(\mathbf{v}|\mathbf{y}) = (\mathbf{Z}^T\mathbf{Z} + \sigma_e^2\mathbf{D}^{-1})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}^T\mathbf{X}\widehat{\boldsymbol{\beta}}).\end{aligned}$$

However, it cannot give MLEs for the variance components $\boldsymbol{\psi}$. For the MLEs of $\boldsymbol{\psi}$, the integrated likelihood has been used from the multivariate normal distribution of $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$,

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log \int \exp\left(\mathcal{J}(\boldsymbol{\theta}, \mathbf{v})\right) d\mathbf{v} \\ &= -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\log|2\pi\mathbf{V}|,\end{aligned}$$

where the marginal covariance matrix $\mathbf{V}$ is

$$\mathbf{V} = \mathbf{V}(\boldsymbol{\psi}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \sigma_e^2\mathbf{I}_N.$$

For given variance components, it is known that the MLEs for $\boldsymbol{\beta}$ from the integrated likelihood $\ell(\boldsymbol{\theta})$ is the same as Henderson's MLE for $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{Z}\widehat{\mathbf{v}}).$$

In LMMs, MLEs for variance components could be biased in finite sample. To reduce the bias, REMLEs for $\boldsymbol{\psi}$ are often used (Patterson and Thompson, 1971). In LMMs, REMLEs maximize the restricted likelihood,

$$\ell_R(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}; \widehat{\boldsymbol{\beta}}) - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|, \tag{5.2}$$

which is an adjusted profile likelihood (Cox and Reid, 1987; Lee et al., 2017). However, both the integrated likelihood $\ell(\boldsymbol{\theta})$ and the restricted likelihood $\ell_R(\boldsymbol{\psi})$ involve the computation of the inverse of $N \times N$ matrix $\mathbf{V}$. In LMMNNs with single independent random effects of Simchoni and Rosset (2021), $\mathbf{V}$ has a block-diagonal form. This allows computation of exact MLEs. Simchoni and Rosset (2023) noted that $\mathbf{V}$ is not a block-diagonal form in general, even for LMMs with single categorical feature, when the random effects have a complex correlation structure. In order to avoid computing $\mathbf{V}^{-1}$, they proposed the use of block-diagonal approximation to $\mathbf{V}$. However, it requires a rigorous theoretical justification and the resulting approximate MLEs can have severe biases.

Further difficulties arise when the model contains multiple categorical features $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_K)$ with corresponding random effects $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_K)$,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{v}_1 + \cdots + \mathbf{Z}_K\mathbf{v}_K + \mathbf{e}, \tag{5.3}$$

where $\mathbf{v}_k \sim N(\mathbf{0}, \mathbf{D}_k)$ is q_k-dimensional vector for $k = 1, ..., K$. Simchoni and Rosset (2023) claimed that the use of block-diagonal approximation can avoid heavy computation in the inverse of $N \times N$ matrix. We found that the integrated likelihood can be computed by using the Woodbury formula,

$$\mathbf{V}^{-1} = (\mathbf{Z}\mathbf{D}\mathbf{Z}^T + \sigma_e^2 \mathbf{I}_N)^{-1} = \frac{1}{\sigma_e^2}\left[I_N - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + \sigma_e^2 \mathbf{D}^{-1})^{-1}\mathbf{Z}^T\right],$$

and the matrix determinant lemma,

$$\begin{aligned}\log|\mathbf{V}| &= \log|\mathbf{Z}\mathbf{D}\mathbf{Z}^T + \sigma_e^2\mathbf{I}_N| \\ &= \log|\mathbf{Z}^T\mathbf{Z}\mathbf{D} + \sigma_e^2 I_Q| + (N-Q)\log\sigma_e^2,\end{aligned}$$

where $\mathbf{D} = \text{block-diag}(\mathbf{D}_1, ..., \mathbf{D}_K)$. This formulation can reduce the computations of integrated likelihood without any approximations. However, $\mathbf{Z}^T\mathbf{Z}$ is not a block-diagonal matrix when $k \neq 1$. Thus, it still requires heavy computation for every mini-batch. We study how the h-likelihood overcomes the computational difficulties of an integrated likelihood approach.

5.3 H-likelihood approach for LMMs

In Henderson's joint likelihood, $\mathbf{v}$ is additive to the fixed effects $\boldsymbol{\beta}$ in the linear predictor of LMMs

$$\mathrm{E}(\mathbf{y}|\mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}.$$

Lee et al. (2017) called the $\mathbf{v}$-scale the weak-canonical scale and Lee and Nelder (1996) proposed the use of Henderson's joint likelihood $\mathcal{J}(\boldsymbol{\theta}; \mathbf{v})$ as the h-

likelihood for general non-normal models. However, its joint maximization cannot give the MLEs for the variance components, which leads to the use of integrated likelihood. Thus, the key to avoid computational difficulty due to integration is to define a new proper h-likelihood whose joint maximization gives the MLEs for the whole parameters including variance components. We define the h-likelihood for LMMs, which contain the multiple categorical features $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_K)$ with corresponding random effects $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_K)$. Since

$$\log f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}) + \log f_{\boldsymbol{\theta}}(\mathbf{v}) = \log f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v}) = \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y}) + \log f_{\boldsymbol{\theta}}(\mathbf{y}),$$

let us define the h-likelihood based on the canonical scale of random effects $\mathbf{v}^c$,

$$h = h(\boldsymbol{\theta}, \mathbf{v}^c) = \ell(\boldsymbol{\theta}) + \log f_{\boldsymbol{\theta}}(\mathbf{v}^c|\mathbf{y})$$

where $\ell(\boldsymbol{\theta}) = \log f_{\boldsymbol{\theta}}(\mathbf{y})$ is the integrated likelihood. Given θ, let $\widetilde{\mathbf{v}}^c$ be mode of h. A sufficient condition for $h(\boldsymbol{\theta}, \mathbf{v}^c)$ to be the h-likelihood is that $f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}^c|\mathbf{y})$ is free of $\boldsymbol{\theta}$. In Appendix 5.9.3, we show that

$$\mathbf{v}^c = \left(\frac{1}{\sigma_e^2}\mathbf{Z}^T\mathbf{Z} + \mathbf{D}^{-1}\right)^{\frac{1}{2}} \mathbf{v}$$

is the canonical scale and the resulting predictive likelihood at $\mathbf{v}^c$,

$$\log f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}^c|\mathbf{y}) = \log f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}|\mathbf{y}) + \log \left|\frac{d\mathbf{v}}{d\mathbf{v}^c}\right| = -\frac{1}{2}\log|2\pi\mathbf{I}_Q|$$

is free of $\boldsymbol{\theta}$. This leads to

$$h(\boldsymbol{\theta}, \widetilde{\mathbf{v}}^c) \propto \ell(\boldsymbol{\theta}),$$

so that the joint maximization of $h(\boldsymbol{\theta}, \mathbf{v})$ gives the MLEs for the whole fixed parameters. Let $h(\boldsymbol{\theta}, \mathbf{v})$ be a reparameterization of $h(\boldsymbol{\theta}, \mathbf{v}^c)$, then the h-likelihood can be expressed as

$$\begin{aligned} h = h(\boldsymbol{\theta}, \mathbf{v}) &= \log f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}) + \log f_{\boldsymbol{\theta}}(\mathbf{v}) + \log \left| \frac{d\mathbf{v}}{d\mathbf{v}^c} \right| \\ &= \mathcal{J}(\boldsymbol{\theta}, \mathbf{v}) - \frac{1}{2} \log \left| \frac{1}{\sigma_e^2} \mathbf{Z}^T \mathbf{Z} + \mathbf{D}^{-1} \right|. \end{aligned}$$

Thus, the h-likelihood $h(\boldsymbol{\theta}, \mathbf{v})$ is not proportional to the Henderson's joint likelihood $\mathcal{J}(\boldsymbol{\theta}, \mathbf{v})$ in (5.1), since $\log |d\mathbf{v}/d\mathbf{v}^c|$ depends upon the variance components. So the h-likelihood is different from the Henderson's joint likelihood. Given $\boldsymbol{\theta}$, the h-likelihood and joint likelihood of $\mathbf{v}$ are proportional. Thus, joint maximization of the h-likelihood provides BLUPs for random effects. With the model (5.3), the h-likelihood is

$$h = - \frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}||^2}{2\sigma_e^2} - \frac{N}{2} \log \sigma_e^2 - \frac{1}{2} \mathbf{v}^T \mathbf{D}^{-1} \mathbf{v} - \frac{1}{2} \log \left| \frac{1}{\sigma_e^2} \mathbf{Z}^T \mathbf{Z} \mathbf{D} + \mathbf{I}_Q \right|. \tag{5.4}$$

In Markov random field models or smoothing splines, the precision matrix of the random effects $\mathbf{P}_k = \mathbf{D}_k^{-1}$ are explicitly expressed and in independent random effect models $\mathbf{P}_k = \lambda_k^{-1} \mathbf{I}_{q_k}$. Let $\mathbf{P} = \text{block-diag}(\mathbf{P}_1, ..., \mathbf{P}_K)$. Then the canonical scale $\mathbf{v}^c$ becomes

$$\mathbf{v}^c = \left(\frac{1}{\sigma_e^2} \mathbf{Z}^T \mathbf{Z} + \mathbf{P} \right)^{\frac{1}{2}} \mathbf{v}$$

and the h-likelihood becomes

$$h = -\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}||^2}{2\sigma_e^2} - \frac{N}{2}\log\sigma_e^2 - \frac{1}{2}\mathbf{v}^T\mathbf{P}\mathbf{v} + \frac{1}{2}\log|\mathbf{P}| - \frac{1}{2}\log\left|\frac{1}{\sigma_e^2}\mathbf{Z}^T\mathbf{Z} + \mathbf{P}\right|,$$

which does not requires the computation of $\mathbf{D}^{-1}$.

It is worth emphasizing that the h-likelihood approach does not require the inverse of $N \times N$ matrix but only $Q \times Q$ matrix where $Q = \sum_{k=1}^{K} q_k$. It is often true that $Q \ll N$. When $\sum_{k=2}^{K} q_K \ll q_1 < N$ and $\mathbf{D}_1 = \lambda_1 \mathbf{I}_{q_1}$, it is not necessary to compute the inverse and the determinant of the whole $Q \times Q$ matrix but $(Q - q_1) \times (Q - q_1)$ matrix. In Appendix 5.9.3, we derive the first and the second derivatives of the h-likelihood, which can be obtained without computing the inverse of full $Q \times Q$ matrix directly.

The h-likelihood has advantage over the Henderson's joint likelihood, equivalent to the h-likelihood of Lee and Nelder (1996), in that it is computationally efficient and gives MLEs for all parameters. Given variance components, the joint likelihood and the h-likelihood provides common estimators. Thus, difference is ML estimation of variance components. In Appendix 5.9.3, we show that the restricted likelihood (5.2) is the adjusted profile h-likelihood,

$$\ell_R(\boldsymbol{\psi}) = h_R(\boldsymbol{\psi}) = h(\boldsymbol{\psi}; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{v}}^c) - \frac{1}{2}\log\left|\frac{1}{\sigma_e^2}\mathbf{X}^T\mathbf{X} - \frac{1}{\sigma_e^4}\mathbf{X}^T\mathbf{Z}\mathbf{A}^{-1}\mathbf{Z}^T\mathbf{X}\right| \qquad (5.5)$$

where $\mathbf{A} = \frac{1}{\sigma_e^2}\mathbf{Z}^T\mathbf{Z} + \mathbf{D}^{-1}$. Since the additional log determinant term involves an inverse of $Q \times Q$ matrix, the REML procedure is computationally harder than the ML procedure.

5.4 Learning algorithm with the h-likelihood

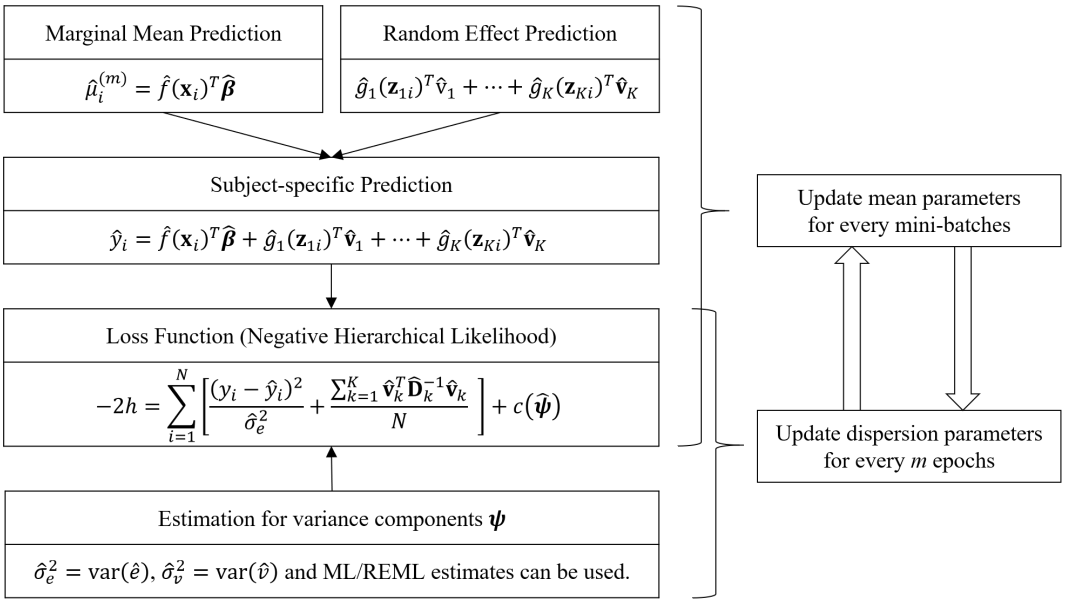


Figure 5.1: A sketch of the proposed model fitting algorithm via h-likelihood.

Following Simchoni and Rosset (2023), we first extend the LMM (5.3) to the LMMNN with random effects for the multiple categorical features,

$$\mathbf{y} = f(\mathbf{X})\boldsymbol{\beta} + g_1(\mathbf{Z}_1)\mathbf{v}_1 + \cdots + g_K(\mathbf{Z}_K)\mathbf{v}_K + \mathbf{e} \tag{5.6}$$

where $f : \mathbb{R}^{p^*} \to \mathbb{R}^p$ and $g_k : \mathbb{R}^{q_k^*} \to \mathbb{R}^{q_k}$ are non-linear functions to be estimated by the neural networks, $\mathbf{X}$ and $\mathbf{Z}_k$ are $n \times p^*$ and $n \times q_k^*$ model matrix, respectively. LMMNN allows complex covariance structures of clustered data due to categorical variables, temporal-spatial structures, and combinations of these. Here, $f(\mathbf{X})$ denotes the last hidden layer including the bias node and $\boldsymbol{\beta}$ is the weight vector from the last hidden layer to the output layer.

The extension of the h-likelihood (5.4) to the proposed model (5.6) is straightforward. By replacing $\mathbf{X}$ and $\mathbf{Z}_k$ to $f(\mathbf{X})$ and $g_k(\mathbf{Z}_k)$ for $k = 1, ..., K$, respectively, the canonical scale $\mathbf{v}^c = (\mathbf{v}_1^c, ..., \mathbf{v}_K^c)$ is given by

$$\mathbf{v}^c = \left(\frac{1}{\sigma_e^2} g(\mathbf{Z})^T g(\mathbf{Z}) + \mathbf{D}^{-1} \right)^{\frac{1}{2}} \mathbf{v}$$

where $g(\mathbf{Z}) = (g_1(\mathbf{Z}_1), ..., g_K(\mathbf{Z}_K))$. Then, the objective function for training the network is defined by the negative h-likelihood,

$$\text{Loss} = -2h = \frac{1}{\sigma_e^2} \sum_{i=1}^{N} \left[y_i - f(\mathbf{x}_i)^T \boldsymbol{\beta} - g(\mathbf{z}_i)^T \mathbf{v} \right]^2 + \sum_{k=1}^{K} \mathbf{v}_k^T \mathbf{D}_k^{-1} \mathbf{v}_k + c(\boldsymbol{\psi}), \tag{5.7}$$

where $c(\boldsymbol{\psi}) = \log \left| \sigma_e^{-2} g(\mathbf{Z})^T g(\mathbf{Z}) \mathbf{D} + \mathbf{I}_Q \right| + N \log \sigma_e^2$ is a function of $\boldsymbol{\psi}$ and $g(\mathbf{Z})$ only. Each component of the negative h-likelihood has straight-forward interpretation:

- $\frac{1}{\sigma_e^2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}||^2$ represents the conditional log-density $-2 \log f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v})$, which can be decomposed for online learning.
- $\mathbf{v}^T \mathbf{D}^{-1} \mathbf{v}$ represents the log-density $-2 \log f_{\boldsymbol{\theta}}(\mathbf{v})$, which can be viewed as a kernel regularizer for the weights of categorical features.
- The remaining term $c(\boldsymbol{\psi})$ is a function of dispersion parameters, which does not affect learning of mean parameters, i.e., all the weights in neural network and random effects.

Therefore, the h-likelihood loss for LMM can be understood as the sum of the squared loss, kernel regularizer for random effects, and an additional function

for yielding MLEs of dispersion parameters.

Let $\widehat{y}_i = \mathrm{E}(y_i|\mathbf{v}) = f(\mathbf{x}_i)^T\boldsymbol{\beta} + g(\mathbf{z}_i)^T\mathbf{v}$, then the loss function becomes

$$\text{Loss} = \sum_{i=1}^{N}\left[\frac{(y_i-\widehat{y}_i)^2}{\sigma_e^2} + \frac{\sum_{k=1}^{K}\mathbf{v}_k^T\mathbf{D}_k^{-1}\mathbf{v}_k}{N}\right] + c(\boldsymbol{\psi}).$$

and its gradient with respect to the mean parameters in $f, g, \boldsymbol{\beta}, \mathbf{v}$ is given by

$$\begin{aligned}\nabla \text{ Loss} &= \nabla\sum_{i=1}^{N}\left[\frac{(y_i-\widehat{y}_i)^2}{\sigma_e^2} + \frac{\sum_{k=1}^{K}\mathbf{v}_k^T\mathbf{D}_k^{-1}\mathbf{v}_k}{N}\right] \\ &\propto \sum_{i=1}^{N}\left[\nabla(y_i-\widehat{y}_i)^2 + \frac{\sigma_e^2}{N}\sum_{k=1}^{K}\nabla\mathbf{v}_k^T\mathbf{D}_k^{-1}\mathbf{v}_k\right],\end{aligned}$$

which does not involve the log-determinant of $Q \times Q$ matrices in $c(\boldsymbol{\psi})$. Note further that the gradient with respect to the random effects is $\nabla_{\mathbf{v}_k}\mathbf{v}_k^T\mathbf{D}_k^{-1}\mathbf{v}_k = 2\mathbf{v}_k/\lambda_k$ when $\mathbf{v}_k$ is independent random effect. Even if every pair of $\mathbf{v}_k$ is correlated, it only involves the inverse of $q_k \times q_k$ matrix. Thus, for given variance components $\boldsymbol{\psi}$, optimization of the negative h-likelihood loss (5.7) with respect to the mean parameters can naturally decompose for online learning frameworks. Furthermore, it can be interpreted as the optimization of the sum of squared error $\sum_i(y_i-\widehat{y}_i)^2$ with the penalty function $\sum_k \sigma_e^2\mathbf{v}_k^T\mathbf{D}_k^{-1}\mathbf{v}_k$. In LMMs, MLEs for mean parameters are robust against estimation of dispersion parameters, whereas MLEs for dispersion parameters are sensitive to estimation of mean parameters. Thus, we update the variance components every m epoch, not every mini-batch.

An advantage of the h-likelihood is that it avoids heavy computation in the integrated likelihood. Figure 5.1 shows our two-step algorithm with the

negative h-likelihood loss. The proposed algorithm allows online learning of mean parameters including random effects while saving the computational cost required for estimation of dispersion parameters.

- **M-step:** Update the mean parameters $(f, g, \boldsymbol{\beta}, \mathbf{v})$ in the neural network for every mini-batch.
- **V-step:** Update the variance components in $\boldsymbol{\psi}$ using the whole training data for every m epoch.

Figure 5.2 shows the MSE vs. time curves of the h-likelihood approach and the improved integrated likelihood approach with the Woodbury formula and the matrix determinant lemma. This assess the relative efficiency of the two methods in terms of computational complexity and accuracy (MSE). The MSE of the h-likelihood approach (blue) decreased more rapidly than that of the integrated likelihood approach (red). These results provide evidence that the proposed h-likelihood approach is computationally more efficient than the integrated likelihood approach, even when the latter is improved by using the Woodbury formula and the matrix determinant lemma.

In early stage of learning, the method-of-moments estimators (MMEs) could be used for training the variance components, because MLEs are often sensitive to the bias in the mean parameters and MMEs take less computational cost. It is worth noting that the MMEs require the random effect predictors $\widehat{\mathbf{v}}$, which are not provided by the integrated likelihood while training the network. When the number of dispersion parameters is small, second order optimization algorithms can be used for the covariance kernel, such as the RBF kernel. Newton-Raphson method is implemented for estimation of

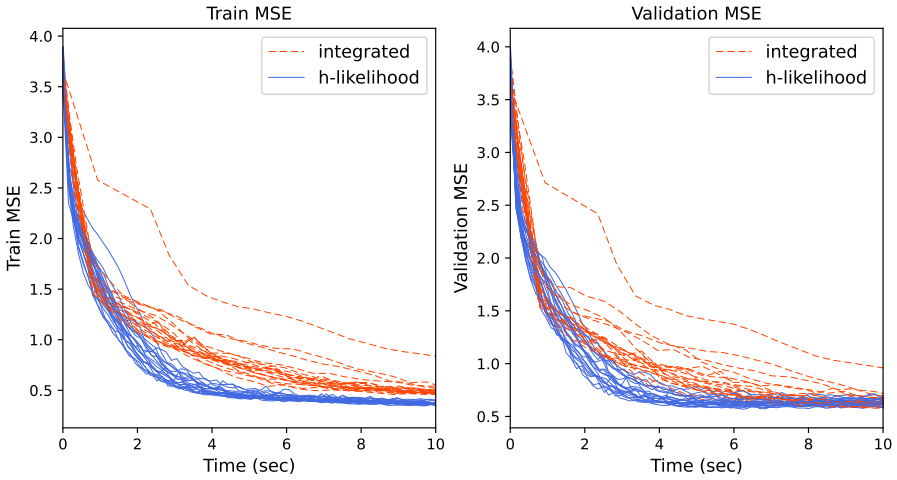


Figure 5.2: MSE curves of the integrated likelihood approach and the proposed h-likelihood approach from 20 repetitions. $N = 10,000$ data are generated from the normal distribution with a nonlinear function $f(\mathbf{x}) = (x_1+x_2)\cos(x_1+x_2) + 2x_1x_2$, $q = 100$ dimensional Gaussian random effects $\mathbf{v}_1$ and $\mathbf{v}_2$ from $N(0, I_{100})$, and $\sigma_e^2 = 1$.

dispersion parameters.

5.4.1 REML procedure

The restricted h-likelihood of the proposed model (5.6) can be obtained by replacing $\mathbf{X}$ and $\mathbf{Z}$ in (5.5) with $\widehat{f}(\mathbf{X})$ and $\widehat{g}(\mathbf{Z})$. For given $\widehat{f}$ and $\widehat{g}$, the restricted h-likelihood is given by

$$h_R(\boldsymbol{\psi}) = h(\boldsymbol{\psi}; \widehat{f}, \widehat{g}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{v}}) - \frac{1}{2}\log\left|\frac{\widehat{f}(\mathbf{X})^T\widehat{f}(\mathbf{X})}{\sigma_e^2} - \frac{\widehat{f}(\mathbf{X})^T\widehat{g}(\mathbf{Z})\mathbf{A}^{-1}\widehat{g}(\mathbf{Z})^T\widehat{f}(\mathbf{X})}{\sigma_e^4}\right|,$$

where $\mathbf{A} = \frac{1}{\sigma_e^2}\widehat{g}(\mathbf{Z})^T\widehat{g}(\mathbf{Z}) + \mathbf{D}^{-1}$, which allows REML procedure for LMMNN models.

5.4.2 Adjustments for random effects

In LMMs, constraints are imposed on the random effects $\mathrm{E}(\mathbf{v}) = \mathbf{0}$. Without the constraints, the proposed model (5.6) has additional parameter $\mu_k = \mathrm{E}(\mathbf{v}_k)$ and the transformation

$$\begin{aligned}\beta_0^* &= \beta + \epsilon_k \\ \mathbf{v}_k^* &= \mathbf{v}_k - \epsilon_k \sim N\left(\mu_k^* = \mu_k - \epsilon_k, \mathbf{D}_k\right)\end{aligned}$$

gives the same h-likelihood, so that the parameters may not be identifiable. Thus, when the DNN models contains the random effects, the bias in local minima can cause poor predictions. Simchoni and Rosset (2023) considered two cases of $g_k(\cdot)$, the identity function $g_k(\mathbf{Z}_k) = \mathbf{Z}_k$ and $g_k(\mathbf{Z}_k) = \mathbf{Z}_k\mathbf{W}_k$ where $\mathbf{W}_k$ is $q_k^* \times q_k$ matrix with $q_k \leq q_k^*$. When $g_k(\cdot)$ is identity function, we propose the following adjustment for local minima by putting constraints on the random effect predictors,

$$\widehat{\mathbf{v}}_k^* = \widehat{\mathbf{v}}_k - \frac{\widehat{\mathbf{v}}_k^T\widehat{\mathbf{D}}^{-1}\mathbf{1}_{q_k}}{\mathbf{1}_{q_k}^T\widehat{\mathbf{D}}_k^{-1}\mathbf{1}_{q_k}} \quad \text{and} \quad \widehat{\beta}_0^* = \widehat{\beta}_0 + \frac{\widehat{\mathbf{v}}_k^T\widehat{\mathbf{D}}^{-1}\mathbf{1}_{q_k}}{\mathbf{1}_{q_k}^T\widehat{\mathbf{D}}_k^{-1}\mathbf{1}_{q_k}}, \tag{5.8}$$

where $\mathbf{1}_{q_k} = (1, ..., 1)^T$. When the random effect $\mathbf{v}_k$ is independent, i.e., $\mathbf{D}_k = \lambda_k\mathbf{I}_{q_k}$, the adjustment becomes

$$\widehat{\mathbf{v}}_k^* = \widehat{\mathbf{v}}_k - \frac{1}{q_k}\sum_{j=1}^{q_k}\widehat{v}_{kj} \quad \text{and} \quad \widehat{\beta}_0 = \widehat{\beta} + \frac{1}{q_k}\sum_{j=1}^{q_k}\widehat{v}_{kj}.$$

Algorithm 1 Two-step Algorithm for H-likelihood

Input: $\mathbf{x}_i$, $\mathbf{z}_i$
Initialize all the fixed and random parameters.
repeat
  < **M-step** >
  for epoch = 1 **to** m **do**
    Update the mean parameters in f, g, $\boldsymbol{\beta}$ and $\mathbf{v}$ for every mini-batch.
  end for
  < **V-step** >
  Update dispersion parameters in $\boldsymbol{\psi}$ by using the whole training data (full batch).
until the loss function is not improved for pre-determined number of times
Adjust the random effect predictors $\widehat{\mathbf{v}}$ as in (5.8).

Following theorem shows that the proposed adjustment (5.8) can always reduce the proposed loss function.

Theorem 5.1. *In the LMMNN* (5.6)*, suppose that* $(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*)$ *is the replacement of* $\widehat{\beta}_0$ *and* $\widehat{\mathbf{v}}_k$ *in* $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})$ *with the adjusted values* $\widehat{\boldsymbol{\theta}}^*$ *and* $\widehat{\mathbf{v}}_k^*$ *in* (5.8)*, then*

$$h(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*) \geq h(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})$$

and the equality holds if and only if $\widehat{\mathbf{v}}_k^T \widehat{\mathbf{D}}^{-1} \mathbf{1}_{q_k} = 0$*, i.e.,* $\widehat{\mathbf{v}}_k^* = \widehat{\mathbf{v}}_k$*.*

5.5 Comparison with existing methods

To show the performance of their integrated likelihood approaches, namely the LMMNN with and without assuming spatial correlation (LMMNN-R and LMMNN-E), Simchoni and Rosset (2023) reported the results from various existing methods, one-hot encoding (OHE), entity embedding (EMB; Guo and

Berkhahn (2016)), convolutional neural network (CNN; LeCun et al. (1998)) and stochastic variational deep kernel learning (SV-DKL; Wilson et al. (2016b)). To study the performance of the proposed model, we first review the existing methods for comparison.

- OHE is a basic approach to handle the categorical features, but it becomes challenging when the number of categories is large.
- EMB is known to improve OHE by mapping the high-cardinality categorical features into the low-dimensional Euclidean spaces.
- CNN is the most widely used method to analyze visual images. For spatial data, CNN can be applied by handling the locations as images.
- SV-DKL is a stochastic variational procedure which generalize the deep kernel learning (Wilson et al., 2016a). It is considered as a SOTA method for handling spatial data. Deep kernel learning combines the non-parametric flexibility of kernel methods with the inductive biases of deep learning architectures. Wilson et al. (2016b) showed that SV-DKL can take advantages over alternative scalable Gaussian process models and stand-alone DNNs.
- LMMNN-E transforms the locations into a 1000 dimensional vector which is treated as a single independent random effects.
- LMMNN-R uses the RBF covariance kernel for the spatial random effects. It has the similar model formulation with our proposed HL methods but different loss function and learning algorithm using the block-diagonal approximation.

The h-likelihood approach gives exact MLEs, whereas SOTA methods such as SV-DKL, LMMNN-E and LMMNN-R provide only approximate MLEs.

5.6 Numerical studies

We present numerical studies using spatial data to demonstrate the performance of the proposed method. Following Simchoni and Rosset (2023), we generate the data as follows. For $i = 1, ..., N$, input variable $\mathbf{x}_i = (x_{i1}, ..., x_{i10})^T$ are sampled from $U(-1, 1)$ distribution and

$$y_i = x_{i+} \cdot \cos x_{i+} + 2x_{i1}x_{i2} + \mathbf{z}_i^T \mathbf{v} + \epsilon_i$$

where $x_{i+} = x_{i1} + \cdots + x_{i10}$, the noise ϵ_i is sampled from $N(0, \sigma_e^2)$, and a vector of random effects $\mathbf{v}$ is sampled from the multivariate normal distribution with zero mean and covariance represented by RBF kernel, for $i, j \in \{1, ..., q\}$,

$$\mathrm{Cov}(v_i, v_j) = \sigma_v^2 \exp\left\{-\frac{(s_i - s_j)^2}{2l^2}\right\},$$

where s_i and s_j are 2D locations sampled from $U(-10, 10) \times U(-10, 10)$ grid.

We generate $N = 100,000$ data points with $q = 1,000$ random effects. We randomly separate the data into training set (60%), validation set (20%) and test set (20%). All experiments are repeated 100 times. To fit the proposed method, Adam optimizer is used for the mean parameters, and Newton-Raphson methods is used for the variance components. Since the MLEs for variance components could be sensitive to the bias in the mean parameters, method-of-moments estimators in Appendix 5.9.2 are used in early stages.

Standard multi-layer perceptrons (MLPs) with 4 hidden layers of 100-50-25-12 neurons and 25% dropout were applied for all the experiments. Sigmoid activation function is used for the last hidden layer to obtain the REMLEs, and ReLU activation function is used for the others. Early stopping criteria with validation loss is employed to prevent overfitting. The proposed method is implemented using Python based on Keras (Chollet et al., 2015) and Tensorflow (Abadi et al., 2015), and all the experiments are made on Nvidia RTX 2080Ti GPU.

We report the mean and standard error of mean squared prediction errors (MSPEs) of test data,

$$\text{MSPE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (y_i - \widehat{y}_i)^2.$$

Since the prediction is insensitive to the estimation of dispersion parameters, the MLE and REMLE have the same prediction MSPE. Difference between LMMNN methods and HL method is dispersion parameter estimation of the HL method. Thus, the exact ML estimation of dispersion parameters enhance the predictability. Table 5.1 shows that the proposed method is better than all the existing methods. Table 5.2 shows the estimation of variance components. For the length-scale parameter l^2 in RBF kernel, block-diagonal approximation of LMMNN method produces severely biased estimates, whereas the proposed method estimates accurately. For both σ_e^2 and σ_v^2, the proposed exact MLEs are slightly better than the approximate MLEs using block-diagonal approximation. Compared to MLE, REMLE is slightly less biased, but the difference is small despite the additional computing cost. In LMMs of finite samples,

Table 5.1: Average of test MSPEs. Results of existing models are cited from Simchoni and Rosset (2023).

l^2	OHE	EMB	CNN	SV-DKL	LMMNN-E	LMMNN-R	**HL**
0.1	1.35	1.34	1.28	1.26	1.26	1.29	**1.11**
1.0	1.33	1.34	1.27	1.12	1.18	1.13	**1.03**
10.0	1.34	1.30	1.22	1.09	1.10	1.10	**1.10**

Table 5.2: Estimated variance components on average when $\sigma_e^2 = \sigma_v^2 = 1$. Results of LMMNN-R are cited from Simchoni and Rosset (2023).

True	LMMNN-R			HL (MLE)			HL (REMLE)		
l^2	$\widehat{\sigma}_e^2$	$\widehat{\sigma}_v^2$	$\widehat{l^2}$	$\widehat{\sigma}_e^2$	$\widehat{\sigma}_v^2$	$\widehat{l^2}$	$\widehat{\sigma}_e^2$	$\widehat{\sigma}_v^2$	$\widehat{l^2}$
0.1	1.12	0.99	0.48	0.934	0.983	0.097	0.934	0.983	0.097
1.0	1.12	1.10	1.49	1.001	1.059	1.009	1.001	1.059	1.008
10.0	1.11	0.74	4.93	0.962	0.712	8.929	0.962	0.713	8.934

REMLEs often reduces the bias of the MLEs, but in LMMNN with large N, the improvement seems negligible.

To demonstrate the usefulness of the adjustment (5.8) of random effects predictor, we report the root mean squared errors (RMSEs) of random effects predictors. Without adjustment, mean and standard error of RMSEs are 0.14 (0.06). With adjustment, mean and standard error of RMSEs are 0.13 (0.05). We have focused that the adjustment improves not only the random effect prediction but also estimation of MLEs for variance components, which gives the good prediction performance of the HL procedure.

In summary, the proposed HL method outperforms the existing methods

including the SOTA methods of variational approach and integrated likelihood approach for the spatial data, including SV-DKL (Wilson et al., 2016b), LMMNN-E and LMMNN-R (Simchoni and Rosset, 2023).

5.7 Real data analysis

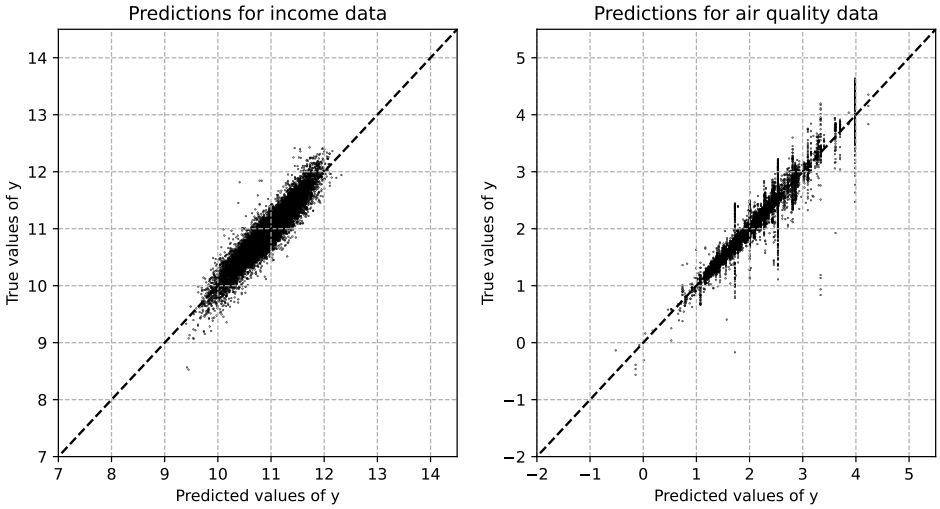


Figure 5.3: The HL predictors from income data (left) and air quality data (right).

Simchoni and Rosset (2023) analyzed several data sets. They used the 5-fold cross validation (CV) procedures where 80% of the data is used to predict and the remaining 20% is test data. Standard MLPs with two hidden layers of 10-3 neurons and ReLU activation function were used for all the data sets. RBF kernel was used for spatial correlation. Instead of OHE, analysis ignoring correlation structure (Ignore) was shown, since OHE perform similarly to EMB in simulation studies.

Income data

Income data (MuonNeutrino, 2019) have mean yearly income in dollars for $71,371$ US census tracts from $3,108$ counties. The response variable is log-income and in addition to the location features (longitude and latitude), the data contain $p = 30$ input variables. Here, $N = 71,371$, $K = 1$, and $q = 3,108$.

Air quality data

Centers for Disease Control and Prevention reported air quality data (CDC, 2020) of PM2.5 particles level in $71,347$ US census tracts. Simchoni and Rosset (2023) analyzed the air quality data by using additional features from the income data. The response variable is PM2.5 particles level with $p = 32$ input variables. Here, $N = 71,347$, $K = 1$, and $q = 3,107$.

Cars data

Cars data (Reese, 2020) have the price of $N = 97,729$ used cars. The response variable is log-price of the cars. It contains $q_1 = 15,226$ models, $q_2 = 12,235$ locations to give $Q = q_1 + q_2 = 27,461$, and $p = 73$ input variables. Since $\mathbf{D}_1 = \lambda_1 \mathbf{I}_{q_1}$, we only need to compute the $q_2 \times q_2$ inverse matrix, instead of either $N \times N$ or $Q \times Q$ inverse matrices.

Prediction results

Table 3 shows the mean of the MSPEs for test data from 5-fold CV procedure. In air quality data, the proposed method has the smallest MSPEs. In income data, it has comparable MSPEs to the smallest MSPE of LMMNN-E without

Table 5.3: Average of test MSPEs from the 5-fold cross validation. Results of existing models are cited from Simchoni and Rosset (2023).

Data	Ignore	EMB	CNN	SV-DKL	LMMNN-E	LMMNN-R	**HL**
Income	.034	.032	.032	.030	.027	.028	.028
Air	.285	.260	.163	.044	.088	.035	.023
Cars	.152	.092	.137	.149	.136	.084	.084

spatial random effects. In cars data, the proposed method and LMMNN-R outperform the other methods. Figure 5.3 shows the predicted values of output variables against the true values for the income data and air quality data. When Simchoni and Rosset (2023)'s block-diagonal approximation works well (income data and cars data), the proposed method and LMMNN-R behave similarly, whereas the approximation does not work well (air quality data), the proposed method outperforms LMMNN-R. However, not only the correlation matrix, but also the data, parameters, and the batch size can affect the accuracy of the approximation. Thus, it is hard to know whether the approximation will work well or not.

5.8 Concluding remarks

In LMMs, the conventional integrated likelihood has been successfully implemented to obtain the MLEs. However, with the surge of DNN models, the integrated likelihood encounters a computational difficulty due to the large size of data. Variational methods and approximate integrated likelihood approach have been proposed to obtain approximate MLEs. However, they could

have non-negligible biases, so the algorithm to obtain the exact MLEs is of interest. Lee and Nelder (1996) proposed the h-likelihood to avoid numerically difficult integration. However, it does not give the exact MLEs for variance components. In this chapter, we introduce a new h-likelihood for LMMs, which gives the MLEs for whole parameters and BLUPs for random effects.

For LMMNN models, the two-step algorithm enables online learning by minimizing the negative h-likelihood loss function. Its joint optimization produces exact MLEs for mean and dispersion parameters and BLUP for the random effects. The algorithm also avoids a difficulty to implement the REMLE procedure for variance components. In LMMNN models, we found that an adjustment for random effect predictors is useful for enhancing the performance of variance component estimation. In this chapter, we only considered simple MLP for the neural network $f(\mathbf{x})$, but more complex architectures can be easily implemented.

Via simulations and real data analyses, we show that predictive performance of HL method outperforms the existing methods, OHE, EMBED, CNN, and SOTA methods, SV-DKL, LMMMNN-E and LMMNN-R.

In the future we hope to make the proposed method more computationally efficient, applicable to non-normal hierarchical models such as hierarchical generalized linear models (Lee and Nelder, 1996) with neural networks.

5.9 Appendix

5.9.1 The computation of h-likelihood when q_1 is large

Suppose that the model contains a large q_1 dimensional independent random effects with $\mathbf{D}_1 = \lambda_1 \mathbf{I}_{q_1}$ Since $\mathbf{Z}_1^T \mathbf{Z}_1$ is diagonal, the determinant $\left| \frac{1}{\sigma_e^2} \mathbf{Z}^T \mathbf{Z} \mathbf{D} + \mathbf{I}_Q \right|$ in the h-likelihood (5.4) can be expressed as

$$\begin{vmatrix} \frac{1}{\sigma_e^2}\mathbf{Z}_1^T\mathbf{Z}_1\mathbf{D}_1 + \mathbf{I}_{q_1} & \frac{1}{\sigma_e^2}\mathbf{Z}_1^T\mathbf{Z}_2\mathbf{D}_2 & \dots & \frac{1}{\sigma_e^2}\mathbf{Z}_1^T\mathbf{Z}_K\mathbf{D}_K \\ \frac{1}{\sigma_e^2}\mathbf{Z}_2^T\mathbf{Z}_1\mathbf{D}_1 & \frac{1}{\sigma_e^2}\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{D}_2 + \mathbf{I}_{q_2} & \dots & \frac{1}{\sigma_e^2}\mathbf{Z}_2^T\mathbf{Z}_K\mathbf{D}_K \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sigma_e^2}\mathbf{Z}_K^T\mathbf{Z}_1\mathbf{D}_1 & \frac{1}{\sigma_e^2}\mathbf{Z}_K^T\mathbf{Z}_2\mathbf{D}_2 & \dots & \frac{1}{\sigma_e^2}\mathbf{Z}_K^T\mathbf{Z}_K\mathbf{D}_K + \mathbf{I}_{q_K} \end{vmatrix},$$

which leads to

$$\left| \frac{1}{\sigma_e^2} \mathbf{Z}^T \mathbf{Z} \mathbf{D} + \mathbf{I}_Q \right| = |\mathbf{B}_{11}| \cdot \left| \mathbf{I}_{Q-q_1} + \frac{\mathbf{Z}_{-1}^T \mathbf{Z}_{-1} \mathbf{D}_{-1}}{\sigma_e^2} - \frac{\mathbf{Z}_{-1}^T \mathbf{Z}_1 \mathbf{D}_1 \mathbf{B}_{11}^{-1} \mathbf{Z}_1^T \mathbf{Z}_{-1} \mathbf{D}_{-1}}{\sigma_e^4} \right|$$

where $\mathbf{Z}_{-1} = (\mathbf{Z}_2, ..., \mathbf{Z}_K)$, $\mathbf{D}_{-1} = \text{block-diag}(\mathbf{D}_2, ..., \mathbf{D}_K)$,

$$\mathbf{B}_{11} = \frac{1}{\sigma_e^2} \mathbf{Z}_1^T \mathbf{Z}_1 \mathbf{D}_1 + \mathbf{I}_{q_1} = \text{diag} \left(\frac{\lambda_1}{\sigma_e^2} n_{1j} + 1 \right)_{j=1,...,q_1},$$

and $n_{1j} = \sum_{t=1}^{N} z_{1jt}$ is the number of observations in the j-th category of the first categorical variable $\mathbf{Z}_1$. Since the first term is the determinant of diagonal matrix $\mathbf{B}_{11}$ and the second term is the determinant of the size $\sum_{k=2}^{K} q_K \ll q_1$ matrix, the h-likelihood can be easily computed without handling the inverse computation of $Q \times Q$ matrices. In Appendix 5.9.3, the first and the second derivatives of the h-likelihood with respect to the variance components are

derived and they can be obtained without computing the inverse of full $Q \times Q$ matrix.

5.9.2 Methods-of-moments estimators

In early stage of learning, including the initial values, MMEs of variance components are used because it is computationally fast and less sensitive to the bias in the mean parameters. For $j = 1, ..., q_k$, each v_{kj} has normal distribution with mean zero and variance λ_k. Thus, we can use

$$\widehat{\lambda}_k = \frac{1}{q_k - 1} \sum_{j=1}^{q_k} (v_{kj} - \bar{v}_k)^2,$$

for the variance of random effects and

$$\widehat{\sigma}_e^2 = \frac{1}{N-1} \sum_{j=1}^{q_k} \left[y_i - \widehat{f}(\mathbf{x}_i)^T \widehat{\boldsymbol{\beta}} - \sum_{k=1}^{K} \widehat{g}_k(\mathbf{z}_{ki})^T \widehat{\mathbf{v}}_k \right]^2$$

for the variance of noise.

5.9.3 Technical details

Since $\mathbf{v}|\mathbf{y}$ has the multivariate normal distribution,

$$\mathbf{v}|\mathbf{y} \sim N\left(\frac{1}{\sigma_e^2} \mathbf{A}^{-1} \mathbf{Z}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),\ \mathbf{A}^{-1} \right),$$

where $\mathbf{A} = \sigma_e^{-2} \mathbf{Z}^T \mathbf{Z} + \mathbf{D}^{-1}$, the distribution of $\mathbf{v}^c|\mathbf{y}$ is given by

$$\mathbf{v}^c|\mathbf{y} \sim N\left(\frac{1}{\sigma_e^2} \mathbf{A}^{-\frac{1}{2}} \mathbf{Z}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),\ \mathbf{I}_Q \right),$$

which leads to $\widetilde{\mathbf{v}}^c = \sigma_e^{-2}\mathbf{A}^{-\frac{1}{2}}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and the predictive likelihood

$$\log f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}^c|\mathbf{y}) = -\frac{1}{2}\log|2\pi\mathbf{I}_Q| = \text{constant}.$$

Thus, $\mathbf{v}^c = \mathbf{A}^{\frac{1}{2}}\mathbf{v}$ is the canonical scale to give the h-likelihood,

$$\begin{aligned} h(\boldsymbol{\theta}, \mathbf{v}^c) = &-\frac{1}{2\sigma_e^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{A}^{-\frac{1}{2}}\mathbf{v}^c\right)^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{A}^{-\frac{1}{2}}\mathbf{v}^c\right) \\ &-\frac{N}{2}\log(2\pi\sigma_e^2) - \frac{1}{2}\mathbf{v}^{cT}\mathbf{A}^{-\frac{1}{2}}\mathbf{D}^{-1}\mathbf{A}^{-\frac{1}{2}}\mathbf{v}^c - \frac{1}{2}\log\left|2\pi\mathbf{A}^{\frac{1}{2}}\mathbf{D}\mathbf{A}^{\frac{1}{2}}\right|. \end{aligned}$$

of which the joint maximization gives the MLEs for the whole parameters. The first derivatives of the h-likelihood with respect to $\boldsymbol{\beta}$ and $\mathbf{v}^c$ are

$$\begin{aligned} \frac{\partial h(\boldsymbol{\theta}, \mathbf{v}^c)}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma_e^2}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{A}^{-\frac{1}{2}}\mathbf{v}^c\right), \\ \frac{\partial h(\boldsymbol{\theta}, \mathbf{v}^c)}{\partial \mathbf{v}^c} &= \frac{1}{\sigma_e^2}\mathbf{A}^{-\frac{1}{2}}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{v}^c, \end{aligned}$$

and the second derivatives are

$$\frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v}^c)}{\partial \boldsymbol{\beta}^2} = -\frac{1}{\sigma_e^2}\mathbf{X}^T\mathbf{X}, \quad \frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v}^c)}{\partial \boldsymbol{\beta}\partial \mathbf{v}^c} = -\frac{1}{\sigma_e^2}\mathbf{A}^{-\frac{1}{2}}\mathbf{Z}^T\mathbf{X}, \quad \frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v}^c)}{\partial \mathbf{v}^{c2}} = -\mathbf{I}_Q,$$

which leads to

$$\left|-\frac{\partial^2 h(\boldsymbol{\theta}, \mathbf{v}^c)}{\partial(\boldsymbol{\beta}, \mathbf{v}^c)^2}\right| = \begin{vmatrix} \mathbf{I}_Q & \frac{1}{\sigma_e^2}\mathbf{A}^{-\frac{1}{2}}\mathbf{Z}^T\mathbf{X} \\ \frac{1}{\sigma_e^2}\mathbf{X}^T\mathbf{Z}\mathbf{A}^{-\frac{1}{2}} & \frac{1}{\sigma_e^2}\mathbf{X}^T\mathbf{X} \end{vmatrix} = \left|\frac{1}{\sigma_e^2}\mathbf{X}^T\mathbf{X} - \frac{1}{\sigma_e^4}\mathbf{X}^T\mathbf{Z}\mathbf{A}^{-1}\mathbf{Z}^T\mathbf{X}\right|.$$

Thus, the adjusted profile h-likelihood is given by

$$h_R(\boldsymbol{\psi}) = h(\boldsymbol{\psi}; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{v}}^c) - \frac{1}{2}\log\left|\frac{1}{\sigma_e^2}\mathbf{X}^T\mathbf{X} - \frac{1}{\sigma_e^4}\mathbf{X}^T\mathbf{Z}\mathbf{A}^{-1}\mathbf{Z}^T\mathbf{X}\right|,$$

which is the integrated likelihood,

$$\begin{aligned} h_R(\boldsymbol{\psi}) &= \log\iint \exp(h(\boldsymbol{\theta}, \mathbf{v}^c))d\mathbf{v}^c d\boldsymbol{\beta} = \log\iint f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{v}^c)d\mathbf{v}^c d\boldsymbol{\beta} \\ &= \log\int f_{\boldsymbol{\theta}}(\mathbf{y})d\boldsymbol{\beta} = \log\int \exp(\ell(\boldsymbol{\theta}))d\boldsymbol{\beta} \\ &= \ell(\boldsymbol{\psi}; \widehat{\boldsymbol{\beta}}) - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| = \ell_R(\boldsymbol{\psi}). \end{aligned}$$

Thus, the restricted likelihood is an adjusted profile h-likelihood.

Let $s_e = \log\sigma_e^2$ be the log-variance of random noise and $\boldsymbol{\lambda}_k = (\lambda_{k1}, ..., \lambda_{kj_k})$ be the vector of j_k dispersion parameters involved in $\mathbf{D}_k$ for $k = 1, ..., K$, then the objective function can be expressed as

$$\begin{aligned} \text{Loss} &= e^{-s_e}(\mathbf{y} - \widehat{\mathbf{y}})^T(\mathbf{y} - \widehat{\mathbf{y}}) + Ns_e + \sum_{k=1}^{K}\mathbf{v}_k^T\mathbf{D}_k^{-1}\mathbf{v}_k + \log|\mathbf{B}| \\ &= a_0(s_e) + \sum_{k=1}^{K} a_k(\boldsymbol{\lambda}_k) + \log|\mathbf{B}(s_e, \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_K)| \end{aligned}$$

where $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{v}}$, $\mathbf{B} = \mathbf{A}\mathbf{D} = e^{-s_e}\mathbf{Z}^T\mathbf{Z}\mathbf{D} + \mathbf{I}_Q$, $a_0(s_e) = e^{-s_e}(\mathbf{y} - \widehat{\mathbf{y}})^T(\mathbf{y} - \widehat{\mathbf{y}}) + Ns_e$ and $a_k(\boldsymbol{\lambda}_k) = \mathbf{v}_k^T\mathbf{D}_k^{-1}\mathbf{v}_k$. Here the derivatives of $\log|\mathbf{B}|$ is difficult to

evaluate. The first deirvatives of $\mathbf{B}$ are given by

$$\begin{aligned}\frac{\partial \mathbf{B}}{\partial s_e} &= -e^{-s_e}\mathbf{Z}^T\mathbf{Z}\mathbf{D} = \mathbf{I}_Q - \mathbf{B},\\ \frac{\partial \mathbf{B}}{\partial \lambda_{kj}} &= e^{-s_e}\mathbf{Z}^T\mathbf{Z}\frac{\partial \mathbf{D}}{\partial \lambda_{kj}} = e^{-s_e}\left(\mathbf{0}_{k-},\ \mathbf{Z}^T\mathbf{Z}_k\frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}},\ \mathbf{0}_{k+}\right),\end{aligned}$$

where $\mathbf{0}_{k-}$ and $\mathbf{0}_{k+}$ are zero matrices of size $Q \times (q_1 + \cdots + q_{k-1})$ and $Q \times (q_{k+1} + \cdots + q_K)$, respectively, so that

$$\mathbf{Z}^T\mathbf{Z}\frac{\partial \mathbf{D}}{\partial \lambda_{kj}} = \left(\mathbf{0}_{k-},\ \mathbf{Z}^T\mathbf{Z}_k\frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}},\ \mathbf{0}_{k+}\right) = \begin{pmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_1^T\mathbf{Z}_k\frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_K^T\mathbf{Z}_k\frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix},$$

and the non-zero second derivatives are given by

$$\begin{aligned}\frac{\partial^2 \mathbf{B}}{\partial s_e^2} &= e^{-s_e}\mathbf{Z}^T\mathbf{Z}\mathbf{D} = \mathbf{B} - \mathbf{I}_Q,\\ \frac{\partial^2 \mathbf{B}}{\partial s_e \partial \lambda_{kj}} &= -e^{-s_e}\mathbf{Z}^T\mathbf{Z}\frac{\partial \mathbf{D}}{\partial \lambda_{kj}} = -\frac{\partial \mathbf{B}}{\partial \lambda_{kj}},\\ \frac{\partial^2 \mathbf{B}}{\partial \lambda_{kj}^2} &= e^{-s_e}\mathbf{Z}^T\mathbf{Z}\frac{\partial^2 \mathbf{D}}{\partial \lambda_{kj}^2} = e^{-s_e}\left(\mathbf{0}_{k-},\ \mathbf{Z}^T\mathbf{Z}_k\frac{\partial^2 \mathbf{D}_k}{\partial \lambda_{kj}^2},\ \mathbf{0}_{k+}\right),\\ \frac{\partial^2 \mathbf{B}}{\partial \lambda_{ki} \partial \lambda_{kj}} &= e^{-s_e}\mathbf{Z}^T\mathbf{Z}\frac{\partial \mathbf{D}}{\partial \lambda_{kj}} = e^{-s_e}\left(\mathbf{0}_{k-},\ \mathbf{Z}^T\mathbf{Z}_k\frac{\partial^2 \mathbf{D}_k}{\partial \lambda_{ki}\partial \lambda_{kj}},\ \mathbf{0}_{k+}\right).\end{aligned}$$

Thus, the first and second derivatives of $\log|\mathbf{B}|$ are

$$\begin{aligned}\frac{\partial \log|\mathbf{B}|}{\partial s_e} &= tr\left[\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial s_e}\right] = tr[\mathbf{B}^{-1} - \mathbf{I}_Q] = tr[\mathbf{B}^{-1}] - Q\\ \frac{\partial \log|\mathbf{B}|}{\partial \lambda_{kj}} &= tr\left[\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial \lambda_{kj}}\right] = e^{-s_e}tr\left[[\mathbf{B}^{-1}]_k\mathbf{Z}^T\mathbf{Z}_k\frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}}\right]\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \log|\mathbf{B}|}{\partial s_e^2} &= tr\left[\mathbf{B}^{-1}\frac{\partial^2 \mathbf{B}}{\partial s_e^2}\right] - tr\left[\left(\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial s_e}\right)^2\right] = tr[\mathbf{B}^{-1} - \mathbf{B}^{-2}] \\
\frac{\partial^2 \log|\mathbf{B}|}{\partial \lambda_{kj}^2} &= tr\left[\mathbf{B}^{-1}\frac{\partial^2 \mathbf{B}}{\partial \lambda_{kj}^2}\right] - tr\left[\left(\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial \lambda_{kj}}\right)^2\right] \\
&= e^{-s_e} tr\left[[\mathbf{B}^{-1}]_k \mathbf{Z}^T\mathbf{Z}_k \frac{\partial^2 \mathbf{D}_k}{\partial \lambda_{kj}^2}\right] - e^{-s_e} tr\left[\left([\mathbf{B}^{-1}]_k \mathbf{Z}^T\mathbf{Z}_k \frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}}\right)^2\right] \\
\frac{\partial^2 \log|\mathbf{B}|}{\partial s_e \partial \lambda_{kj}} &= tr\left[\mathbf{B}^{-1}\frac{\partial^2 \mathbf{B}}{\partial s_e \partial \lambda_{kj}}\right] - tr\left[\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial s_e}\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial \lambda_{kj}}\right] \\
&= -e^{-s_e} tr\left[[\mathbf{B}^{-2}]_k \mathbf{Z}^T\mathbf{Z}_k \frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}}\right] \\
\frac{\partial^2 \log|\mathbf{B}|}{\partial \lambda_{ki} \partial \lambda_{kj}} &= tr\left[\mathbf{B}^{-1}\frac{\partial^2 \mathbf{B}}{\partial \lambda_{ki} \partial \lambda_{kj}}\right] - tr\left[\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial \lambda_{ki}}\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial \lambda_{kj}}\right] \\
&= e^{-s_e} tr\left[[\mathbf{B}^{-1}]_k \mathbf{Z}^T\mathbf{Z}_k \frac{\partial^2 \mathbf{D}_k}{\partial \lambda_{ki} \partial \lambda_{kj}}\right] \\
&\quad - e^{-s_e} tr\left[[\mathbf{B}^{-1}]_k \mathbf{Z}^T\mathbf{Z}_k \frac{\partial \mathbf{D}_k}{\partial \lambda_{ki}}[\mathbf{B}^{-1}]_k \mathbf{Z}^T\mathbf{Z}_k \frac{\partial \mathbf{D}_k}{\partial \lambda_{kj}}\right]
\end{aligned}
$$

where $[\mathbf{B}^{-1}]_k$ is the submatrix of $\mathbf{B}^{-1}$ from $(q_1+\cdots+q_{k-1}+1)$-th row to $(q_1+\cdots+q_k)$-th row. In real data analyses, one of the categorical features has sometimes extremely high cardinality $q_1 \gg \sum_{k=2}^K q_k$. In such cases, the corresponding random effect $\mathbf{v}_1$ is assumed to be independent but $\mathbf{B} = \frac{1}{\sigma_e^2}\mathbf{Z}^T\mathbf{Z}\mathbf{D} + \mathbf{I}_Q$ is not a diagonal, so the computation of the derivatives involves the inverse of extremely high dimensional matrix. However, $\mathbf{B}$ is a sparse matrix such that

$$
\mathbf{B}^{-1} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathrm{diag}\left(\frac{\lambda_1 n_{1j}}{\sigma_e^2} + 1\right) & \frac{1}{\sigma_e^2}\mathbf{Z}_1^T\mathbf{Z}\mathbf{D} \\ \frac{1}{\sigma_e^2}\mathbf{Z}^T\mathbf{Z}_1\mathbf{D}_1 & \frac{1}{\sigma_e^2}\mathbf{Z}_{-1}^T\mathbf{Z}_{-1}\mathbf{D}_{-1} + \mathbf{I}_{Q-q_1} \end{pmatrix}^{-1},
$$

so the inverse $\mathbf{B}^{-1}$ can be computed by using decomposition of the submatrices.

Note here that the second derivative of $\mathbf{D}_k$ becomes zero when $\mathbf{v}_k$ is in-

dependent. Suppose that $\mathbf{v}_1, ..., \mathbf{v}_{K-1}$ are independent random effects and the last random effect $\mathbf{v}_K$ is correlated, i.e., $\mathbf{D}_k = \lambda_k \mathbf{I}_{q_k}$ for $k = 1, ..., K-1$ and $\mathbf{D}_K = \mathbf{D}_K(\boldsymbol{\lambda}_K)$ for $\boldsymbol{\lambda}_K = (\lambda_{K1}, ..., \lambda_{KJ})$. The computation can be further reduced, because for $k = 1, ..., K-1$, the first and the second derivatives of $\mathbf{D}_k$ are the identity matrix and zero matrix, respectively.

Proof of Theorem 1

Let $\widehat{\beta}_0^* = \widehat{\beta}_0 + \delta$ and $\widehat{\mathbf{v}}_k^* = \widehat{\mathbf{v}}_k - \delta$. Note here that δ does not affect the predicted values of the output variable $\widehat{\mathbf{y}}$, because $\widehat{\beta}_0^* + \mathbf{Z}_k \widehat{\mathbf{v}}_k^* = \widehat{\beta}_0 + \mathbf{Z}_k \widehat{\mathbf{v}}_k$. The first derivative of h-likelihood with respect to δ is given by

$$\frac{\partial h(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*)}{\partial \delta} = \frac{\partial}{\partial \delta}\left(-\frac{1}{2}(\widehat{\mathbf{v}}_k - \delta)^T \widehat{\mathbf{D}}_k^{-1}(\widehat{\mathbf{v}}_k - \delta)\right) = \widehat{\mathbf{v}}_k^T \widehat{\mathbf{D}}_k^{-1} \mathbf{1}_{q_k} - \delta \cdot \mathbf{1}_{q_k}^T \widehat{\mathbf{D}}_k^{-1} \mathbf{1}_{q_k},$$

which leads to the solution $\delta = \widehat{\mathbf{v}}_k^T \widehat{\mathbf{D}}_k^{-1} \mathbf{1}_{q_k} / \mathbf{1}_{q_k}^T \widehat{\mathbf{D}}_k^{-1} \mathbf{1}_{q_k}$, where $\mathbf{1}_{q_k} = (1, ..., 1)^T$. The second derivative is given by

$$\frac{\partial^2 h(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*)}{\partial \delta^2} = -\mathbf{1}_{q_k} \widehat{\mathbf{D}}_k^{-1} \mathbf{1}_{q_k} < 0,$$

since $\widehat{\mathbf{D}}_k$ should be positive definite. Thus, for given $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{v}}$, the h-likelihood has the unique maximum at $\delta = \widehat{\mathbf{v}}_k^T \widehat{\mathbf{D}}^{-1} \mathbf{1}_{q_k} / \mathbf{1}_{q_k}^T \widehat{\mathbf{D}}^{-1} \mathbf{1}_{q_k}$. This implies that the adjustment (5.8) can always increase the h-likelihood,

$$h(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*) \geq h(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}}),$$

and the equality holds if and only if $\widehat{\mathbf{v}}_k^T \widehat{\mathbf{D}}_k^{-1} \mathbf{1}_{q_k} = 0$.

Chapter 6

DNN for clustered count data via h-likelihood

6.1 Introduction

Deep neural networks (DNNs) have been proposed to capture the nonlinear relationship between input and output variables (Goodfellow et al., 2016; LeCun et al., 2015). However, DNNs provide efficient marginal predictions for independent outputs, while in practice, they can be correlated, over-dispersed, or clustered. On the other hand, random effect models have been employed to make subject-specific predictions. Lee and Nelder (1996) proposed hierarchical generalized linear models (HGLMs), which allow random effects from an arbitrary conjugate distribution of the generalized linear model (GLM) family of the outputs.

Both DNNs and random effect models have successfully improved prediction accuracy. Recently, there has been a growing interest in combining these

two extensions. Simchoni and Rosset (2021, 2023) proposed the linear mixed model neural network for Gaussian (continuous) outputs with the normal random effects, which are conjugates of the Gaussian outputs. These Gaussian DNNs allow explicit expressions for likelihoods. Lee and Lee (2023) introduced the hierarchical likelihood (h-likelihood) approach, which provides the most efficient likelihood-based procedure. For non-Gaussian (discrete) outputs, Tran et al. (2020) proposed a Bayesian approach for DNNs with normal random effects using the variational approximation method (Bishop and Nasrabadi, 2006; Blei et al., 2017). Mandel et al. (2021) used a quasi-likelihood approach (Breslow and Clayton, 1993) for DNNs, but the quasi-likelihood method has been criticized for poor predictions. Lee and Nelder (2001) proposed the use of the Laplace approximation to have approximate maximum likelihood estimators (MLEs). Although Mandel et al. (2021) also adapted Laplace approximation for DNNs, their method ignored many terms in the second derivatives due to computational expense, which could lead to inconsistent estimations (Han and Lee, 2022). Therefore, a new approach is desired for non-Gaussian DNNs to have the exact MLEs.

Clustered count outputs are widely encountered in various fields (Henderson and Shimakura, 2003; Henry et al., 1998; Roulin and Bersier, 2007; Thall and Vail, 1990), but to the best of our knowledge, there appears to be no available source code for the Poisson DNN with random effects. In this chapter, we introduce Poisson-gamma DNN for the clustered count data. We propose the use of the h-likelihood approach, which allows simultaneous estimation of MLEs for fixed parameters and best unbiased predictors (BUPs) for random effects. In contrast to the ordinary DNN framework, we found that the local minima

can cause poor prediction when the network has subject-specific random effects. To address this issue, we propose an adjustment to the random effect prediction that prevents from violation of the constraints in random effects for identifiability. Additionally, we introduce the method-of-moments estimators for pretraining the variance components.

In Section 6.2, we present the Poisson-gamma DNN. In Section 6.3, we derive the h-likelihood for the Poisson-gamma DNN. In Section 6.4, we present the algorithm for online learning, which includes an adjustment of random effect predictors and pretraining procedure for variance components. In Section 6.5, we provide experimental studies to compare the proposed method with various existing methods. The results of the experimental studies clearly show that the proposed method improves predictions. In Section 6.6, real data analyses demonstrate that the proposed method has the best prediction accuracy in various clustered count data. In particular, introducing the subject-specific effects enhances the ability of the neural network to identify the nonlinear effects of the input variables. All the proofs are in Appendix 6.8.

6.2 Model Descriptions

6.2.1 Poisson DNN

Let $\mathcal{D} = \{(y_{ij}, \mathbf{x}_{ij}) : i = 1, ..., n,\ j = 1, ..., q_i\}$ be a dataset with output variable y_{ij} and p-dimensional vector of input variables $\mathbf{x}_{ij}$, where the subscript (i, j) denotes the jth outcome of the ith subject (cluster). For the prediction of count outputs, Poisson DNN (Rodrigo and Tsokos, 2020) gives the marginal

predictor,

$$\eta_{ij}^m = \log \mu_{ij}^m = \text{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta}) = \sum_{k=1}^{p_L} g_k(\mathbf{x}_{ij}; \mathbf{w})\beta_k + \beta_0, \tag{6.1}$$

where $\mu_{ij}^m = \text{E}(Y_{ij}|\mathbf{x}_{ij})$ is the marginal mean, $\text{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})$ is the neural network predictor, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p_L})^T$ is the vector of weights and bias between the last hidden layer and the output layer, $g_k(\mathbf{x}_{ij}; \mathbf{w})$ denotes the k-th node of the last hidden layer, and $\mathbf{w}$ is the vector of all the weights and biases before the last hidden layer. Here the inverse of the log function, $\exp(\cdot)$, becomes the activation function of the output layer. Poisson DNNs allow highly non-linear relationship between input and output variables, but only provide the marginal predictions. Thus, Poisson DNN can be viewed as an extension of Poisson GLM, where

$$\eta_{ij}^m = \mathbf{x}_{ij}^T \beta.$$

6.2.2 Poisson-gamma DNN

To allow subject-specific prediction, we propose the Poisson-gamma DNN,

$$\eta_{ij}^c = \log \mu_{ij}^c = \text{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta}) + \mathbf{z}_{ij}^T \mathbf{v}, \tag{6.2}$$

where $\mu_{ij}^c = \text{E}(Y_{ij}|\mathbf{x}_{ij}, v_i)$ is the conditional mean, $\text{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})$ is the marginal predictor of the Poisson DNN in (6.1), $\mathbf{v} = (v_1, ..., v_n)^T$ is the vector of random effects from the log-gamma distribution, and $\mathbf{z}_{ij}$ is the (i, j)-th vector of model

matrix for random effects. Here the conditional mean μ_{ij}^c can be expressed as

$$\mu_{ij}^c = \exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\} \cdot u_i,$$

where $u_i = \exp(v_i)$ is the gamma random effect. Using the random effects $\mathbf{v}$, subject-specific predictions can be made. Note here that, for any $\epsilon \in \mathbb{R}$, the model equation (6.2) can be expressed as

$$\log \mu_{ij}^c = \sum_{k=1}^{p_L} g_k(\mathbf{x}_{ij}; \mathbf{w})\beta_k + \beta_0 + v_i = \sum_{k=1}^{p_L} g_k(\mathbf{x}_{ij}; \mathbf{w})\beta_k + (\beta_0 - \epsilon) + (v_i + \epsilon),$$

or equivalently, for any $\delta = \exp(\epsilon) > 0$,

$$\mu_{ij}^c = \exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\} \cdot u_i = \left[\frac{1}{\delta}\exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\}\right] \cdot (\delta u_i),$$

which leads to an identifiability problem. Thus, it is necessary to place constraints on either the fixed effects or the random effects. Lee and Nelder (1996) proposed imposing constraints on the random effects rather than fixed effects. In this chapter, we use the constraints $\mathrm{E}(u_i) = \mathrm{E}(\exp(v_i)) = 1$ as described in Lee et al. (2017). The use of constraints $\mathrm{E}(u_i) = 1$ has an advantage that the marginal predictions can be directly obtained, because

$$\mu_{ij}^m = \mathrm{E}[\mathrm{E}(Y_{ij}|\mathbf{x}_{ij}, u_i)] = \mathrm{E}\left[\exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\} \cdot u_i\right] = \exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\}.$$

Thus, we employ $v_i = \log u_i$ to the equation (6.2), where $u_i \sim \mathrm{Gamma}(\lambda^{-1}, \lambda^{-1})$ with $\mathrm{E}(u_i) = 1$ and $\mathrm{var}(u_i) = \lambda$. By allowing two separate output nodes, the

Poisson-gamma DNN provides both marginal and subject-specific predictions,

$$\widehat{\mu}_{ij}^{m} = \exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \widehat{\mathbf{w}}, \widehat{\boldsymbol{\beta}})\right\} \quad \text{and} \quad \widehat{\mu}_{ij}^{c} = \exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \widehat{\mathbf{w}}, \widehat{\boldsymbol{\beta}}) + \mathbf{z}_{ij}^{T}\widehat{\mathbf{v}}\right\},$$

where the hats denote the predicted values. Subject-specific prediction can be achieved by multiplying the marginal mean predictor $\widehat{\mu}_{ij}^{m}$ from Poisson DNN and the subject-specific predictor of random effect $\widehat{u}_i = \exp(\widehat{v}_i)$. Note that

$$\mathrm{var}(Y|\mathbf{x}) = \mathrm{E}(\mathrm{var}(Y|\mathbf{x}, \mathbf{v})) + \mathrm{var}(\mathrm{E}(Y|\mathbf{x}, \mathbf{v})) \geq \mathrm{E}(\mathrm{var}(Y|\mathbf{x}, \mathbf{v})),$$

where $\mathrm{var}(\mathrm{E}(Y|\mathbf{x}, \mathbf{v}))$ represents between-subject variance and $\mathrm{E}(\mathrm{var}(Y|\mathbf{x}, \mathbf{v}))$ represents within-subject variance. To enhance the predictions, Poisson-gamma DNN uses the conditional predictor $\mathrm{E}(Y|\mathbf{x}, \mathbf{v})$ having only within-subject variance, whereas Poisson DNN improves the marginal predictor $\mathrm{E}(Y|\mathbf{x})$ by allowing highly nonlinear function of $\mathbf{x}$. By replacing $\mathrm{NN}(\cdot)$ with a linear model $\eta_{ij}^{m} = \mathbf{x}_{ij}^{T}\beta$, in the Poisson-gamma DNN, the model becomes the Poisson-gamma HGLM.

6.3 Construction of h-likelihood

For statistical models with random effects, it is important to define the objective function for obtaining MLEs of fixed parameters $\boldsymbol{\theta} = (\mathbf{w}, \boldsymbol{\beta}, \lambda)$. Consider an extended likelihood for different scale $(\boldsymbol{\theta}, \mathbf{u})$,

$$\ell_e(\boldsymbol{\theta}, \mathbf{u}) = \sum_{i,j} \log f_{\boldsymbol{\theta}}(y_{ij}|u_i) + \sum_{i} \log f_{\boldsymbol{\theta}}(u_i), \tag{6.3}$$

and an extended likelihood for $(\boldsymbol{\theta}, \mathbf{v})$,

$$\ell_e(\boldsymbol{\theta}, \mathbf{v}) = \sum_{i,j} \log f_{\boldsymbol{\theta}}(y_{ij}|v_i) + \sum_i \log f_{\boldsymbol{\theta}}(v_i).$$

Note here that nonlinear transformation of random effects leads to a different extended likelihood. Due to the Jacobian terms,

$$\ell_e(\boldsymbol{\theta}, \mathbf{v}) = \ell_e(\boldsymbol{\theta}, \mathbf{u}) + \sum_i \log \left| \frac{du_i}{dv_i} \right| \neq \ell_e(\boldsymbol{\theta}, \mathbf{u}).$$

The two extended likelihoods $\ell_e(\boldsymbol{\theta}, \mathbf{u})$ and $\ell_e(\boldsymbol{\theta}, \mathbf{v})$ lead to different MLEs, raising the question on how to obtain the true MLE. In Poisson-gamma HGLMs, Lee and Nelder (1996) proposed the use of $\ell_e(\boldsymbol{\theta}, \mathbf{v})$. Given the variance component λ, their approach can give MLEs for $\boldsymbol{\beta}$ and BUPs for $\mathbf{u}$ by the joint maximization of $\ell_e(\boldsymbol{\theta}, \mathbf{v})$. However, it could not yield MLE for the variance component λ. In this chapter, we derive the new h-likelihood whose joint maximization leads to MLEs of the whole fixed parameters including the variance component λ and BUPs of the random effects $\mathbf{u}$ and conditional mean $\boldsymbol{\mu}^c$.

Consider an objective function of the form

$$h(\boldsymbol{\theta}, \mathbf{v}) = \ell_e(\boldsymbol{\theta}, \mathbf{v}) + c(\boldsymbol{\theta}; \mathbf{y}), \tag{6.4}$$

where $c(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n c_i(\boldsymbol{\theta}; \mathbf{y}_i)$ is a function of $\boldsymbol{\theta}$ and $\mathbf{y}_i = (y_{i1}, ..., y_{iq_i})^T$ for each subject $i = 1, ..., n$. Then, the equation (6.4) can be expressed as

$$h(\boldsymbol{\theta}, \mathbf{v}) = \{\log f_{\boldsymbol{\theta}}(\mathbf{y}) + \log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})\} + c(\boldsymbol{\theta}; \mathbf{y}) = \ell(\boldsymbol{\theta}; \mathbf{y}) + \log f_{\boldsymbol{\theta}}(\mathbf{v}^*|\mathbf{y}),$$

where

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \log \int \exp\{\ell_e(\boldsymbol{\theta}, \mathbf{v})\}\, d\mathbf{v} \tag{6.5}$$

is the classical marginal log-likelihood for MLEs of $\boldsymbol{\theta}$, and $\mathbf{v}^* = (v_1^*, ..., v_n^*)^T$ with elements

$$v_i^* = v_i \cdot \exp\{-c_i(\boldsymbol{\theta}; \mathbf{y}_i)\}.$$

A sufficient condition for $h(\boldsymbol{\theta}, \mathbf{v})$ to yield exact MLEs of all the fixed parameters in $\boldsymbol{\theta}$ is that $f_{\boldsymbol{\theta}}(\widetilde{\mathbf{v}}^*|\mathbf{y})$ is independent of $\boldsymbol{\theta}$, where $\widetilde{\mathbf{v}}^*$ is the mode,

$$\widetilde{\mathbf{v}}^* = \arg\max_{\mathbf{v}^*} h(\boldsymbol{\theta}, \mathbf{v}^*) = \arg\max_{\mathbf{v}^*} \log f_{\boldsymbol{\theta}}(\mathbf{v}^*|\mathbf{y}).$$

Let $c_i(\boldsymbol{\theta}; \mathbf{y}_i) = (y_{i+} + \lambda^{-1}) + \log\Gamma(y_{i+} + \lambda^{-1}) - (y_{i+} + \lambda^{-1})\log(y_{i+} + \lambda^{-1})$ and y_{i+} be the sum of responses in $\mathbf{y}_i$, then $f(\widetilde{\mathbf{v}}^*|\mathbf{y})$ becomes free of $\boldsymbol{\theta}$,

$$\log f(\widetilde{\mathbf{v}}^*|\mathbf{y}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(\widetilde{v}_i^*|\mathbf{y}) = \sum_{i=1}^{n} \{\log f_{\boldsymbol{\theta}}(\widetilde{v}_i|\mathbf{y}) + c_i(\boldsymbol{\theta}; \mathbf{y}_i)\} = 0,$$

which leads to yielding MLEs because

$$h(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) = \ell(\boldsymbol{\theta}, \mathbf{y}).$$

Thus, joint maximization of the h-likelihood (6.4) provides exact MLEs for the fixed parameters $\boldsymbol{\theta}$, including the variance component λ. Furthermore, BUPs of $\mathbf{u}$ and $\boldsymbol{\mu}^c$ can be obtained from the h-likelihood,

$$\widetilde{\mathbf{u}} = \exp(\widetilde{\mathbf{v}}) = \mathrm{E}(\mathbf{u}|\mathbf{y}) \quad \text{and} \quad \widetilde{\boldsymbol{\mu}}^c = \exp(\widetilde{\mathbf{v}}) \cdot \mathrm{NN}(\mathbf{X}; \mathbf{w}, \boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{\mu}^c|\mathbf{y}).$$

It is worth emphasizing that the h-likelihood differs from the Henderson's joint likelihood (Henderson, 1973) for linear mixed models whose joint maximization cannot yield MLEs of variance components.

6.4 Learning algorithm with the h-likelihood

6.4.1 Loss function for online learning

The proposed Poisson-gamma DNN can be trained by optimizing the negative h-likelihood loss,

$$\begin{aligned}\text{Loss} = -h(\boldsymbol{\theta}, \mathbf{v}) &= -\log f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}) - \log f_{\boldsymbol{\theta}}(\mathbf{v}) - c(\boldsymbol{\theta}; \mathbf{y}) \\ &= -\sum_{i,j} \left[y_{ij} \left(\log \mu_{ij}^m + v_i \right) - e^{v_i} \mu_{ij}^m \right] \\ &\quad - \sum_{i=1}^{n} \left[\frac{v_i - e^{v_i} - \log \lambda}{\lambda} - \log \Gamma \left(\lambda^{-1} \right) + c_i(\lambda; y_{i+}) \right],\end{aligned}$$

which is a function of the two separate output nodes $\mu_{ij}^m = \text{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})$ and $v_i = \mathbf{z}_{ij}^T \mathbf{v}$. To apply online stochastic optimization methods, the proposed loss function is expressed as

$$\text{Loss} = \sum_{i,j} \left[-y_{ij} \left(\log \mu_{ij}^m + v_i \right) + e^{v_i} \mu_{ij}^m - \frac{v_i - e^{v_i}}{q_i \lambda} + a_i(\lambda; \mathbf{y}_i) \right], \tag{6.6}$$

where $a_i(\lambda; \mathbf{y}_i) = q_i^{-1} \left\{ \lambda^{-1} \log \lambda + \log \Gamma(\lambda^{-1}) - c_i(\lambda, y_{i+}) \right\}$.

6.4.2 The local minima problem

Though DNNs often encounter the local minima, Dauphin et al. (2014) claimed that local minima may not produce poor predictions in ordinary DNNs. However, we observed that the local minima can lead to poor prediction when the network reflects subject-specific effects. In Poisson-gamma DNNs, we impose the constraint $\mathrm{E}(u_i) = 1$ for identifiability, because

$$\mu_{ij}^c = \exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\} \cdot u_i = \left[\frac{1}{\delta}\exp\left\{\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\}\right] \cdot (\delta u_i)\,.$$

However, in practice, Poisson-gamma DNNs often end with local minima that violate the constraint. To prevent poor prediction due to local minima, we introduce an adjustment to the predictors of u_i,

$$\widehat{u}_i \leftarrow \frac{\widehat{u}_i}{\frac{1}{n}\sum_{i=1}^n \widehat{u}_i} \quad \text{and} \quad \widehat{\beta}_0 \leftarrow \widehat{\beta}_0 + \log\left(\frac{1}{n}\sum_{i=1}^n \widehat{u}_i\right) \tag{6.7}$$

to satisfy $\sum_{i=1}^n \widehat{u}_i/n = 1$. The following theorem shows that the proposed adjustment improves the local h-likelihood prediction.

Theorem 6.1. *In Poisson-gamma DNNs, suppose that $\widehat{\beta}_0$ and $\widehat{u}_i$ are estimates of β_0 and u_i such that*

$$\frac{1}{n}\sum_{i=1}^n \widehat{u}_i = 1 + \epsilon$$

for some $\epsilon \in \mathbb{R}$. Let $\widehat{u}_i^*$ and $\widehat{\beta}_0^*$ be the adjusted estimators in (6.7)*. Then,*

$$h(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*) \geq h(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}}),$$

and the equality holds if and only if $\epsilon = 0$, where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^*$ are vectors of the

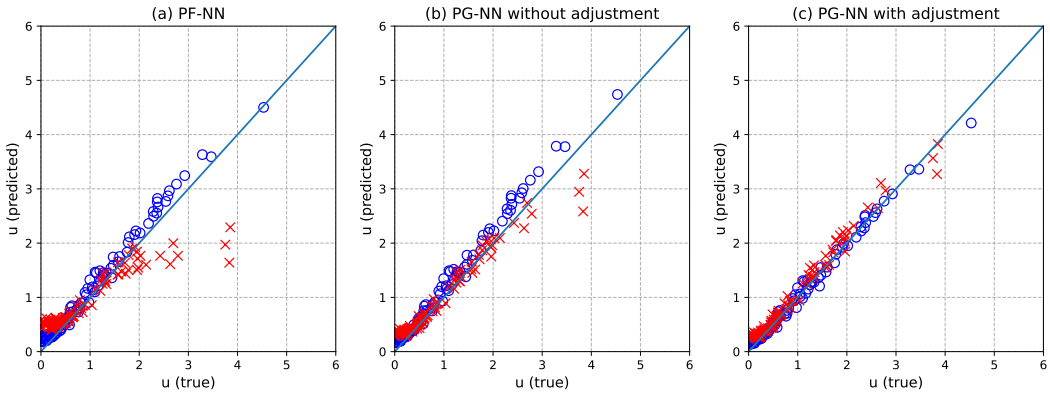


Figure 6.1: Predicted values of u_i from two replications (marked as o and x for each) when u_i is generated from the Gamma distribution with $\lambda = 1$, $n = 100$, $q = 100$.

same fixed parameter estimates but with different $\widehat{\beta}_0$ *and* $\widehat{\beta}_0^*$ *for* β_0*, respectively.*

Theorem 1 shows that the adjustment (6.7) improves the random effect prediction. According to our experience, even though limited, this adjustment becomes important, especially when the cluster size is large. Figure 6.1 is the plot of $\widehat{u}_i$ against the true u_i under $(n, q) = (100, 100)$ and $\lambda = 1$. Figure 6.1 (a) shows that the fixed effect estimator of subject-specific effects using Poisson DNN (PF-NN) produces poor prediction of u_i. Figure 6.1 (b) and (c) show that random effect prediction (PG-NN) improves the subject-specific prediction, and the proposed adjustment improves it further.

6.4.3 Pretraining variance components

We found that the MLE for $\lambda = \text{var}(u_i)$ could be sensitive to the choice of initial value, giving a slow convergence. We propose the use of method-of-

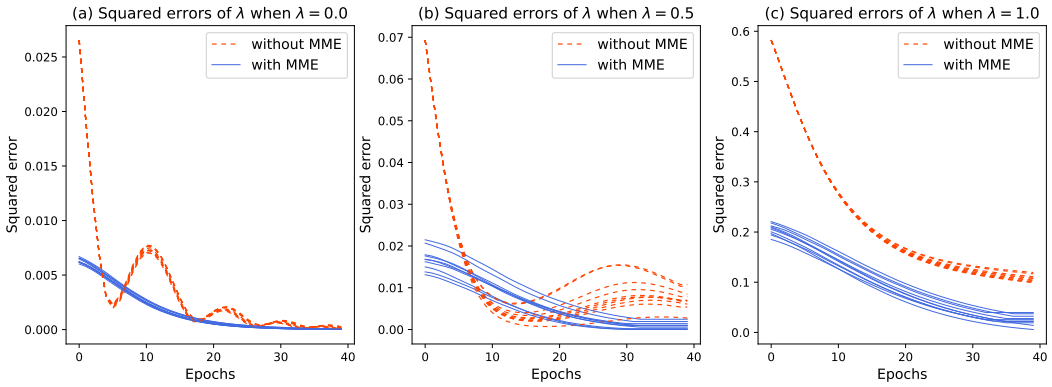


Figure 6.2: Learning curve for the variance component λ when (a) $\lambda = 0$, (b) $\lambda = 0.5$, and (c) $\lambda = 1$.

moments estimator (MME) for pretraining λ,

$$\widehat{\lambda} = \left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{u}_i - 1)^2\right]\left[\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{n\sum_i^n \widehat{\mu}_{i+}^{-1}(\widehat{u}_i - 1)^2}{\{\sum_i^n(\widehat{u}_i - 1)^2\}^2}}\right]. \tag{6.8}$$

Convergence of the method-of-moments estimator (6.8) is shown in Appendix 6.8.1. Figure 6.2 shows that the proposed pretraining procedure accelerates the convergence in various settings. The entire learning algorithm of the proposed method is briefly described in Algorithm 2.

6.5 Experimental Studies

To investigate the performance of the Poisson-gamma DNN, we conducted experimental studies. The five input variables $\mathbf{x}_{ij} = (x_{1ij}, ..., x_{5ij})^T$ are generated from the AR(1) process with autocorrelation $\rho = 0.5$ for each $i = 1, ..., n$ and $j = 1, ..., q$. The random effects are generated from either $u_i \sim \text{Gamma}(\lambda^{-1}, \lambda^{-1})$ or $v_i \sim \text{N}(0, \lambda)$ where $v_i = \log u_i$. In normal random ef-

Algorithm 2 Learning algorithm for Poisson-gamma DNN via h-likelihood

Input: $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$

< **Stage 1** >
for epoch = 0 **to** method-of-moments epochs **do**
  Train $\mathbf{w}$, $\boldsymbol{\beta}$ and $\mathbf{v}$ by minimizing the negative h-likelihood.
  Compute method-of-moments estimator of λ.
  Adjust the random effect predictors.
end for

< **Stage 2** >
for epoch = 0 **to** maximum epochs **do**
  Train all the fixed and random parameters by minimizing the negative h-likelihood.
  Adjust the random effect predictors.
end for

Compute MLE of λ.

fects (Lee and Nelder, 1996), it is common that the constraint $\mathrm{E}(v_i) = 0$ is imposed. When $\lambda = 0$, the conditional mean μ_{ij}^c becomes the marginal mean μ_{ij}^m. The output variable y_{ij} is generated from a Poisson distribution with the conditional mean

$$\mu_{ij}^c = u_i \cdot \exp\left\{0.2 + 0.2(\cos x_{1ij} + \cos x_{2ij} + \cos x_{3ij}) + \frac{0.2}{x_{4ij}^2 + 1} + \frac{0.2}{x_{5ij}^2 + 1}\right\}.$$

Results are based on the 100 sets of simulated data. The data consist of $q = 10$ observations for $n = 1,000$ subjects. For each subject, 6 observations are assigned to the training set, 2 are assigned to the validation set, and the remaining 2 are assigned to the test set. To evaluate the prediction perfor-

mances, the root mean squared Pearson residuals (RMSR) for Poisson outputs and the root mean squared errors (RMSE) for continuous outputs are often considered,

$$\text{RMSR} = \sqrt{\sum_{(i,j)\in\text{test}} \frac{(y_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij} N_{\text{test}}}} \quad \text{and} \quad \text{RMSE} = \sqrt{\sum_{(i,j)\in\text{test}} \frac{(y_{ij} - \widehat{\mu}_{ij})^2}{N_{\text{test}}}},$$

respectively. Note that RMSE is the RMSR for continuous outputs, because $\text{var}(y_{ij})$ is constant in the Gaussian distribution, whereas $\text{var}(y_{ij}) = \mu_{ij}$ in the Poisson distribution. The RMSE could be used for count outputs, whereas the RMSR would be used for count outputs. We report both for comparison.

For comparison, we consider the following models.

- **P-GLM** : Classic Poisson GLM for count data using `R`.
- **N-NN** : Conventional DNN for continuous data.
- **P-NN** : Poisson DNN for count outputs.
- **PN-GLM** : Poisson-normal HGLM using `lme4` (Bates et al., 2015).
- **PG-GLM** : Poisson-gamma HGLM using the proposed method.
- **NF-NN** : N-NN with fixed subject-specific effects for continuous data.
- **NN-NN** : N-NN with normal random effects for continuous data.
- **PF-NN** : P-NN with fixed subject-specific effects for count data.
- **PG-NN** : The proposed Poisson-gamma DNN for count data.

P-GLM, N-NN, and P-NN are used for marginal prediction, while the others are used for subject-specific prediction. N-NN, NF-NN, and NN-NN are models for continuous outputs, whereas the remaining models are designed for count outputs. For NF-NN and PF-NN, MLEs are obtained by maximizing the conditional likelihood $\sum_{i,j} \log f_{\boldsymbol{\theta}}(y_{ij}|v_i)$. On the other hand, for PN-GLM, PG-GLM, NN-NN, and PG-NN, subject-specific predictions are made by maximizing the h-likelihood.

PN-GLM is the generalized linear mixed model with random effects $v_i \sim N(0, \lambda)$. Current statistical software for PN-GLM and PG-GLM (`lme4` and `dhglm`) cannot provide the exact MLEs because of the intractable integration for computing the marginal likelihood (6.5) using the extended likelihood (6.3). The proposed PG-NN can be used to give exact MLEs for PG-GLM. Among various methods for NN-NN (Lee and Lee, 2023; Mandel et al., 2021; Simchoni and Rosset, 2021, 2023; Tran et al., 2020), we applied the state-of-the-art method proposed by Lee and Lee (2023). All the DNNs and PG-GLMs were implemented in Python using Keras (Chollet et al., 2015) and TensorFlow (Abadi et al., 2015). For all DNNs, we employed a standard multi-layer perceptron (MLP) consisting of 3 hidden layers with 10 neurons and leaky ReLU activation function. We applied the Adam optimizer with a learning rate of 0.001 and an early stopping process based on the validation loss while training the DNNs. NVIDIA Quadro RTX 6000 were used for computations.

Table 6.1 shows the average of test RMSEs and RMSRs from the experimental studies. When the true model does not have random effects (G(0)), the PG-NN is comparable to the P-NN, which should perform the best (marked by the bold face) in terms of RMSE and RMSR. N-NN (P-NN) is also better than

Table 6.1: Averages of test RMSEs and RMSRs of simulation studies over 100 replications of each scenario. G(0) implies the absence of random effects, i.e., $v_i = 0$ for all i. Bold numbers are the minimum values.

Model		Distribution of random effects (λ)				
		G(0)	G(0.5)	G(1)	N(0.5)	N(1)
P-GLM	RMSE	1.560	2.240	2.751	2.954	5.381
	RMSR	1.046	1.501	1.839	1.747	2.804
N-NN	RMSE	1.505	2.206	2.727	2.914	2.407
	RMSR	1.013	1.473	1.816	1.713	1.143
P-NN	RMSE	**1.503**	2.205	2.727	2.913	2.469
	RMSR	**1.011**	1.470	1.812	1.711	1.161
PN-GLM	RMSE	1.561	1.680	1.727	1.945	5.351
	RMSR	1.048	1.106	1.105	1.118	2.772
PG-GLM	RMSE	1.561	1.704	1.753	1.978	2.469
	RMSR	1.048	1.123	1.106	1.139	1.161
NF-NN	RMSE	1.638	1.666	1.748	2.021	4.664
	RMSR	1.152	1.301	1.136	1.241	1.256
NN-NN	RMSE	1.516	1.785	2.062	2.360	5.354
	RMSR	1.020	1.121	1.209	1.256	2.773
PF-NN	RMSE	1.629	1.634	1.653	1.854	**2.183**
	RMSR	1.147	1.135	1.128	1.129	1.128
PG-NN	RMSE	1.507	**1.622**	**1.647**	**1.850**	2.280
	RMSR	1.016	**1.079**	**1.084**	**1.061**	**1.085**

NF-NN and NN-NN (PF-NN and PG-NN). When the distribution of random effects is correctly specified (G(0.5) and G(1)), the PG-NN performs the best. Even when the distribution of random effects is misspecified (N(0.5), N(1)), the PG-NN performs the best in terms of RMSR. This result is in accordance with the simulation results of McCulloch and Neuhaus (2011), namely, in GLMMs, the prediction accuracy is little affected for violations of the distributional assumption for random effects. However, in N(1), the PF-NN performs the best in terms of RMSE. We believe that in count data, the RMSR should be considered as a performance measure because the variance increases with the mean, and the RMSE is sensitive to large prediction values.

6.6 Real Data Analysis

To investigate prediction performance for clustered count outputs in real data, we considered the following datasets:

- **Epilepsy data:** Epilepsy data are reported by Thall and Vail (1990) from a clinical trial of $n = 59$ patients with epilepsy. The data contain $N = 236$ observations with $q_i = 4$ repeated measures from each patient and $p = 4$ input variables.

- **CD4 data:** CD4 data are from a study of AIDS patients with advanced immune suppression, reported by Henry et al. (1998). The data contain $N = 4612$ observations from $n = 1036$ patients with $q_i \geq 2$ repeated measurements and $p = 4$ input variables.

- **Bolus data:** Bolus data are from a clinical trial following abdominal

surgery for $n = 65$ patients with $q_i = 12$ repeated measurements, reported in Henderson and Shimakura (2003). The data have $N = 780$ observations with $p = 2$ input variables.

- **Owls data:** Owls data are reported by Roulin and Bersier (2007), which can be found in the `R` package `glmmTMB` (Brooks et al., 2023). The data contain $N = 599$ observations and $n = 27$ nests with $p = 3$ input variables. The cluster size q_i in each nest varies from 4 to 52.

- **Fruits data:** Fruits data are reported in Banta et al. (2010). The data have $N = 625$ observations clustered by $n = 24$ types of maternal seed family with $p = 3$ input variables. The cluster size q_i varies from 11 to 47.

For all the DNNs, a standard MLP with one hidden layer of 10 neurons and a sigmoid activation function were employed. For longitudinal data (Epilepsy, CD4, Bolus), the last observation for each patient was used as the test set. For clustered data (Owls, Fruits), an observation was randomly selected as the test set from each cluster. RMSRs and RMSEs are reported in Table 6.2. Except for Fruits data, non-linear effects may not improve the marginal prediction from the P-GLM. DNNs are widely acknowledged for improving predictions in large-sized datasets. We faced challenges in finding large-sized count data, but our analyses show that the non-linear effects improve the subject-specific predictions, maintaining the superior performance of the PG-NN. This implies that introducing subject-specific random effects in DNNs helps to identify the nonlinear effects of the input variables in moderately sized data. For Fruits data, PG-GLM performs the best in terms of RMSR. However, PF-NN has

the best performance in RMSE. PG-NN performs comparably.

6.7 Concluding Remarks

Lee and Lee (2023) showed that the h-likelihood provides the most efficient likelihood-based subject-specific procedure for continuous outputs (NN-NN). For non-normal outputs, the Laplace approximation has often been used to obtain approximate MLEs. With the new h-likelihood (6.4), both MLEs of fixed parameters and BUPs of the random effects can be directly obtained from the single objective function. This enables a fast end-to-end learning algorithm. By introducing subject-specific random effects, DNNs can effectively identify the nonlinear effects of the input variables for moderately-sized data. Though we focus on introducing the h-likelihood procedure for clustered count data and use the standard MLP for experimental studies and real data analyses, the proposed method (PG-NN) can be adapted to the state-of-the-art network architectures by using the negative h-likelihood as the loss function.

6.8 Appendix

6.8.1 Convergence of the method-of-moments estimator

As derived in Section 6.8.2, for given λ and μ_{i+}, maximization of the h-likelihood leads to

$$\widehat{u}_i = \widehat{u}_i(\mathbf{y}_i) = \exp\left(\widehat{v}_i(\mathbf{y}_i)\right) = \frac{y_{i+} + \lambda^{-1}}{\mu_{i+} + \lambda^{-1}}.$$

Table 6.2: Test RMSEs and RMSRs of real data analyses. Bold numbers are the minimum values.

Model		Dataset				
		Epilepsy	CD4	Bolus	Owls	Fruits
P-GLM	RMSE	4.951	36.57	4.646	5.856	39.19
	RMSR	1.520	6.115	2.110	2.307	6.818
N-NN	RMSE	7.522	37.51	5.071	5.850	27.93
	RMSR	2.119	8.516	1.982	2.297	6.573
P-NN	RMSE	6.072	39.06	4.782	6.110	30.01
	RMSR	1.712	6.830	2.354	3.076	6.854
PN-GLM	RMSE	3.943	24.73	3.810	7.120	38.43
	RMSR	1.242	3.422	1.727	5.791	6.795
PG-GLM	RMSE	4.009	37.14	3.799	6.692	28.83
	RMSR	1.229	4.424	1.714	4.479	**5.786**
NF-NN	RMSE	6.663	36.04	3.847	5.560	27.07
	RMSR	1.750	6.921	1.718	2.215	5.897
NN-NN	RMSE	6.744	36.53	3.872	5.539	27.80
	RMSR	1.770	7.640	1.727	2.674	5.825
PF-NN	RMSE	3.890	23.77	3.865	5.992	**26.60**
	RMSR	1.238	3.558	1.816	2.951	6.430
PG-NN	RMSE	**3.126**	**23.17**	**3.771**	**5.215**	28.83
	RMSR	**1.135**	**3.513**	**1.677**	**2.000**	6.376

Thus, $\mathrm{E}(\widehat{u}_i) = 1$ and $\mathrm{Var}(\widehat{u}_i) = \lambda\{1 - (\lambda\mu_{i+} + 1)^{-1}\}$. Define d_i as

$$d_i = \frac{\widehat{u}_i - 1}{\sqrt{1 - (\lambda\mu_{i+} + 1)^{-1}}} = \frac{y_{i+} - \mu_{i+}}{\mu_{i+} + \lambda^{-1}}\sqrt{1 + \lambda^{-1}\mu_{i+}^{-1}}$$

to have $\mathrm{E}(d_i) = 0$ and $\mathrm{Var}(d_i) = \lambda$ for any $i = 1, ..., n$. Then, by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 \to \mathrm{E}(d_i^2) = \mathrm{Var}(d_i) + \mathrm{E}(d_i)^2 = \lambda.$$

Note here that

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \left\{\frac{1}{n}\sum_{i=1}^{n}(\widehat{u}_i - 1)^2\right\} + \frac{1}{\lambda}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{(\widehat{u}_i - 1)^2}{\mu_{i+}}\right\}.$$

Then, solving the following equation,

$$\lambda - \left\{\frac{1}{n}\sum_{i=1}^{n}(\widehat{u}_i - 1)^2\right\} - \frac{1}{\lambda}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{(\widehat{u}_i - 1)^2}{\widehat{\mu}_{i+}}\right\} = 0,$$

leads to an estimate $\widehat{\lambda}$ in (6.8) and $\widehat{\lambda} \to \lambda$ as $n \to \infty$.

6.8.2 Technical details

Let $\widetilde{v}_i = \arg\max_{v_i}\{\log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})\}$. Then maximization of the $\log f_{\boldsymbol{\theta}}(\mathbf{v}|\mathbf{y})$ yields

$$\tilde{v}_i = \arg\max_{v_i}\left[\sum_{j=1}^{q_i}\left(y_{ij}v_i - \mu_{ij}^m e^{v_i}\right) + \frac{v_i - e^{v_i}}{\lambda}\right] = \log\left(\frac{y_{i+} + \lambda^{-1}}{\mu_{i+} + \lambda^{-1}}\right),$$

where $y_{i+} = \sum_{j=1}^{q_i} y_{ij}$ and $\mu_{i+} = \sum_{j=1}^{q_i} \mu_{ij}^m$. This leads to

$$\log f(\widetilde{\mathbf{v}}^*|\mathbf{y}) = \sum_{i=1}^{n} \{\log f_{\boldsymbol{\theta}}(\widetilde{v}_i|\mathbf{y}) + c_i(\boldsymbol{\theta}; \mathbf{y}_i)\} = 0.$$

Thus, maximization of the h-likelihood gives MLEs for fixed parameters,

$$\arg\max_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \widetilde{\mathbf{v}}) = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{y}).$$

Furthermore, from the distribution of $u_i|\mathbf{y}_i$,

$$\widetilde{u}_i = \exp(\widetilde{v}_i) = \frac{y_{i+} + \lambda^{-1}}{\mu_{i+} + \lambda^{-1}} = \mathrm{E}(u_i|\mathbf{y}_i)$$

and

$$\widetilde{\mu}_{ij}^c = \exp(\widetilde{v}_i) \cdot \mathrm{NN}(\mathbf{X}; \mathbf{w}, \boldsymbol{\beta}) = \mu_{ij}^m \cdot \mathrm{E}(u_i|\mathbf{y}_i) = \mathrm{E}(\mu_{ij}^c|\mathbf{y})$$

become the BUPs of u_i and μ_{ij}^c, respectively.

Proof of Theorem 1.

The adjustment (6.7) transports

$$\widehat{u}_i^* = \widehat{u}_i/(1+\epsilon) \quad \text{and} \quad \widehat{v}_i^* = \widehat{v}_i - \log(1+\epsilon).$$

Since $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})$ and $(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*)$ have the same conditional expectation $\widehat{\mu}_{ij}$, equation (6.4) yields

$$\begin{aligned} h(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*) - h(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}}) &= \sum_{i=1}^{n} \left[\frac{\widehat{v}_i^* - \exp(\widehat{v}_i^*)}{\widehat{\lambda}} \right] - \sum_{i=1}^{n} \left[\frac{\widehat{v}_i - \exp(\widehat{v}_i)}{\widehat{\lambda}} \right] \\ &= \widehat{\lambda}^{-1} \left\{ \sum_{i=1}^{n} (\widehat{v}_i^* - \widehat{v}_i) - \sum_{i=1}^{n} (\widehat{u}_i^* - \widehat{u}_i) \right\} \\ &= n\widehat{\lambda}^{-1} \left\{ \epsilon - \log(1+\epsilon) \right\} \geq 0, \end{aligned}$$

and the equality holds if and only if $\epsilon = 0$. Thus, $h(\widehat{\boldsymbol{\theta}}^*, \widehat{\mathbf{v}}^*) \geq h(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{v}})$.

Chapter 7

DNN for semi-parametric frailty models via h-likelihood

7.1 Introduction

Recently, deep neural network (DNN) has provided a major breakthrough to enhance prediction in various areas (Goodfellow et al., 2016; LeCun et al., 2015). The DNN models allow extensions of Cox proportional hazards (PH) models (Kvamme et al., 2019; Sun et al., 2020). Recently, subject-specific prediction of the DNN models has been studied by including random effects in neural network (NN) predictor (Mandel et al., 2021; Tran et al., 2020). However, these DNN random-effect models have been studied for only complete data. In this chapter, we propose a new DNN-FM. To the best of our knowledge, there is no literature on the DNN-FM for censored survival data. Lee and Nelder (1996) introduced the h-likelihood for the inference of general models with random effects and Ha et al. (2001) extended it to the semi-parametric frailty mod-

els. We reformulate the h-likelihood to obtain maximum likelihood estimators (MLEs) for fixed unknown parameters and best unbiased predictors (BUPs) for random frailties (Searle et al., 1992) by a simple joint maximization of the profiled h-likelihood (Lee et al., 2017), which is constructed by profiling out the non-parametric baseline hazard for semi-parametric DNN-FMs. Thus, the proposed DNN-FM can be trained by using a negative profiled h-likelihood as a loss function. Experimental studies show that the proposed method enhances the prediction performance of the existing DNN-Cox and FM in terms of Brier score and C-index, which are popular predictive measures in survival analysis.

In Section 7.2, we review the DNN-Cox model. We propose the DNN-FM and introduce its h-likelihood in Section 7.3 and learning algorithm in Section 7.4. The experimental study is presented to compare its predictive performance with various methods in Section 7.5. A real data analysis is in Section 7.6, followed by concluding remarks in Section 7.7. A theoretical framework for an online learning and all the technical details are in Appendix 7.8.

7.2 A review of DNN-Cox model

7.2.1 DNN-Cox model

Let T_i be the survival time (time-to-event) for subject $i = 1, \ldots, n$, and let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ be a p-dimensional vector of input variables (covariates or features). The semi-parametric Cox model has a hazard function of T_i,

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t)\exp(\eta_i), \quad \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} \tag{7.1}$$

where $\lambda_0(\cdot)$ is a non-parametric baseline hazard function, the linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ is a parametric model for risk function (or risk score) of covariates $\mathbf{x}_i$, and $\boldsymbol{\beta}$ is a vector of p-dimensional regression parameters without intercept (or bias) term. The survival function for T_i given $\mathbf{x}_i$ is

$$S(t|\mathbf{x}_i) = P(T_i > t|\mathbf{x}_i) = \exp\{-\Lambda_0(t)e^{\eta_i}\},$$

where $\Lambda_0(t)$ is the baseline cumulative hazard.

The Cox model (7.1) is extended to the DNN-Cox model, by relaxing the parametric linear model $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ with a non-linear function of $\mathbf{x}_i$,

$$\eta_i = \sum_{k=1}^{p_L} \beta_k g_k^{(L)}(\mathbf{x}_i; \mathbf{w}) = \mathrm{NN}(\mathbf{x}_i; \mathbf{w}, \boldsymbol{\beta}), \tag{7.2}$$

where $\mathrm{NN}(\cdot)$ denotes neural network risk predictor of the output layer with the last hidden layer; $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{p_L})^T$ is a vector of the output weights, with p_L number of nodes of the Lth hidden layer; and $\mathbf{w} = (\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_L^T)^T$ is a combined vectorization consisting of a vector $\mathbf{w}_l$ of the lth hidden weights. Here, the $g_k^{(L)}(\mathbf{x}_i; \mathbf{w})$ is the k-th node of the last hidden layer $g^{(L)}(\mathbf{x}_i; \mathbf{w}) = (g_1^{(L)T}(\mathbf{x}_i; \mathbf{w}), \ldots, g_{p_L}^{(L)T}(\mathbf{x}_i; \mathbf{w}))^T$, which depends on the input variables $\mathbf{x}_i$ and the hidden weights $\mathbf{w}$, and the last hidden layer can be expressed as the form of compositional functions

$$g^{(L)}(\mathbf{x}_i; \mathbf{w}) = \sigma^{(L)}(\cdots \sigma^{(2)}(\sigma^{(1)}(\mathbf{x}_i; \mathbf{w}_1); \mathbf{w}_2) \cdots ; \mathbf{w}_L),$$

where $\sigma^{(l)}(\cdot)$ denotes the activation function of hidden layer.

In survival analysis, the observable random variables are, for $i = 1, \ldots, n$,

$$y_i = \min(T_i, C_i) \text{ and } \delta_i = I(T_i \leq C_i)$$

where C_i is the censoring time corresponding to T_i. The DNN weights $(\mathbf{w}, \boldsymbol{\beta})$ in (7.2) can be estimated by maximizing the Breslow's log-likelihood (Breslow, 1972; Kvamme et al., 2019; Tarkhan and Simon, 2022),

$$\ell = \sum_i \delta_i \eta_i - \sum_k d_{(k)} \log\Big\{ \sum_{i \in R_{(k)}} \exp(\eta_i) \Big\}, \tag{7.3}$$

where $\eta_i = \text{NN}(\mathbf{x}_i; \mathbf{w}, \boldsymbol{\beta})$ is the NN predictor which also represents an output node of the DNN-Cox model, $R_{(k)} = \{i : y_i \geq y_{(k)}\}$ is the risk set at time $y_{(k)}$ which is the kth smallest distinct event time among the y_i's, and $d_{(k)}$ is the number of events at $y_{(k)}$.

7.2.2 Prediction measures

For censored data, the Brier score and concordance index (C-index) have been widely used to evaluate the predictive performance of DNN-Cox model (7.2) (Kvamme et al., 2019).

Brier score

The time-dependent Brier score is defined as

$$BS(t) = E\left\{I(t) - S(t|\mathbf{x})\right\}^2,$$

where $BS(t)$ is the mean squared error of the difference between $I(t)$ and $S(t|\mathbf{x})$. Here, $I(t)$ is the event status at the time point t (i.e. $I(t) = I(T > t) = 1$ if $T > t$ and 0 otherwise) and $S(t|\mathbf{x})$ is a model-based survival function. The estimated Brier score (Graf et al., 1999) is given by

$$\widehat{BS}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_i(t)\left\{y_i(t) - \widehat{S}(t|\mathbf{x}_i)\right\}^2,$$

where $y_i(t) = I(y_i > t)$ at a specific time point t and $\widehat{S}(t|\mathbf{x}_i)$ is estimated survival function given $\mathbf{x}_i$. Here, $\widehat{w}_i(t)$ is the inverse probability of censoring weights (IPCW),

$$\widehat{w}_i(t) = \frac{(1 - y_i(t))\delta_i}{\widehat{G}(y_i)} + \frac{y_i(t)}{\widehat{G}(t)},$$

where $\widehat{G}(t) = \widehat{P}(C > t)$ indicates the estimated survival function of censoring time. Thus, the estimated Brier score can be viewed as the mean squared error between the observed event status $y_i(t)$ and the predicted survival function $\widehat{S}(t|\mathbf{x}_i)$. The lower Brier score indicates a better predictive performance. For overall predictive performance, the integrated Brier score (IBS) is widely used with the maximum survival time t_{max},

$$\text{IBS} = \frac{1}{t_{max}}\int_0^{t_{max}} \text{BS}(s)ds.$$

C-index

The definition of C-index is based on the property that a survival model should predict a shorter survival time for subjects that fail earlier and a longer survival time for subjects that fail later. Let T_i and T_j be independent survival times with corresponding covariate vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Then the C-index

is defined by

$$C = P(S(t|\mathbf{x}_i) > S(t|\mathbf{x}_j)|T_i > T_j) = P(\eta_i < \eta_j|T_i > T_j).$$

where $\eta_k = \text{NN}(\mathbf{x}_k; \mathbf{w}, \boldsymbol{\beta})$ are the NN predictors of the DNN-Cox model (7.2). Following Harrell Jr et al. (1996), the C-index can be estimated by

$$\widehat{C} = \frac{\sum_i \sum_j \delta_i I(y_i < y_j)\{I(\widehat{\eta}_i > \widehat{\eta}_j) + 0.5I(\widehat{\eta}_i = \widehat{\eta}_j)\}}{\sum_i \sum_j \delta_i I(y_i < y_j)},$$

where $\widehat{\eta}_k = \text{NN}(\mathbf{x}_k; \widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\beta}})$. The range of C-index is from 0 to 1, and a larger value indicates a better performance.

7.3 Proposed DNN for frailty model

The FMs have been introduced for prediction of clustered survival time. Consider a clustered survival dataset

$$D_N = \{(y_{ij}, \delta_{ij}, \mathbf{x}_{ij}), i = 1, \ldots, n; j = 1, \ldots, n_i\},$$

where $y_{ij} = \min(T_{ij}, C_{ij})$ is the jth observation of the ith subject (or cluster), T_{ij} and C_{ij} are the corresponding survival and censoring times, respectively, and $\delta_{ij} = I(T_{ij} \leq C_{ij})$ is censoring indicator, and $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ is a vector of p covariates corresponding to T_{ij}. Here, n is the number of clusters, n_i is cluster size and $N = \sum_{i=1}^{n} n_i$ is the total sample size. The dependency among T_{ij}'s can be modelled via a frailty in the hazard function. Let u_i denote the unobserved frailty of the ith cluster. Then, the semi-parametric FM has

the conditional hazard function,

$$\lambda_{ij}(t|u_i, \mathbf{x}_{ij}) = \lambda_0(t)\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})u_i = \lambda_0(t)\exp(\eta_{ij}), \quad \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i, \tag{7.4}$$

where η_{ij} is linear predictor and $v_i = \log u_i$.

7.3.1 DNN-FM

The FM (7.4) is extended to a new DNN-FM by replacing $\mathbf{x}_{ij}^T\boldsymbol{\beta}$ with

$$\mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta}) = \sum_{k=1}^{p_L} g_k^{(L)}(\mathbf{x}_{ij}; \mathbf{w})\beta_k. \tag{7.5}$$

In this chapter, we assume the gamma distribution for the frailty u_i with $\mathrm{E}(u_i) = 1$ and $\mathrm{var}(u_i) = \alpha$, which is denoted by $\mathrm{Gamma}(1/\alpha, 1/\alpha)$. Figure 7.1 presents a schematic diagram of architecture of the DNN-FM, which is constructed by allowing for output nodes $\mathrm{NN}(\mathbf{x}_{ij}; \widehat{\mathbf{w}}, \widehat{\boldsymbol{\beta}})$ and $\widehat{u}_i$ from the two separate input layers, namely input vector $\mathbf{x}_{ij}$ and one-hot encoding vector of subjects $\mathbf{z}_i$, respectively. In the DNN-FM, subject-specific prediction can be made by multiplying the risk predictor $\exp\{\mathrm{NN}(\mathbf{x}_{ij}; \widehat{\mathbf{w}}, \widehat{\boldsymbol{\beta}})\}$ and frailty predictors $\widehat{u}_i$.

7.3.2 Construction of h-likelihood

In the FM (7.4), it is important to define the likelihood to obtain the exact MLEs for fixed parameters and BUPs for random frailties. Let $y = \min(T, C)$, $\mathbf{y}^* = (y, \delta)$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \alpha)^T$, and $\boldsymbol{\psi}$ be the vector of whole fixed parameters. Under the conditional independence and non-informative censoring given v_i,

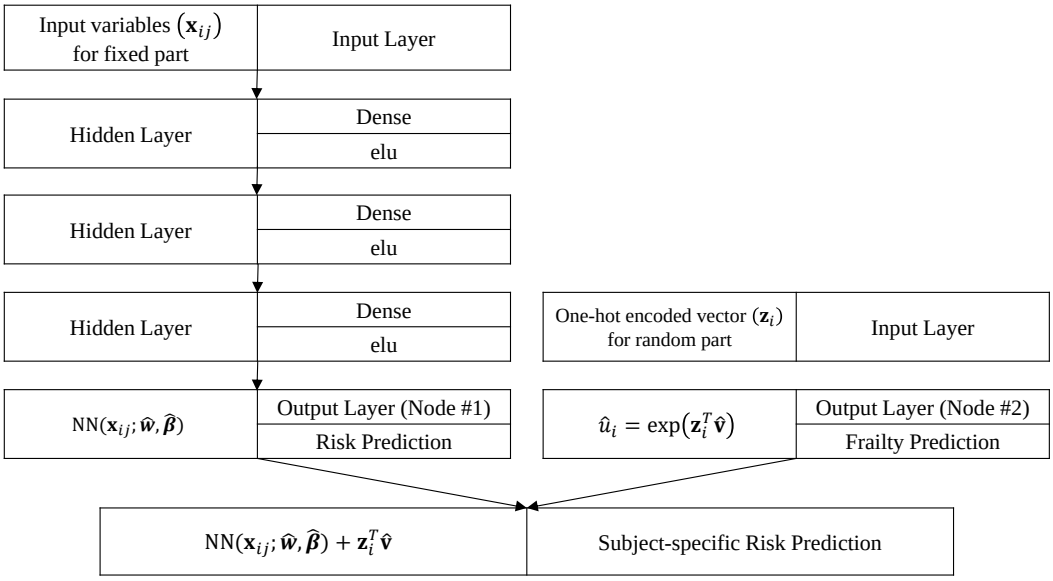


Figure 7.1: An example of model architecture for DNN-FM.

Ha et al. (2001) proposed the use of an extended likelihood

$$\ell_e(\boldsymbol{\psi}, \mathbf{v}; \mathbf{y}^*, \mathbf{v}) = \sum_{i,j} \log f_{\boldsymbol{\psi}}(y_{ij}, \delta_{ij}|v_i) + \sum_i \log f_{\boldsymbol{\psi}}(v_i), \tag{7.6}$$

where

$$\begin{aligned}\log f_{\boldsymbol{\psi}}(y_{ij}, \delta_{ij}|v_i) &= \delta_{ij}\{\log \lambda_{ij}(y_{ij}|v_i)\} - \Lambda_{ij}(y_{ij}|v_i) \\ &= \delta_{ij}\{\log \lambda_0(y_{ij}) + \eta_{ij}\} - \Lambda_0(y_{ij}) \exp(\eta_{ij})\end{aligned}$$

is the conditional censored log-likelihood of y_{ij} and δ_{ij} given v_i, $\Lambda_0(\cdot)$ is the cumulative baseline hazard function, $f_{\boldsymbol{\psi}}(v_i)$ is a density function of v_i with the parameter $\boldsymbol{\psi}$ and $v_i = \log(u_i)$. Lee and Nelder's (1996) original h-likelihood was aimed to obtain MLEs for all fixed parameters and good predictors for

random effects by the joint maximization. However, their h-likelihood cannot give an exact MLE for variance component α. In this chapter, we introduce a new h-likelihood for the gamma FM (7.4). Consider an extended likelihood with $\mathbf{v}^c$ scale,

$$h(\boldsymbol{\psi}, \mathbf{v}^c) = \ell(\boldsymbol{\psi}; \mathbf{y}^*) + \log f_{\boldsymbol{\psi}}(\mathbf{v}^c|\mathbf{y}^*), \tag{7.7}$$

where $\ell(\boldsymbol{\psi}; \mathbf{y}^*) = \log \int f_{\boldsymbol{\psi}}(\mathbf{y}^*, \mathbf{v}) d\mathbf{v}$ is the marginal log-likelihood. Given $\boldsymbol{\psi}$, let

$$\tilde{\mathbf{v}}^c = \arg\max_{\mathbf{v}^c} h(\boldsymbol{\psi}, \mathbf{v}^c) = \arg\max_{\mathbf{v}^c} f_{\boldsymbol{\psi}}(\mathbf{v}^c|\mathbf{y}^*).$$

From (7.7), a sufficient condition for $h(\boldsymbol{\psi}, \mathbf{v}^c)$ to give the exact MLEs for $\boldsymbol{\psi}$ is that $f_{\boldsymbol{\psi}}(\tilde{\mathbf{v}}^c|\mathbf{y}^*)$ is independent of $\boldsymbol{\psi}$. Let

$$v_i^c = v_i \exp\left\{a_i(\alpha, \delta_{i+})\right\}, \tag{7.8}$$

where $a_i(\alpha, \delta_{i+}) = (\delta_{i+} + \alpha^{-1})\left(\log(\delta_{i+} + \alpha^{-1}) - 1\right) - \log\Gamma(\delta_{i+} + \alpha^{-1})$, then the predictive likelihood becomes

$$\log f(\tilde{\mathbf{v}}^c|\mathbf{y}^*) = \sum_{i=1}^{n} \log f_{\boldsymbol{\psi}}(\tilde{v}_i^c|\mathbf{y}^*) = \sum_{i=1}^{n} \left\{\log f_{\boldsymbol{\psi}}(\tilde{v}_i|\mathbf{y}^*) - a_i(\alpha, \delta_{i+})\right\} = 0,$$

which is free from $\boldsymbol{\psi}$. Thus, $\ell(\boldsymbol{\psi}, \mathbf{y}^*) = h(\boldsymbol{\psi}, \tilde{\mathbf{v}}^c)$. Let $h(\boldsymbol{\psi}, \mathbf{v})$ be a reparameterization of the h-likelihood (7.7), then

$$h = h(\boldsymbol{\psi}, \mathbf{v}) = \ell_e(\boldsymbol{\psi}, \mathbf{v}; \mathbf{y}^*, \mathbf{v}) + \log\left|\frac{d\mathbf{v}}{d\mathbf{v}^c}\right| = h(\boldsymbol{\psi}, \mathbf{v}^c),$$

where $\ell_e(\boldsymbol{\psi}, \mathbf{v}; \mathbf{y}^*, \mathbf{v})$ is the h-likelihood (7.6) of Ha et al. (2001) and the Jacobian term is $\log\left|\frac{d\mathbf{v}}{d\mathbf{v}^c}\right| = -\sum_{i=1}^{n} a_i(\alpha, \delta_{i+})$. Note here that $h(\boldsymbol{\psi}, \mathbf{v}) \neq \ell_e(\boldsymbol{\psi}, \mathbf{v}; \mathbf{y}^*, \mathbf{v})$.

Given $\boldsymbol{\psi}$, we have the BUP for $\mathbf{u}$, $\mathrm{E}(\mathbf{u}|\mathbf{y}^*)$, by solving $\partial h/\partial \mathbf{v} = 0$ (or $\partial h/\partial \mathbf{u} = 0$), where $\mathbf{u} = \exp(\mathbf{v})$. The joint maximization of the new h-likelihood gives MLEs for the whole fixed parameters including variance component and BUPs for the random frailties. Technical details are derived in Appendix 7.8.1.

7.4 Learning algorithm using the profiled h-likelihood

For the DNN-FM, the new h-likelihood is

$$h = h(\boldsymbol{\psi}, \mathbf{v}) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big[\delta_{ij}\{\log\lambda_0(y_{ij}) + \eta_{ij}\} - \Lambda_0(y_{ij})\exp(\eta_{ij})\Big] + \sum_{i=1}^{n}\left[\frac{\log u_i - u_i}{\alpha} - \alpha^{-1}\log\alpha - \log\Gamma(\alpha^{-1}) - a_i(\alpha, \delta_{i+})\right] \quad (7.9)$$

where

$$\eta_{ij} = \mathrm{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta}) + v_i$$

and $a_i(\alpha, \delta_{i+})$ is given in (7.8). For eliminating the non-parametric baseline hazard $\lambda_0(\cdot)$ in (7.9), following Ha et al. (2001), we have a profiled h-likelihood,

$$h_p = h_p(\boldsymbol{\theta}, \mathbf{v}) = h_{\mathrm{PL}} - \sum_{i=1}^{n} a_i(\alpha, \delta_{i+}), \quad (7.10)$$

where

$$h_{\mathrm{PL}} = \sum_{ij} \delta_{ij}\eta_{ij} - \sum_{k} d_{(k)} \log \left[\sum_{(i,j)\in R_{(k)}} \exp(\eta_{ij}) \right] + \sum_{i=1}^{n} \left[\frac{\log u_i - u_i}{\alpha} - \frac{\log \alpha}{\alpha} - \log \Gamma\left(\frac{1}{\alpha}\right) \right]$$

is the penalized partial likelihood (Ripatti and Palmgren, 2000; Therneau et al., 2003). Here, $R_{(k)} = \{(i,j) : y_{ij} \geq y_{(k)}\}$ is risk set at time $y_{(k)}$, and $d_{(k)}$ is the number of events at $y_{(k)}$ which is the kth smallest distinct event times among the y_{ij}'s. However, direct maximization of the penalized partial likelihood cannot provide MLEs. To obtain the MLEs, Gu et al. (2004) proposed the use of the marginal partial log-likelihood

$$\ell_p = \log \int \exp(h_{\mathrm{PL}}) dv.$$

However, this integration is often numerically intractable. Ha et al. (2017, 2001) and Ripatti and Palmgren (2000) proposed the use of the Laplace approximation of ℓ_p, but it is still numerically difficult and does not give the exact MLEs. The Laplace approximation can yield a biased estimation for frailty models with a small cluster size or under heavy censoring (Gorfine and Zucker, 2023; Jeon et al., 2012).

An advantage of the h-likelihood approach is that the nuisance parameters associated with the non-parametric hazard $\lambda_0(\cdot)$ can be eliminated by profiling. Since the joint maximization of h_p gives the MLEs for fixed parameters and BUPs for random frailties, the DNN-FM can be trained by using the negative profiled h-likelihood $-h_p$ as loss function, which contains $\mathrm{NN}(\mathbf{x}; \mathbf{w}, \boldsymbol{\beta})$ and u_i.

7.4.1 Local minima problem

In the FM, we impose the constraints $\mathrm{E}(u_i) = 1$ for identifiability. For any ϵ,

$$\begin{aligned}\lambda_{ij}(t|u_i, \mathbf{x}_{ij}) &= \lambda_0(t) \exp\left\{\mathrm{NN}\ \ (\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})\right\} u_i \\ &= \lambda_0(t) \exp\left\{\mathrm{NN}\left(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta}\right) + \epsilon\right\} \left(u_i / \exp(\epsilon)\right).\end{aligned}$$

However, DNN models often encounter local minima which violates the constraints. This causes a computational difficulty in the DNN-FM. To prevent poor prediction due to the local minima, we introduce an adjustment on the predictor of u_i,

$$\widehat{u}_i \leftarrow \frac{\widehat{u}_i}{\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i} \tag{7.11}$$

to satisfy

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = 1.$$

7.4.2 ML learning algorithm

We propose a h-likelihood learning algorithm:

- **Inner loop:** Given $\widehat{\alpha}$, find $(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}})$ under a loss function $-h_p$ in (7.10).
- **Adjustment:** Transport $\widehat{\mathbf{u}}$ as in (7.11).
- **Outer loop:** Given $(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}})$, find $\widehat{\alpha}$ under a loss function $-h_p$ in (7.10).

This algorithm describes double loop iterative procedures with an additional adjustment on the frailty predictors: for details, see Algorithm 3. Figure 7.2 displays a schematic diagram of the h-likelihood learning procedure of the DNN-FM.

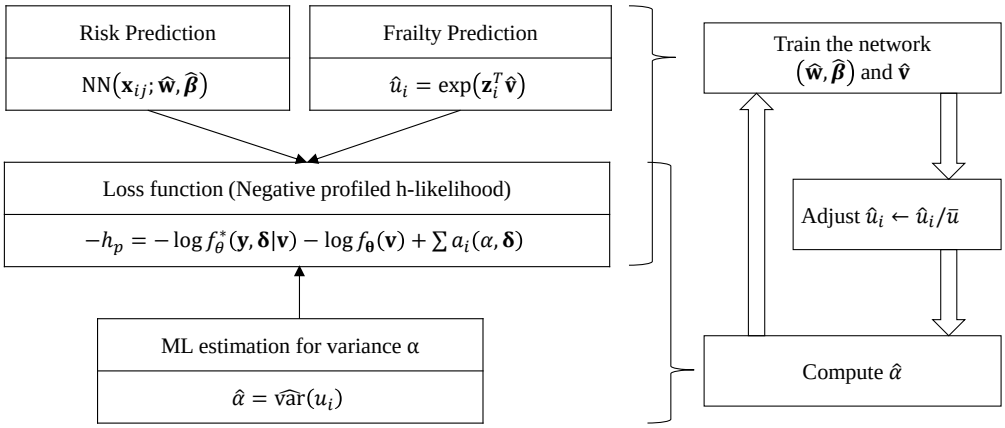


Figure 7.2: A schematic diagram of DNN-FM fitting procedure.

Algorithm 3 H-likelihood Learning Algorithm.

REPEAT UNTIL α CONVERGES:

TRAIN THE NETWORK:

$$\widehat{\mathbf{w}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{v}} \leftarrow \arg\min_{\mathbf{w},\boldsymbol{\beta},\mathbf{v}} \{-h_p(\mathbf{w},\boldsymbol{\beta},\widehat{\alpha},\mathbf{v})\}$$

return $\widehat{\mathbf{w}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{v}}$

ADJUST THE FRAILTIES:

$$\bar{u} \leftarrow \textstyle\sum_{i=1}^{n} \exp(\widehat{v}_i)/n$$

$$\widehat{v}_i \leftarrow \widehat{v}_i - \log \bar{u} \quad \textbf{for} \quad i = 1,...,n$$

return $\widehat{\mathbf{v}}$

COMPUTE VARIANCE COMPONENT:

$$\widehat{\alpha} \leftarrow \arg\min_{\alpha} \left\{-h_p(\widehat{\mathbf{w}},\widehat{\boldsymbol{\beta}},\alpha,\widehat{\mathbf{v}})\right\}$$

return $\widehat{\alpha}$

7.5 Experimental studies

To evaluate the performance of the proposed method, experimental studies are conducted based on 100 replications of simulated data. We use the extended forms of the IBS and C-index of FMs (Van Oirbeek and Emmanuel, 2016; Van Oirbeek and Lesaffre, 2010) as performance measures. Details are derived in Appendix 7.8.2.

7.5.1 Experimental design

Given u_i and $\mathbf{x}_{ij}$, survival times T_{ij} are generated from the hazard function

$$\lambda_{ij}(t|u_i, \mathbf{x}_{ij}) = \lambda_0(t) \exp\{f(\mathbf{x}_{ij})\} u_i,$$

where $f(\mathbf{x}_{ij})$ is an unknown true risk function of x_{ij} and $\lambda_0(t) = \phi t^{\phi-1}$ is set to be a Weibull baseline hazard with shape parameter $\phi = 2$. The DNN-FM fits the true but unknown $f(\mathbf{x}_{ij})$ by $\text{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})$. Here, the five input variables $\mathbf{x}_{ij} = (x_{1ij}, ..., x_{5ij})^T$ are generated from AR(1) process with autocorrelation $\rho = 0.5$ and frailties u_i are generated from gamma distribution with $\text{E}(u_i) = 1$ and $\text{Var}(u_i) = \alpha$, and

$$\begin{aligned} f(\mathbf{x}_{ij}) = {} & 0.4\cos(x_{1ij}) + 0.3\cos(x_{2ij}) + 0.6\cos(x_{3ij}) + 0.5x_{2ij} \cdot x_{3ij} \\ & + 0.4/(x_{4ij}^2 + 1) + 0.5/(x_{5ij}^2 + 1). \end{aligned}$$

We set the frailty variance α to be 0, 0.5, 1 and 2, where $\alpha = 0$ means the DNN-Cox model without frailty. The censoring times are generated from an exponential distribution with parameter empirically determined to achieve

approximately two right censoring rates, low (around 15%) and high (around 45%). We set the total sample size $N = 8000$ with $(n, n_i) = (1000, 8)$ for all i. Thus, the dataset contains 1000 subjects and each subject has 8 observations. For each subject i, we assign 4 observations ($j = 1, 2, 3, 4$) to the training set, 2 observations ($j = 5, 6$) to the validation set and the remaining 2 observations ($j = 7, 8$) to the test set.

For comparison, we consider the fitting of the following four models.

- **Cox:** Cox proportional hazard model
- **DNN-Cox:** DNN-Cox proportional hazard model
- **FM:** Gamma frailty model
- **DNN-FM:** The proposed DNN-frailty model

The network architecture and hyper-parameters are tuned by using the vanilla DNN-Cox. As an optimal result, we set all the DNN models to have 3 hidden layers of 10 nodes with relu activation function. We use the full batch and AdamW optimizer with learning rate 0.01. Early stopping with validation loss is employed to prevent overfitting. The Cox model is implemented using `lifelines` package in Python, gamma FM is implemented using `frailtyEM` (Balan and Putter, 2019) package in `R`, and the DNN models (DNN-Cox, DNN-FM) are implemented using Python based on Keras (Chollet et al., 2015) and Tensorflow (Abadi et al., 2015).

7.5.2 Experimental results

For evaluation of the prediction performances, IBS (7.12) and C-index (7.13) in Appendix 7.8.2 are computed on the test set. Figure 4 shows box plots of IBS for each model under 15% censoring. Figure 7.3 (a) shows that all models have comparable results when there is no frailty. Even if there is no frailty ($\alpha = 0$), the proposed DNN-FM model is still comparable to the vanilla DNN-Cox, which should have the smallest IBC. Figure 7.3 (b), (c) and (d) show that the DNN-FM has the smallest IBS values when frailty is presented. Figure 7.4 shows box plots of C-index for each model under 15% censoring. When there is no frailty term, Figure 7.4 (a) shows that the two DNN models (DNN-Cox, DNN-FM) have comparable results, but that the two non-DNN models (Cox, FM) give poor results, which means they do not capture the nonlinear effect of input variables in terms of C-index. As expected, Figure 7.4 (b), (c) and (d) show that the DNN-FM has the highest C-index. Next, Figures 7.5 and 7.6 present box plots of IBS and C-index under 45% censoring, respectively and they overall show similar trends to Figures 7.3 and 7.4. However, the trends in Figure 7.5 (a) are somewhat different. That is, the two standard models (Cox and FM) in Figure 7.5 (a) give poor results as compared to those in Figure 7.3 (a), meaning that under 45% censoring, they do not again capture the nonlinear effect of input variables in terms of IBC.

Mean and standard deviation of IBS and C-index for each model with two censoring rates are summarized in Table 7.1. This confirms that the DNN-FM outperforms three existing models (Cox, DNN-Cox and FM). Table 7.2 reports mean and standard deviation of estimated frailty variance ($\widehat{\alpha}$) from train sets under 100 replications of simulated data. When $\alpha = 0$, the true model does

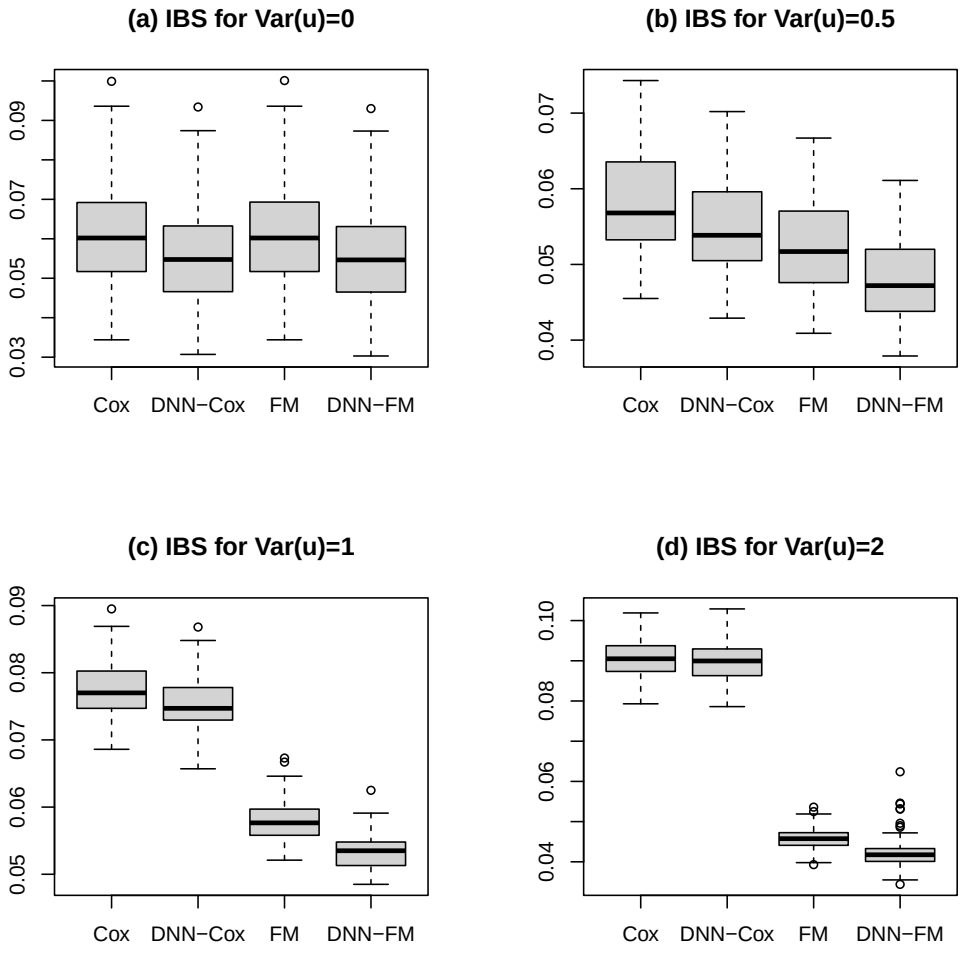


Figure 7.3: 15% censoring: Box plot of IBS from 100 replications for each frailty variance, $\text{var}(u) = \alpha$.

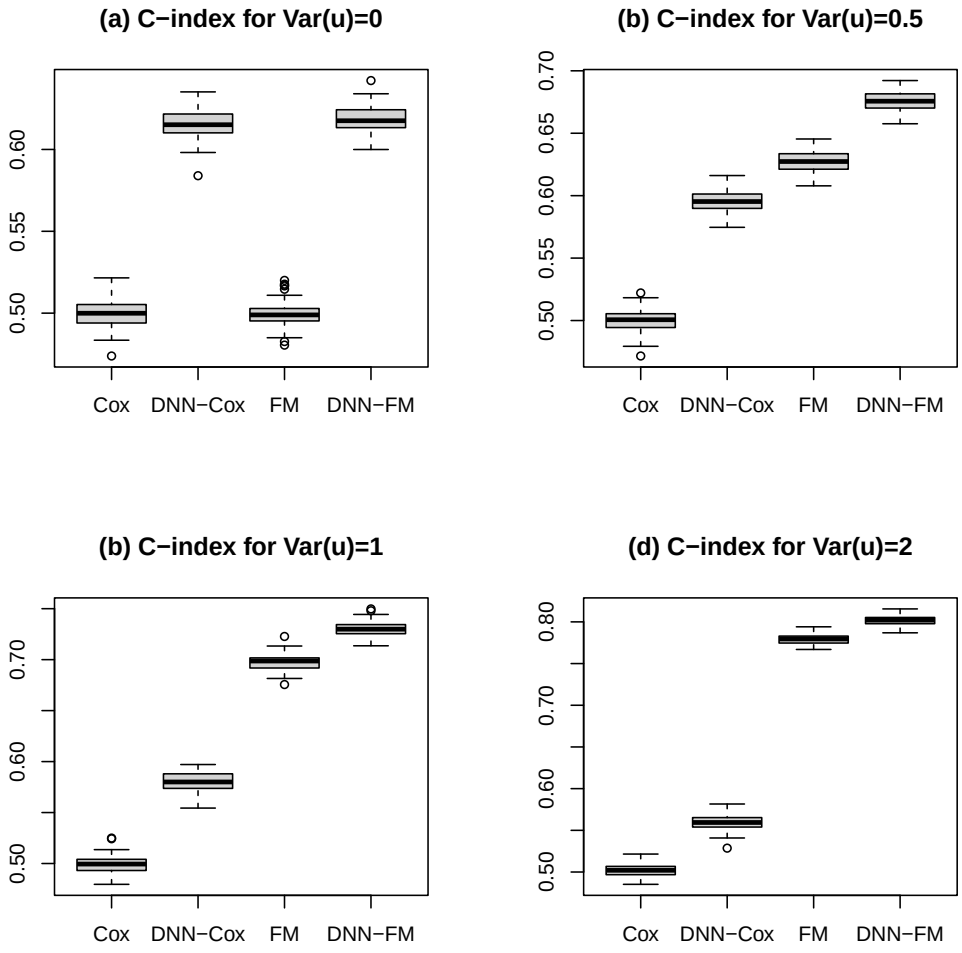


Figure 7.4: 15% censoring: Box plot of C-index from 100 replications for each frailty variance, $\text{var}(u) = \alpha$.

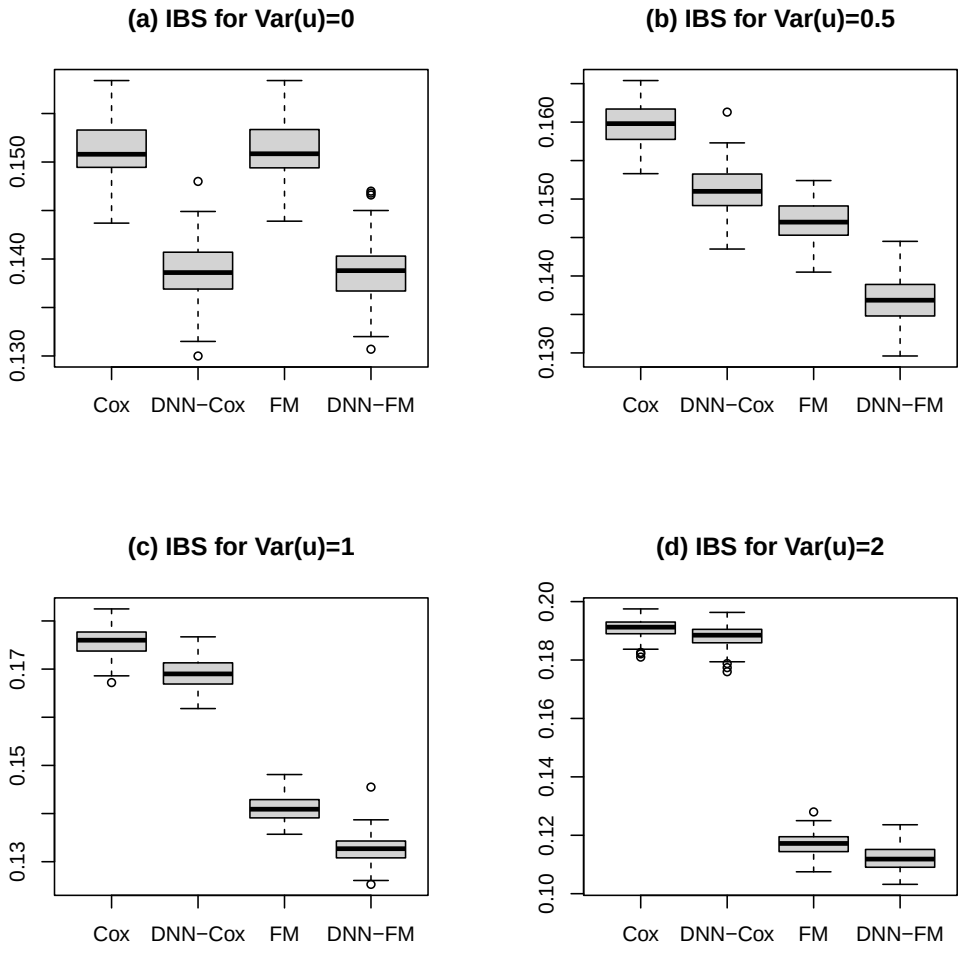


Figure 7.5: 45% censoring: Box plot of IBS from 100 replications for each frailty variance, $\text{var}(u) = \alpha$.

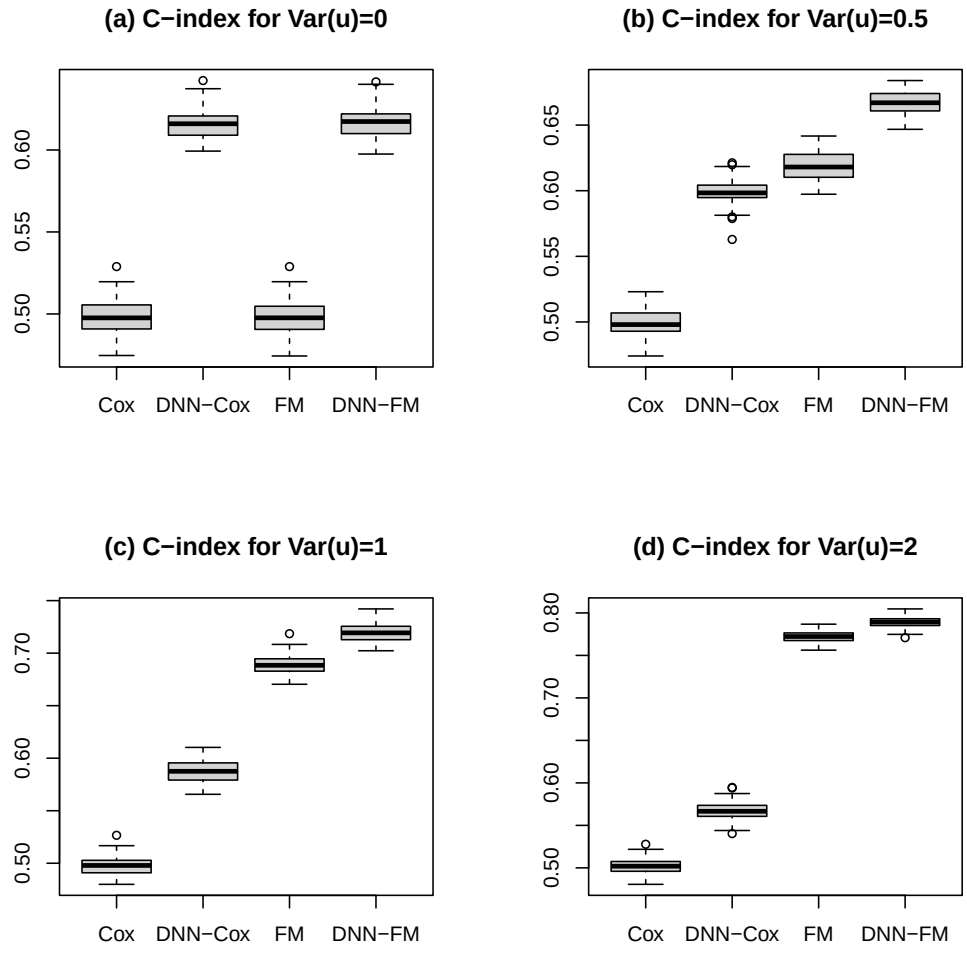


Figure 7.6: 45% censoring: Box plot of C-index from 100 replications for each frailty variance, $\text{var}(u) = \alpha$.

Table 7.1: Mean (standard deviation) of IBS and C-index from 100 replications for each frailty variance α.

Censoring	Measure	α	Cox	DNN-Cox	FM	DNN-FM
15%	IBS	0	0.062 (0.013)	0.056 (0.012)	0.062 (0.013)	0.056 (0.012)
		0.5	0.058 (0.007)	0.055 (0.006)	0.053 (0.006)	0.048 (0.006)
		1	0.077 (0.004)	0.075 (0.004)	0.058 (0.003)	0.053 (0.003)
		2	0.090 (0.005)	0.090 (0.005)	0.046 (0.003)	0.042 (0.004)
	C-index	0	0.499 (0.009)	0.615 (0.008)	0.499 (0.007)	0.618 (0.008)
		0.5	0.500 (0.008)	0.596 (0.009)	0.627 (0.008)	0.675 (0.008)
		1	0.499 (0.008)	0.580 (0.009)	0.697 (0.008)	0.730 (0.007)
		2	0.502 (0.007)	0.559 (0.009)	0.779 (0.006)	0.802 (0.005)
45%	IBS	0	0.151 (0.003)	0.139 (0.003)	0.151 (0.003)	0.139 (0.003)
		0.5	0.160 (0.003)	0.151 (0.003)	0.147 (0.003)	0.137 (0.003)
		1	0.176 (0.003)	0.169 (0.003)	0.141 (0.003)	0.133 (0.003)
		2	0.191 (0.003)	0.188 (0.004)	0.117 (0.004)	0.112 (0.004)
	C-index	0	0.498 (0.011)	0.616 (0.009)	0.498 (0.011)	0.617 (0.009)
		0.5	0.500 (0.011)	0.599 (0.010)	0.619 (0.011)	0.667 (0.009)
		1	0.498 (0.009)	0.587 (0.011)	0.689 (0.009)	0.720 (0.008)
		2	0.502 (0.009)	0.567 (0.010)	0.772 (0.007)	0.789 (0.006)

Table 7.2: Mean and standard deviation of estimated frailty variance $\widehat{\alpha}$ from 100 replications.

Censoring	α	FM	DNN-FM
15%	0	0.006 (0.009)	0.008 (0.010)
	0.5	0.390 (0.035)	0.485 (0.050)
	1	0.823 (0.051)	1.000 (0.062)
	2	1.711 (0.084)	2.043 (0.094)
45%	0	0.008 (0.012)	0.011 (0.014)
	0.5	0.417 (0.047)	0.496 (0.054)
	1	0.859 (0.060)	0.995 (0.090)
	2	1.748 (0.102)	2.065 (0.123)

not have frailties, and the estimates of α under FM and DNN-FM with two censoring rates (15% and 45%) are closed to zero. As α increases, the MLE of α under FM is downward biased, whereas that under DNN-FM is consistent. As expected, we see that the standard deviations of $\widehat{\alpha}$ tend to increase as α or censoring rate increases.

7.6 Multi-center bladder cancer data

We illustrate the DNN-FM method using a bladder cancer multi-center trial conducted by the EORTC (Sylvester et al., 2006). We consider 392 bladder-cancer patients from 21 centers that participated in EORTC trial 30791. The primary endpoint (event of interest) was time (day) to the first bladder cancer recurrence from randomization. Of the 392 patients, 200 (51.02%) had recurrence of bladder cancer (event of interest) and 81 (20.66%) died prior to recurrence (a competing event). 111 (28.32%) patients who were still alive and without recurrence were censored at the date of the last available follow-up.

Following Park and Ha (2019), we regarded the 81 competing risk events as censored, resulting that censoring rate is 49.98% with 192 censored patients. The data are unbalanced due to different number of patients in each center. In this chapter, we used the data with 373 patients from 16 centers which have more than 5 patients in each center. The numbers of patients per center varied from 6 to 78, with mean 23.3 and median 17.5. In each center, we used two randomly selected patients as test set, another two randomly selected patients as validation set, and the remaining patients as training set. We consider the following 12 categorical input variables (**x**):

- Chemotherapy (main covariate): yes, no
- Age: less than or equal to 65, greater than 65
- Gender: male, female
- Prior recurrent rate: primary, less than 1/yr, greater than 1/yr
- Number of tumors: single tumor, 2-7 tumors, more than 7 tumors
- Tumor size: less than 3cm, greater than or equal to 3cm
- T category: Ta=0, T1=1
- Carcinoma in situ: yes, no
- G grade: G1, G2, G3

Table 7.3 presents IBS and C-index on the test set of the bladder cancer data. The DNN-FM shows the smallest IBS and the highest C-index which indicate the best prediction performance, and DNN-Cox outperforms the two non-DNN

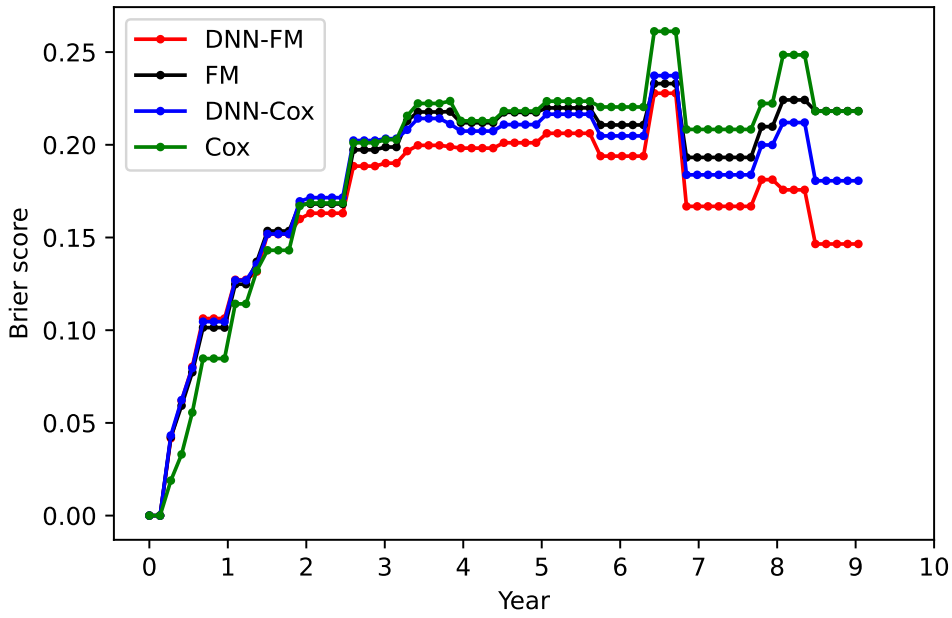


Figure 7.7: Time-dependent Brier score for four survival prediction models on the test set of the bladder cancer data.

models (Cox and FM). In the train set, the estimated frailty variances are small with $\widehat{\alpha} = 0.069$ for DNN-FM and $\widehat{\alpha} = 0.086$ for FM, leading to similar values of IBS and C-index from the Cox and FM. Thus, in this dataset the nonlinear effect of input variables are important in predicting the survival probability of patients in each center. Figure 7.7 shows the time-dependent Brier scores on the test set under the four models. Here, the Brier scores of the four models are similar at almost time points before 3 years. However, after 3 years, the Brier scores of the proposed DNN-FM are always noticeably lower than other three models. Accordingly, the DNN-FM improves prediction of the DNN-Cox model.

Table 7.3: IBS and C-index for four survival prediction models on the test set of the bladder cancer data

Measure	Cox	DNN-Cox	FM	DNN-FM
IBS	0.189	0.183	0.187	0.168
C-index	0.675	0.682	0.668	0.693

7.7 Concluding remarks

We have presented a new DNN-FM. The joint maximization of its profiled h-likelihood provides MLEs for fixed parameters and BUPs for random frailties. Our empirical results demonstrate that the proposed method improves the prediction performance of the existing DNN-Cox and FMs in terms of IBS and C-index. The specification of the gamma frailty distribution in semi-parametric FMs is insensitive to the estimates of fixed regression parameters if the variance of frailty is not very large (Gorfine and Zucker, 2023; Ha et al., 2017, 2001; Hsu et al., 2007). Extension of the proposed method to other frailty distribution such as parametric (e.g. log-normal) or non-parametric distribution (Chee et al., 2021) would be an interesting further work. The proposed DNN-FM can be trained for very large clustered survival data by using an online learning, whose theoretical framework is in Appendix 7.8.3.

7.8 Appendix

7.8.1 Derivation for the predictive likelihood

Recall that $\mathbf{y}^* = (\mathbf{y}, \boldsymbol{\delta})$, where the (i,j)th component of $\mathbf{y}$ is $y_{ij} = \min(T_{ij}, C_{ij})$. Note that $\tilde{\mathbf{v}}$ is given by

$$\begin{aligned}
\tilde{\mathbf{v}} &= \arg\max_{\mathbf{v}} \{\log f_{\boldsymbol{\psi}}(\mathbf{v}|\mathbf{y}^*)\} \\
&= \arg\max_{\mathbf{v}} \{\log f_{\boldsymbol{\psi}}(\mathbf{y}^*|\mathbf{v}) + \log f_{\boldsymbol{\psi}}(\mathbf{v}) - \log f_{\boldsymbol{\psi}}(\mathbf{y}^*)\} \\
&= \arg\max_{\mathbf{v}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(\delta_{ij} v_i - \Lambda_{ij}^{(m)} e^{v_i} \right) + \sum_{i=1}^{n} \left(\frac{v_i - e^{v_i}}{\alpha} \right) \right\} \\
&= \arg\max_{\mathbf{v}} \sum_{i=1}^{n} \left\{ v_i \left(\delta_{i+} + \alpha^{-1} \right) - e^{v_i} \left(\Lambda_{i+} + \alpha^{-1} \right) \right\} \\
&= \log \left(\frac{\delta_{i+} + \alpha^{-1}}{\Lambda_{i+} + \alpha^{-1}} \right)
\end{aligned}$$

where $\delta_{i+} = \sum_{j=1}^{n_i} \delta_{ij}$, $\Lambda_{i+} = \sum_{j=1}^{n_i} \Lambda_{ij}^{(m)} = \sum_{j=1}^{n_i} \Lambda_0(y_{ij}) \exp(f(\mathbf{x}_{ij}))$. This implies that

$$\tilde{u}_i = \exp(\tilde{v}_i) = \frac{\delta_{i+} + \alpha^{-1}}{\Lambda_{i+} + \alpha^{-1}} = E(u_i|\mathbf{y}_i^*)$$

is the BUP (Searle et al., 1992) for $u_i (= \exp(v_i))$ in sense that it gives minimum mean squared error of prediction (best) and $E(\tilde{u}_i - u_i) = 0$ (unbiased) with $E(\tilde{u}_i) = E(u_i) = 1$, since

$$u_i|\mathbf{y}_i^* \sim \text{Gamma}\left(\delta_{i+} + \alpha^{-1}, (\Lambda_{i+} + \alpha^{-1})^{-1} \right),$$

which is easily derived from the fact that the gamma distribution is conjugate of the frailty model. From the density function of gamma distribution above,

the predictive likelihood at $\tilde{\mathbf{v}}^c$ is given by

$$\begin{aligned}\log f(\tilde{\mathbf{v}}^c|\mathbf{y}^*) &= \sum_{i=1}^{n} \log f_{\boldsymbol{\psi}}(\tilde{v}_i^c|\mathbf{y}^*) = \sum_{i=1}^{n} \{\log f_{\boldsymbol{\psi}}(\tilde{v}_i|\mathbf{y}^*) - a_i(\alpha, \delta_{i+})\} \\ &= \sum_{i=1}^{n} \{\log f_{\boldsymbol{\psi}}(\tilde{u}_i|\mathbf{y}^*) + \log \tilde{u}_i - a_i(\alpha, \delta_{i+})\} = 0,\end{aligned}$$

where $\tilde{u}_i = \exp(\tilde{v}_i)$.

7.8.2 Evaluation measures for DNN-FM

The Brier score can be extended to the DNN-FM as a conditional form:

$$BS_c(t) = E\left\{Y(t) - S(t|u,x)\right\}^2,$$

where $S(t|u,x)$ is the conditional survival function given u, and the estimated conditional BS (Van Oirbeek and Emmanuel, 2016) is given by

$$\widehat{BS}_c(t) = \frac{1}{N} \sum_{ij \in D_N} \widehat{w}_{ij}(t) \left\{ y_{ij}(t) - \widehat{S}(t|\widehat{u}_i, x_{ij}) \right\}^2, \tag{7.12}$$

where $N = \sum_{i=1}^{N}$ is the total sample size and the IPCW is

$$\widehat{w}_{ij}(t) = \frac{(1 - y_{ij}(t))\delta_{ij}}{\widehat{G}(y_{ij})} + \frac{y_{ij}(t)}{\widehat{G}(t)}, \text{ with } \widehat{G}(t) = \widehat{P}(C > t).$$

The BS can be also summarized as the integrated Brier score (IBS).

The C-index can be also extended to the DNN-FM with clustered survival data (Van Oirbeek and Lesaffre, 2010). For the clustered data, we consider the overall conditional C-index, i.e., the concordant probability defined for all

comparable pairs; it can distinguish two different types of pairs, within-cluster pairs and between-cluster pairs, i.e. pairs whose members belong to the same cluster or to different clusters, respectively. Thus, the overall C-index (C_O) can be split up into a between-cluster C-index (C_B) and a within-cluster C-index (C_W). Let $i = 1, \ldots, n$ define the cluster and let ij be the subset j of the cluster i ($j = 1, \ldots, n_i$). We also denote by ij and ij' two patients from the same cluster i and by ij and $i'j'$ two patients from two different clusters i and i' ($i \neq i'$). For simplicity, we consider no ties, even if it can handle similarly to the case in Section 7.2.2 with presence of ties. Then the estimated within-cluster C-index ($\widehat{C}_W$) is given by

$$\widehat{C}_W = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \delta_{ij} I(y_{ij} < y_{ij'}) I\left(\widehat{\eta}_{ij}^{(m)} > \widehat{\eta}_{ij'}^{(m)}\right)}{\sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \{\delta_{ij} I(y_{ij} < y_{ij'})\}} \right],$$

where $\widehat{\eta}_{ij}^{(m)} = \text{NN}(\mathbf{x}_{ij}; \widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\beta}})$ and the frailty terms are not included directly in the calculation of the within-cluster concordance since they are the same for the compared patients in each pair. Next, the estimated between-cluster C-index ($\widehat{C}_B$) considers only comparison between patients of different clusters and includes the estimated frailty terms; it is given by

$$\widehat{C}_B = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\sum_{i'=1}^{n} \sum_{j'=1}^{n'_i} \delta_{ij} I(y_{ij} < y_{i'j'}) I\left(\widehat{\eta}_{ij} > \widehat{\eta}_{i'j'}\right) \right]}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\sum_{i'=1}^{n} \sum_{j'=1}^{n'_i} \{\delta_{ij} I(y_{ij} < y_{i'j'})\} \right]},$$

where $\widehat{\eta}_{ij} = \widehat{\eta}_{ij}^{(m)} + \widehat{v}_i = \text{NN}(\mathbf{x}_{ij}; \widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\beta}}) + \widehat{v}_i$ and $\widehat{v}_i = \log \widehat{u}_i$. Thus, the estimated overall C-index ($\widehat{C}_O$) can be expressed as a weighted mean of $\widehat{C}_B$ and $\widehat{C}_W$

(Van Oirbeek and Lesaffre, 2010), given by

$$\widehat{C}_O = \frac{n_{T,comp}}{n_{W,comp}}\widehat{C}_W + \frac{n_{T,comp}}{n_{B,comp}}\widehat{C}_B, \tag{7.13}$$

where $n_{T,comp}$ is the number of comparable pairs, and $n_{W,comp}$ and $n_{B,comp}$ are the number of comparable within-and between-cluster pairs, respectively. Note that $\widehat{C}_B$ can be easily calculated based on the function, concordance-index, with a Python library `lifelines`.

7.8.3 Online learning for the DNN-FM

Since the loss function of the DNN-Cox model does not naturally decouple, it causes computational difficulties in large data sets. To overcome this difficulty, Tarkhan and Simon (2022) proposed an online framework. In this section, we extend the online framework to DNN-FM by a simple modification (7.14) of h_p in (7.10).

Let D_s be a set of random samples of size $s_i \geq 0$ drawn from the population of each patient (or cluster) $i = 1, ..., n$, where $s_i (\leq n_i)$ are non-negative integers. Under the assumption of no ties and no censoring, define the profiled h-likelihood from the mini-batch D_s as

$$h_p^{(s)} = \sum_{i:s_i>0} \sum_{j=1}^{s_i} \left[\eta_{ij} - \log \left\{ \sum_{(k,l)\in R_{ij}^{(s)}} \exp(\eta_{kl}) \right\} + \frac{v_i - \exp(v_i)}{n_i \alpha} + c_i(\alpha, n_i) \right], \tag{7.14}$$

where $c_i(\alpha, n_i) = \{-\alpha^{-1} \log \alpha - \log \Gamma(\alpha^{-1}) - a_i(\alpha, n_i)\}/n_i$ and $R_{ij}^{(s)} = \{(k,l) : y_{kl} \geq y_{ij}$ and $(i,j,k,l) \in D_s\}$ is the risk set at the (i,j)th ordered failure

time y_{ij}. Note here that profiled h-likelihood (7.14) from the mini-batch gives $h_p^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = h_p(\boldsymbol{\theta}, \mathbf{v})$ when $D_s = D_n$. Let $U_{\boldsymbol{\beta}}^{(s)}(\boldsymbol{\theta}, \mathbf{v})$, $U_{\alpha}^{(s)}(\boldsymbol{\theta}, \mathbf{v})$, and $U_{\mathbf{v}}^{(s)}(\boldsymbol{\theta}, \mathbf{v})$ be the score functions of profiled h-likelihood from D_s with respect to $\boldsymbol{\beta}$, α, and $\mathbf{v}$, respectively,

$$U_{\boldsymbol{\beta}}^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = \frac{\partial h_p^{(s)}(\boldsymbol{\theta}, \mathbf{v})}{\partial \boldsymbol{\beta}}, \quad U_{\alpha}^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = \frac{\partial h_p^{(s)}(\boldsymbol{\theta}, \mathbf{v})}{\partial \alpha}, \quad U_{\mathbf{v}}^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = \frac{\partial h_p^{(s)}(\boldsymbol{\theta}, \mathbf{v})}{\partial \mathbf{v}}.$$

Then, if $s_i > 0$ for some i and $s_j = 0$ for all $j \neq i$, we have the following Theorem 7.1.

Theorem 7.1. *Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \alpha^*)$ be the vector of true values of fixed parameters and $\widetilde{\mathbf{v}}$ be the mode of profiled h-likelihood at $\theta = \theta^*$, then*

$$E\left[U_{\boldsymbol{\beta}}^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})\right] = 0 \quad \textit{and} \quad E\left[U_{\alpha}^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})\right] = 0.$$

Remark 1: Tarkhan and Simon (2022) studied the online learning framework for the DNN-Cox model. Theorem 7.1 extends the framework to the DNN-FM, with restriction that a mini-batch should be sampled within a cluster.

Theorem 7.2. *Let $\mathbf{v}^*$ be the vector of realized values of random parameters (i.e. log-frailties), then*

$$E\left[U_{\mathbf{v}}^{(s)}(\boldsymbol{\theta}^*, \mathbf{v}^*)\right] \to 0 \quad \textit{as } n_i \to \infty \text{ for all } i.$$

Remark 2: When the cluster size n_i approaches infinity for all i, Theorem 7.2 shows that the frailty predictors converge in probability to their true realized values. It implies that the frailty predictors approach the fixed effect estimators of $\mathbf{v}$ of the Cox model with fixed parameters $\mathbf{v}$. Therefore, the online learning

framework of Tarkhan and Simon (2022) can be directly used for DNN-FM when $n_i \to \infty$. In this case, mini-batches can be drawn from multiple clusters.

7.8.4 Proofs

Proof of Theorem 1

(a) Here, it is enough to consider

$$h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = \sum_{i:s_i>0} \sum_{j=1}^{s_i} \left[\eta_{ij} - \log \left\{ \sum_{(k,l)\in R_{ij}^{(s)}} \exp(\eta_{kl}) \right\} \right],$$

since $\boldsymbol{\beta}$ does not involve the remaining terms of profiled h-likelihood. Here, $\eta_{ij} = \eta_{ij}^{(m)} + v_i$ and $\eta_{ij}^{(m)} = \eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta}) = \text{NN}(\mathbf{x}_{ij}; \mathbf{w}, \boldsymbol{\beta})$. Analogous to Tarkhan and Simon (2022), we define a counting process $dN_{ij}(t)$ as

$$\int_a^b g(t) dN_{ij}(t) = g(t_{ij}) I(t_{ij} \in [a, b]),$$

and define $dN^{(s)}(t) = \sum_{i:s_i>0} \sum_{j=1}^{s_i} dN_{ij}(t)$ to be a counting process for failure times over all patients in D_s under the assumption that the failure time process is absolutely continuous with respect to Lebsegue measure on time, which implies that there is no ties at any time t. Then, $h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v})$ can be expressed as

$$h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = \sum_{i:s_i>0} \sum_{j=1}^{s_i} \eta_{ij} - \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau \log \left\{ \sum_{(k,l)\in R_{ij}^{(s)}} M_{kl}(t) \exp(\eta_{kl}) \right\} dN_{ij}(t),$$

where τ is the duration of the study, and its derivative is

$$\begin{aligned} U_\beta^{(s)}(\boldsymbol{\theta}, \mathbf{v}) &= \frac{\partial h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i:s_i>0} \sum_{j=1}^{s_i} \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}) - \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau \sum_{k,l} w_{kl}(\boldsymbol{\theta}, \mathbf{v}) \eta'(\mathbf{x}_{kl}; \boldsymbol{\beta}) dN_{ij}(t) \\ &= \sum_{i:s_i>0} \sum_{j=1}^{s_i} \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}) - \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau w_{ij}(\boldsymbol{\theta}, \mathbf{v}) \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}) dN^s(t), \end{aligned}$$

where $\eta'(\mathbf{x}_{ij}; \boldsymbol{\beta})$ is the gradient of $\eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta})$ with respect to β,

$$w_{ij}(\boldsymbol{\theta}, \mathbf{v}) = \frac{M_{ij}(t) \exp\{\eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta}) + v_i\}}{\sum_{k,l} M_{kl}(t) \exp\{\eta^{(m)}(\mathbf{x}_{kl}; \boldsymbol{\beta}) + v_i\}}$$

is a weight proportional to the hazard of failure and $M_{ij}(t)$ is an indicator representing whether ij is at risk at time t, i.e., $t_{ij} \geq t$. Thus, the score function $U_\beta^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})$ is given by

$$U_\beta^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}}) = \sum_{i:s_i>0} \sum_{j=1}^{s_i} \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) - \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau w_{ij}(\boldsymbol{\beta}^*, \widetilde{\mathbf{v}}) \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) dN^s(t),$$

and it is enough to show that $E(U_\beta^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})) = E(E(U_\beta^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})|\mathbf{v} = \mathbf{v}^*)) = 0$. As in Tarkhan and Simon (2022), we have

$$E\left(\sum_{i:s_i>0} \sum_{j=1}^{s_i} \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) \Big| \mathbf{v} = \mathbf{v}^* \right) = \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau E\left(w_{ij}(\boldsymbol{\beta}^*, \mathbf{v}^*) \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) dN^s(t) \right),$$

Then the score function becomes

$$
\begin{aligned}
&E(U_\beta^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})|\mathbf{v} = \mathbf{v}^*) \\
&= E\left(\sum_{i:s_i>0} \sum_{j=1}^{s_i} \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) - \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau w_{ij}(\boldsymbol{\beta}^*, \widetilde{\mathbf{v}}) \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) dN^s(t) \right) \\
&= \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau E\left[w_{ij}(\boldsymbol{\beta}^*, \mathbf{v}^*) \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) dN^s(t) \right] \\
&\quad - \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau E\left[w_{ij}(\boldsymbol{\beta}^*, \widetilde{\mathbf{v}}) \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) dN^s(t) | \mathbf{v} = \mathbf{v}^* \right] \\
&= \sum_{i:s_i>0} \sum_{j=1}^{s_i} \int_0^\tau \eta'(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) E\left[\{ w_{ij}(\boldsymbol{\beta}^*, \mathbf{v}^*) - w_{ij}(\boldsymbol{\beta}^*, \widetilde{\mathbf{v}}) \} dN^s(t) | \mathbf{v} = \mathbf{v}^* \right].
\end{aligned}
$$

If the mini-batch is sampled within the i-th cluster only,

$$
\begin{aligned}
&w_{ij}(\boldsymbol{\beta}^*, \mathbf{v}^*) - w_{ij}(\boldsymbol{\beta}^*, \widetilde{\mathbf{v}}) \\
&= \frac{M_{ij}(t) \exp\{\eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) + v_i^*\}}{\sum_{l=1}^{s_i} M_{il}(t) \exp\{\eta^{(m)}(\mathbf{x}_{il}; \boldsymbol{\beta}^*) + v_i^*\}} - \frac{M_{ij}(t) \exp\{\eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta}^*) + \widetilde{v}_i\}}{\sum_{l=1}^{s_i} M_{il}(t) \exp\{\eta^{(m)}(\mathbf{x}_{il}; \boldsymbol{\beta}^*) + \widetilde{v}_i\}} \\
&= \frac{M_{ij}(t) \exp\{\eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta}^*)\}}{\sum_{l=1}^{s_i} M_{il}(t) \exp\{\eta^{(m)}(\mathbf{x}_{il}; \boldsymbol{\beta}^*)\}} - \frac{M_{ij}(t) \exp\{\eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta}^*)\}}{\sum_{l=1}^{s_i} M_{il}(t) \exp\{\eta^{(m)}(\mathbf{x}_{il}; \boldsymbol{\beta}^*)\}} \\
&= 0,
\end{aligned}
$$

which leads to $E(U_\beta^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})) = 0$.

(b) The score function with respect to α is

$$
U_\alpha^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = \frac{\partial h_p^{(s)}(\boldsymbol{\theta}, \mathbf{v})}{\partial \alpha} = \frac{s_i}{n_i} \frac{\partial}{\partial \alpha} \left[\frac{\log u_i - u_i}{\alpha} - \frac{\log \alpha}{\alpha} - \log \Gamma\left(\alpha^{-1}\right) - a_i(\alpha, n_i) \right]
$$

where $a_i(\alpha, n_i) = (n_i + \alpha^{-1}) \{\log(n_i + \alpha^{-1}) - 1\} - \log\Gamma(n_i + \alpha^{-1})$. Since

$$u_i|\mathbf{y}_i^* \sim \text{Gamma}\left(\delta_{i+} + \alpha^{-1}, (\Lambda_{i+} + \alpha^{-1})^{-1}\right)$$

and $\widetilde{u}_i = \widetilde{u}_i(\alpha) = (n_i + \alpha^{-1})/(\Lambda_{i+} + \alpha^{-1}) = E(u_i|\mathbf{y}_i^*)$,

$$\begin{aligned}
&U_\alpha^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}}) = U_\alpha^{(s)}(\boldsymbol{\theta}, \mathbf{v})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*, \mathbf{v}=\widetilde{\mathbf{v}}} \\
&= \frac{s_i}{n_i}\frac{1}{\alpha^2}\left[(u_i - \log u_i - 1) + \log(\alpha n_i + 1) - \psi\left(n_i + \frac{1}{\alpha}\right) + \psi\left(\frac{1}{\alpha}\right)\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*, \mathbf{v}=\widetilde{\mathbf{v}}} \\
&= \frac{s_i}{n_i}\frac{1}{\alpha^{*2}}\left[(\widetilde{u}_i - \log\widetilde{u}_i - 1) + \log(\alpha^* n_i + 1) - \psi\left(n_i + \frac{1}{\alpha^*}\right) + \psi\left(\frac{1}{\alpha^*}\right)\right] \\
&= \frac{s_i}{n_i}\frac{1}{\alpha^{*2}}\left[\widetilde{u}_i - 1 + \log\left(\Lambda_{i+} + \frac{1}{\alpha^*}\right) - \log\left(\frac{1}{\alpha^*}\right) - \psi\left(n_i + \frac{1}{\alpha^*}\right) + \psi\left(\frac{1}{\alpha^*}\right)\right] \\
&= \frac{s_i}{n_i}\frac{1}{\alpha^{*2}}\left[\text{E}(u_i|\mathbf{y}^*) - 1 - \text{E}(\log u_i|\mathbf{y}^*) - \log\left(\frac{1}{\alpha^*}\right) + \psi\left(\frac{1}{\alpha^*}\right)\right]
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function. Thus, we have

$$\text{E}\left[U_\alpha^{(s)}(\boldsymbol{\theta}^*, \widetilde{\mathbf{v}})\right] = \frac{s_i}{n_i}\frac{1}{\alpha^{*2}}\left[\text{E}(u_i) - 1 - E(\log u_i) - \log\left(\frac{1}{\alpha^*}\right) + \psi\left(\frac{1}{\alpha^*}\right)\right] = 0$$

since $\text{E}(u_i) = 1$ and $\text{E}(\log u_i) = \psi(1/\alpha^*) + \log(\alpha^*)$.

Proof of Theorem 2

Let $\mathbf{z}_{ij}$ be the one-hot encoded vector of cluster number, so that $\mathbf{z}_{ij}^T\mathbf{v} = v_i$, then the predictor η_{ij} can be expressed as $\eta_{ij} = \eta^{(m)}(\mathbf{x}_{ij}; \boldsymbol{\beta}) + v_i = \eta^*(\mathbf{x}_{ij}, \mathbf{z}_{ij}; \boldsymbol{\beta}, \mathbf{v})$. For example,

$$\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i = (\mathbf{x}_{ij}^T, \mathbf{z}_{ij}^T)(\boldsymbol{\theta}, \mathbf{v}).$$

Define $h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v})$ as

$$h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = \sum_{i:s_i>0} \sum_{j=1}^{s_i} \left[\eta_{ij} - \log \left\{ \sum_{(k,l)\in R_{ij}^{(s)}} \exp(\eta_{kl}) \right\} \right],$$

then the profiled h-likelihood in (7.14) can be expressed as

$$h_p^{(s)}(\boldsymbol{\theta}, \mathbf{v}) = h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v}) + \sum_{i:s_i>0} \sum_{j=1}^{s_i} \left[\frac{v_i - \exp(v_i)}{n_i \alpha} + c_i(\alpha, n_i) \right].$$

Thus, $h_1^{(s)}(\boldsymbol{\theta}, \mathbf{v})$ is equivalent to the log-partial likelihood (Tarkhan and Simon, 2022) when $\mathbf{v}$ is treated as the fixed parameters, and the remaining terms does not depend on $\boldsymbol{\beta}$. Therefore, by the results of Tarkhan and Simon (2022),

$$E\left[U_{\boldsymbol{\beta}}^{(s)}(\boldsymbol{\theta}^*, \mathbf{v}^*)\right] = 0,$$

and

$$E\left[U_{\mathbf{v}}^{(s)}(\boldsymbol{\theta}^*, \mathbf{v}^*)\right] = E\left[\sum_{i:s_i>0} \sum_{j=1}^{s_i} \frac{\partial}{\partial \mathbf{v}} \left[\frac{v_i - \exp(v_i)}{n_i \alpha} + c_i(\alpha, n_i) \right] \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*, \mathbf{v}=\mathbf{v}^*}.$$

When $n_i \to \infty$,

$$\frac{\partial}{\partial v_i} \left[\frac{v_i - \exp(v_i)}{n_i \alpha^*} + c_i(\alpha^*, n_i) \right] = \frac{1}{n_i} \left[\frac{1 - \exp(v_i^*)}{\alpha^*} \right] \to 0.$$

Thus,

$$E\left[U_{\mathbf{v}}^{(s)}(\boldsymbol{\theta}^*, \mathbf{v}^*)\right] \to 0.$$

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems.

Arrow, K. J. and Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society*, pages 265–290.

Balan, T. A. and Putter, H. (2019). frailtyem: An r package for estimating semiparametric shared frailty models. *Journal of Statistical Software*, 90:1–29.

Balch, M. S. (2012). Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning*, 53(7):1003–1019.

Balch, M. S. (2020). New two-sided confidence intervals for binomial inference derived using walley's imprecise posterior likelihood as a test statistic. *International Journal of Approximate Reasoning*, 123:77–98.

Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A*, 475(2227):20180565.

Banta, J. A., Stevens, M. H., and Pigliucci, M. (2010). A comprehensive test of the 'limiting resources' framework applied to plant tolerance to apical meristem damage. *Oikos*, 119(2):359–369.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bayarri, M., DeGroot, M., and Kadane, J. (1988). What is the likelihood function? (with discussion). In Gupta, S. S. and Berger, J. O., editors, *Ststistical Decision Theory and Related Topics IV*. Springer.

Berger, J. O. and Wolpert, R. L. (1988). *The likelihood principle*. Institute of Mathematical Statistics.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):113–128.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91(434):791–806.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Breslow, N. E. (1972). Discussion of professor cox's paper. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34:216.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91.

Brooks, M., Bolker, B., Kristensen, K., Maechler, M., Magnusson, A., McGillycuddy, M., Skaug, H., Nielsen, A., Berg, C., Bentham, o. v., Sadat, N., Lüdecke, D., Lenth, R., O'Brien, J., Geyer, C. J., Jagan, M., Wiernik, B., and Stouffer, D. B. (2023). *glmmTMB: Generalized Linear Mixed Models using Template Model Builder*. R package version 3.2.0.

Buehler, R. J. (1959). Some validity criteria for statistical inferences. *The Annals of Mathematical Statistics*, 30(4):845–863.

Butler, R. W. (1986). Predictive likelihood inference with applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 48(1):1–23.

CDC (2020). Daily census tract-level pm2.5 concentrations.

Chee, C.-S., Do Ha, I., Seo, B., and Lee, Y. (2021). Semiparametric estimation for nonparametric frailty models using nonparametric maximum likelihood approach. *Statistical methods in medical research*, 30(11):2485–2502.

Chollet, F. et al. (2015). Keras.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 49(1):1–18.

Cunen, C., Hjort, N. L., and Schweder, T. (2020). Confidence in confidence distributions! *Proceedings of the Royal Society A*, 476(2237):20190781.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in bayesian and structural inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 35(2):189–213.

De Finetti, B. (1931). On the subjective meaning of probability. In *Induction and Probability*, pages 291–321. Springer.

Dempster, A. P. (1968). Upper and lower probabilities generated by a random closed interval. *The Annals of Mathematical Statistics*, pages 957–966.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*), 39(1):1–22.

Denoeux, T. and Li, S. (2018). Frequency-calibrated belief functions: Review and new insights. *International Journal of Approximate Reasoning*, 92:232–254.

Edwards, A. W. F. (1977). Discussion of mr. wilkinson's paper. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(2):144–145.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80(1):3–26.

Efron, B. (1998). Ra fisher in the 21st century. *Statistical Science*, pages 95–114.

Evans, M. (2013). What does the proof of birnbaum' s theorem prove?

Firth, D. (2006). Contribution to the discussion of" double hierarchical generalized linear models" by y. lee and ja nelder. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 55(2):168–170.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368.

Fisher, R. A. (1930). Inverse probability. In *Mathematical proceedings of the Cambridge philosophical society*, volume 26(4), pages 528–535. Cambridge University Press.

Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 139(838):343–348.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 144(852):285–307.

Fisher, R. A. (1973). *Statistical methods and scientific inference*. Hafner Press.

Ghosh, M., Reid, N., and Fraser, D. (2010). Ancillary statistics: A review. *Statistica Sinica*, pages 1309–1332.

Gilmour, A., Anderson, R. D., and Rae, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, 72(3):593–599.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gorfine, M. and Zucker, D. M. (2023). Shared frailty methods for complex survival data: A review of recent advances. *Annual Review of Statistics and Its Application*, 10:51–73.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Gu, M. G., Sun, L., and Huang, C. (2004). A universal procedure for parametric frailty models. *Journal of statistical computation and simulation*, 74(1):1–13.

Guo, C. and Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.

Ha, I. D., Jeong, J.-H., and Lee, Y. (2017). *Statistical modelling of survival data with random effects*. Springer.

Ha, I. D. and Lee, Y. (2003). Estimating frailty models via poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*, 12(3):663–681.

Ha, I. D., Lee, Y., and Song, J.-k. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, 88(1):233–243.

Han, J. and Lee, Y. (2022). Enhanced laplace approximation. *arXiv preprint arXiv:2207.09871*.

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, pages 491–544.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Harville, D. A. and Mee, R. W. (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, pages 393–408.

Hejduk, M. and Snow, D. (2019). Satellite conjunction "probability", "possibility", and "plausibility": A categorization of competing conjunction assessment risk assessment paradigms. In *AAS/AIAA Astrodynamics Specialist Conference*.

Hejduk, M. D., Snow, D., and Newman, L. (2019). Satellite conjunction assessment risk analysis for "dilution region" events: issues and operational approaches. In *Space Traffic Management Conference*.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, 1973(Symposium):10–41.

Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192–218.

Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, 90(2):355–366.

Henry, K., Erice, A., Tierney, C., Balfour Jr, H., Fischl, M., Kmack, A., Liou, S., Kenton, A., Hirsch, M., Phair, J., et al. (1998). A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced aids. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 19:339–349.

Hinkley, D. V. (1977). Conditional inference about a normal mean with known coefficient of variation. *Biometrika*, 64(1):105–108.

Hsu, L., Gorfine, M., and Malone, K. (2007). Effect of frailty distribution misspecification on marginal regression estimates and hazard functions in multivariate survival analysis. *Statistics in Medicine*, 26:4657–4678.

Jeon, J., Hsu, L., and Gorfine, M. (2012). Bias correction in the hierarchical likelihood approach to the analysis of multivariate survival data. *Biostatistics*, 13(3):384–397.

Kaminsky, K. S. and Rhodin, L. S. (1985). Maximum likelihood prediction. *The Annals of the Institute of Statistical Mathematics*, 37:507–517.

Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20:1–30.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, H., Ha, I. D., Hwang, C., and Lee, Y. (2023a). Deep neural networks for clustered count data via hierarchical likelihood. *manuscript prepared for publication*.

Lee, H., Ha, I. D., and Lee, Y. (2023b). Deep neural networks for semiparametric frailty models via h-likelihood. *manuscript prepared for publication*.

Lee, H. and Lee, Y. (2023). H-likelihood approach to deep neural networks with temporal-spatial random effects for high-cardinality categorical features. In *International Conference on Machine Learning*. PMLR.

Lee, Y. and Bjørnstad, J. F. (2013). Extended likelihood approach to large-scale multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(3):553–575.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(4):619–656.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006.

Lee, Y. and Nelder, J. A. (2002). Analysis of ulcer data using hierarchical generalized linear models. *Statistics in Medicine*, 21(2):191–202.

Lee, Y. and Nelder, J. A. (2005). Likelihood for random-effect models. *Sort*, 29(2):141–164.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized linear models with random effects: unified analysis via H-likelihood*. Chapman and Hall/CRC, 1. edition.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2017). *Generalized linear models with random effects: unified analysis via H-likelihood*. Chapman and Hall/CRC, 2. edition.

Lewis, D. (1980). A subjectivist' s guide to objective chance. In *IFS: Conditionals, Belief, Decision, Chance and Time*, pages 267–297. Springer.

Lindley, D. V. (1958). Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pages 102–107.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Mandel, F., Ghosh, R. P., and Barnett, I. (2021). Neural networks for clustered and longitudinal data using mixed effects models. *Biometrics*.

Martin, R., Balch, M. S., and Ferson, S. (2021). Response to the comment confidence in confidence distributions! *Proceedings of the Royal Society A*, 477(2250):20200579.

Martin, R. and Liu, C. (2015). *Inferential models: reasoning with uncertainty*, volume 145. CRC Press.

McCulloch, C. E. and Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, 67(1):270–279.

Meng, X.-L. (2009). Decoding the h-likelihood. *Statistical Science*, 24(3):280–293.

MuonNeutrino (2019). Us census demographic data.

Park, E. and Ha, I. D. (2019). Penalized variable selection for accelerated failure time models with random effects. *Statistics in Medicine*, 38(5):878–892.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Pawitan, Y., Lee, H., and Lee, Y. (2023). Epistemic confidence in the observed confidence interval. *Scandinavian Journal of Statistics*.

Pawitan, Y. and Lee, Y. (2017). Wallet game: Probability, likelihood, and extended likelihood. *The American Statistician*, 71(2):120–122.

Pawitan, Y. and Lee, Y. (2021). Confidence as likelihood. *Statistical Science*, 36(4):509–517.

Pedersen, J. (1978). Fiducial inference. *International Statistical Review/Revue Internationale de Statistique*, pages 147–170.

Ramsey, F. P. (1926). Truth and probability. In *Readings in Formal Epistemology: Sourcebook*, pages 21–45. Springer.

Reese, A. (2020). Used cars dataset - vehicles listings from craigslist.org.

Reid, N. (1995). The roles of conditioning in inference. *Statistical Science*, 10(2):138–157.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022.

Robinson, G. (1979). Conditional properties of statistical procedures. *The Annals of Statistics*, pages 742–755.

Rodrigo, H. and Tsokos, C. (2020). Bayesian modelling of nonlinear poisson regression with artificial neural networks. *Journal of Applied Statistics*, 47(5):757–774.

Ross, S. A. et al. (1976). Return, risk and arbitrage. In *Risk and Return in Finance*, pages 189–218. Rodney L. White Center for Financial Research, The Wharton School, University of Pennyslvania,.

Roulin, A. and Bersier, L.-F. (2007). Nestling barn owls beg more intensely in the presence of their mother than in the presence of their father. *Animal Behaviour*, 74(4):1099–1106.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Schweder, T. (2018). Confidence is epistemic probability for empirical science. *Journal of statistical Planning and Inference*, 195:116–125.

Schweder, T. and Hjort, N. L. (2016). *Confidence, likelihood, probability*, volume 41. Cambridge University Press.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance components*. John Wiley & Sons.

Shafer, G. and Vovk, V. (2019). *Game-theoretic foundations for probability and finance*, volume 455. John Wiley & Sons.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4):749–760.

Simchoni, G. and Rosset, S. (2021). Using random effects to account for high-cardinality categorical features and repeated measures in deep neural networks. *Advances in Neural Information Processing Systems*, 34:25111–25122.

Simchoni, G. and Rosset, S. (2023). Integrating random effects in deep neural networks. *Journal of Machine Learning Research*, 24(156):1–57.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *The Annals of Mathematical Statistics*, 30(4):877–880.

Sun, T., Wei, Y., Chen, W., and Ding, Y. (2020). Genome-wide association study-based deep learning for survival prediction. *Statistics in Medicine*, 39(30):4605–4620.

Sylvester, R. J., Van Der Meijden, A. P., Oosterlinck, W., Witjes, J. A., Bouffioux, C., Denis, L., Newling, D. W., and Kurth, K. (2006). Predicting recurrence and progression in individual patients with stage ta t1 bladder cancer using eortc risk tables: a combined analysis of 2596 patients from seven eortc trials. *European Urology*, 49(3):466–477.

Tarkhan, A. and Simon, N. (2022). An online framework for survival analysis: reframing cox proportional hazards model for large data sets and neural networks. *Biostatistics*, pages 1–20.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, pages 657–671.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics*, 12(1):156–175.

Tran, M.-N., Nguyen, N., Nott, D., and Kohn, R. (2020). Bayesian deep net GLM and GLMM. *Journal of Computational and Graphical Statistics*, 29(1):97–113.

Van Oirbeek, R. and Emmanuel, L. (2016). Exploring the clustering effect of the frailty survival model by means of the brier score. *Communications in Statistics*, 45:3294–3306.

Van Oirbeek, R. and Lesaffre, E. (2010). An application of harrell's c-index to ph frailty models. *Statistics in Medicine*, 29(30):3160–3171.

Van Trees, H. L. (1968). *Detection, Estimation and Modulation Theory*, volume 793. John Wiley & Sons.

Wilkinson, G. N. (1977). On resolving the controversy in statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(2):119–144.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016a). Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR.

Wilson, A. G., Hu, Z., Salakhutdinov, R. R., and Xing, E. P. (2016b). Stochas-

tic variational deep kernel learning. *Advances in Neural Information Processing Systems*, 29.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, 80(4):791–795.

국문초록

관측할 수 없는 변량 효과에 관한 통계적 추론

이 논문은 관측할 수 없는 변량 효과에 대한 통계적 추론에 관련된 여섯 개의 주제로 구성되어 있으며, 각각의 주제는 관측할 수 없는 변량에 관한 정보를 갖고 있는 확장된 가능도(extended likelihood)를 중심으로 연결되어 있다. 전반부의 두 주제는 신뢰도(confidence) 이론에 관한 내용으로, 신뢰분포의 밀도함수(confidence density)를 확장된 가능도로 해석하여, 신뢰도에 관한 이론적 성질을 규명하였다. 후반부의 네 주제는 특수한 스케일에서의 확장된 가능도로 정의되는 계층적 가능도(hierarchical likelihood)에 관한 이론적 성질과 딥러닝으로의 확장 및 응용에 관한 내용을 다룬다.

첫번째 주제에서는 관측된 신뢰구간에 대해 인식론적 신뢰도(epistemic confidence)를 정의하고 이를 계산하기 위한 방법을 제시하였다. 또한, 빈도주의적 관점에서 정의되는 신뢰도가 갖는 relevant subset 문제를 Ramsey-De Finetti의 Dutch book 논의에 betting market의 존재를 도입함으로써 설명하고, 인식론적 신뢰도가 이러한 문제로부터 자유로울 수 있음을 보였다. 두번째 주제에서는 Stein의 역설과 인공위성 충돌 문제를 통해 신뢰분포가 특정 지점에서 point mass를 갖는 문제를 새로운 관점에서 해석하여, 역설적으로 여겨지던 point mass의 존재가 신뢰분포의 핵심적인 성질을 유지하도록 만들어주는 데 중요한 역할을 한다는 것을 밝혔다. 이와 더불어, 제안한 형태의 신뢰 분포가 확률 형태의 추론이 갖는 근본적인 한계점으로 지적된 거짓 신뢰(false confidence) 문제에서 (적어도 목표 가설에 한해) 자유롭다는 것을 밝히고, 기존의 다른 방법론들과 달리 Stein 문제 및 인공위성 충돌 문제에서 적정 신뢰도를 유지할 수 있음을 보였다.

세번째 주제에서는 계층 가능도를 새롭게 정의하고 이론적 성질을 규명하였다. 새로운 계층 가능도는 고정 효과의 최대 가능도 추정량과 변량 효과의 점근적 최소분산 불편추정량을 제공할 수 있으며, 기존의 계층 가능도가 갖고 있던 이론적 모호성을 해소할 수 있다. 마지막 세 주제는 새롭게 정의한 계층 가능도를 기반으로 한 딥러닝 모형을 다루고 있다. 대부분의 딥러닝 모형들이 데이터의 독립성을 암묵적으로 가정하고 있지만, 실제 데이터는 시공간적 상관관계를 갖는 경우가 많다. 딥러닝 모형에 변량 효과를 도입함으로써 이러한 문제를 해결할 수 있으며, 계층 가능도 기반 접근법은 기존의 방법론에 비해 여러가지 장점을 갖는다. 네번째 주제에서는 시공간 상관관계를 갖는 연속형 데이터를 다루기 위한 딥러닝 모형을 다루었고, 다섯번째 주제에서는 비정규 변량 효과를 갖는 가산형 데이터를 다루기 위한 딥러닝 모형을 다루었다. 여섯번째 주제에서는 군집화된 절단자료를 분석하기 위한 계층 가능도 기반 준모수적(semi-parameteric) 접근법에 대해 다루었다. 세 주제 모두, 제안한 방법을 통해 딥러닝 모형에 변량 효과를 도입함으로써 기존 방법론의 예측 성능을 향상시킬 수 있었다.

주요어: 변량 효과, 계층 가능도, 딥러닝, 신뢰도, 인식론적 신뢰도, 신뢰분포, 반복측정자료, 시공간 자료, 생존자료분석, 프레일티 모형

학번: 2015-20310