Likelihood prediction for generalized linear mixed models under covariate uncertainty

Md Moudud Alam[*]

Sep. 22, 2010.

Abstract

This paper presents the techniques of likelihood prediction for the generalized linear mixed models. Methods of likelihood prediction is explained through a series of examples; from a classical one to more complicated ones. The examples show, in simple cases, that the likelihood prediction (LP) coincides with already known best frequentist practice such as the best linear unbiased predictor. The paper outlines a way to deal with the covariate uncertainty while producing predictive inference. Using a Poisson error-in-variable generalized linear model, it has been shown that in complicated cases LP produces better results than already know methods.

Key words: Predictive likelihood, Profile predictive likelihood, Stochastic covariate, Coverage interval, Future value prediction, Credit risk prediction.

1 Introduction

Predictive inference is a tricky task, especially for non-Bayesian statisticians (Bjørnstad, 1990 and Hinkley, 1979). The core of the problem was understood during the foundational period of statistics (see e.g. Pearson 1920) but it took a long time for the non-Bayesian statisticians to come up with a set of reasonable proposals on the predictive tools with Lauritzen (1974) and Hinkley (1979) being credited as the earliest, theoretically most sound, references. Unless otherwise stated, by prediction we mean the prediction of one or more unobserved (observable or not) variables or some function of them after having observed the observable variables. Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ be the vector observations on the response, Y, $\mathbf{X}_{n\times p}$ be the matrix of associated observed covariates, $\mathbf{y}^* = (y_1^*, y_2^*, \cdots, y_m^*)$ be the future observations on Y which are not observed and $\dot{\mathbf{X}}^*_{m\times p}$ be the associated covariate matrix where some of its elements are known and some are unknown.

Note that we use an asteric ("*") in the superscript (e.g. X^*) to indicate that the whole variable, vector or matrix or a part of it is not observed, but they are observable. The covariates and the design matrices associated y^* are denoted with an over head accent- dot (".", e.g. $\dot{X}$). As per convention, we use upper case letters to indicate variables, lower cases to indicate their realized values and bold faces to indicate vectors and matrices. We use the subscripts, i $(i = 1, 2, ..., n)$ to indicate observed data, j $(j = 1, 2, ..., m)$ to indicate future observations and their sum is $n + m = l$.

The unknown elements in $\dot{\mathbf{X}}^*$ are not necessarily missing values in the ordinary sense, e.g. non-response in a survey as in Bjørnstad (1996) and Bjørnstad and Sommervoll (2001), rather

[*]School of Technology and Business Studies, Dalarna University and Swedish Business School, Örebro University. Contact: Dalarna University, SE 781 88 Borlänge, Sweden; maa@du.se .

they might be some future values which can be observed only in future time while the prediction is made at current time.

We further assume that, given $\mathbf{X}_{n\times p}$ and information on the clustering of Y, the response can be modeled with a suitable generalized linear mixed model (GLMM). The unknown future covariates can also be modelled with a suitable stochastic model. The problem of interest is to predict Y^* itself or some function $S = s\left(Y^*\right)$ and provide a measure of uncertainty of those predictions based on observed data on Y and $\mathbf{X}$. Some illustrations of the above problem with known $\dot{\mathbf{X}}$ are given in Lee et al. (2006).

Natural examples of stochastic covariates with generalized linear models come from the time series models (Slud and Kedem, 1994; Startz, 2008), dynamic panel discrete choice models (Honoré and Kyriazidou, 2000) and measurement error models (Buzas and Stefanski, 1996). Here we motivate the application of unknown future covariates from the credit risk modeling's view point. Assume that Y represents whether a credit is default or not and $\mathbf{X}$ consists of the respective firm level accounting data, industry classification of the firm, credit bureau observation (comments) and macro variables e.g. slope of yield curve, output gap etc. (see e.g. Carling et al. (2004) and Duffie et al. (2007)). Some of the covariates, e.g. firm's total debt, sales and macro economic indices, are stochastic and their future values can not be observed at current time when the prediction is being made. Assume that we model Y given $\mathbf{X}$ using a suitable GLMM and the unobserved components of $\dot{\mathbf{X}}^*$ with missing future values are modelled with a suitable time-series model. Then, the remaining problem is to predict Y^* (or S) and to provide a measure of uncertainty associated with the prediction.

In the literature of credit risk modelling, the issue of stochastic covariates is handled by the so called doubly-stochastic models using the framework of survival analysis (Duffie et al., 2007; Pesaran et al., 2006). However, those works do not give proper attention to the uncertainties caused by the stochastic covariates nor do they distinguish the problem of estimation from the problem of prediction. Thus the predictive methods presented in this paper may also be applied to those early works with a view to improve the predictive performances of their models.

Given a prediction problem in hand, one can either try to find a frequentist point prediction, e.g. the best linear unbiased predictor (BLUP), and associated prediction error or try to produce a likelihood prediction (Bjørnstad, 1990) or follow the Bayesian approach. The first approach does not have a common analytical framework and the existence of the BLUP is not guaranteed, in general. The Bayesian approach is, in principle, rather straightforward although the choice of a particular prior as well as the concept of the prior distribution may be criticized. The likelihood principle (Berger and Wolpert, 1988) provides a unified principle and an analytical framework to deal with any statistical inference including the prediction of future and unobserved values. This paper explores the likelihood prediction in the context of GLMM.

The contributions of this paper are as follows. It offers a short overview of the likelihood prediction through a series of standard prediction problems. The examples show that the likelihood prediction can be implemented in a straightforward way and its solutions often coincide with already known best frequentist prediction, where such a best prediction exists. The paper also gives the likelihood prediction in more complicated problems such as error in variable generalized linear models and GLMM where a best frequentist prediction such as BLUP is not available. Through an example with a Poisson error-in-variable model it is shown, through simulation, that the likelihood prediction does a better job than the already existing solutions. The paper also outlines an analytical guideline to implement the likelihood prediction with GLMM under covariate uncertainty.

The rest of the paper is organized as follows. Section 2 briefly introduces the principles of likelihood prediction through two classical examples. Section 3 extends the likelihood prediction for GLMM with covariate uncertainty. Section 4 presents three examples of the likelihood prediction under covariate uncertainties. Section 5 offers a comparative discussion on the several proper predictive likelihoods that have been proposed in the literature. Section 6 concludes.

2 Likelihood prediction

An elegant survey on the methods of likelihood prediction is given in Bjørnstad (1990). Often, the prediction statement is summarized in terms of probability inequality which is called the prediction interval. A review of the different methods of producing non-Bayesian prediction interval is presented in Patel (1989). To illustrate the likelihood prediction we take a classic example (see Example 1) that was presented in Pearson (1920), with a reference to Laplace (1774) as the originator, and also discussed by many others including Hinkley (1979), Bjørnstad (1990) and Pawitan (2001).

Example 1. An event has occurred p times out of $p+q=n$ trials, where we have no *apriori* knowledge of the frequency of the events in the total population of occurrences. What is the probability of its occurring r times in a further $r+s=m$ trials?

Example 1 can be translated in terms of the notation system given in Section 1 as: $Y = (Y_1, Y_2, \cdots, Y_n)$ are iid Bernoulli distributed with $E(Y_i) = \theta$, $Y^* = (Y_1^*, ..., Y_m^*)$, $Y_i \perp Y_j \ \forall i \& j$, $\sum_{i=1}^{n} y_i = p$, $S = \sum_{j=1}^{m} y_j^* = r$ and the interest is to predict r given p, n and m. Example 1 qualifies as a fundamental statistical problem which was solved in Laplace (1774) with some difficulty (see Pearson, 1920; Stigler, 1986) using the Bayesian approach. The Bayesian solution to the problem is straightforward and with a flat prior for θ the posterior predictive distribution of r is given as (see Bjørnstad(1990))

$$p(r|p,n) = \frac{\binom{m}{r}\binom{n}{p}}{\binom{m+n}{r+p}} \frac{n+1}{n+m+1}, r = 0, 1, ..., m \tag{1}$$

Due to the unavailability of the concept of prior distribution, a non-Bayesian solution is not easy to formulate. If θ were known, the distribution of r would be Binomial with mean $m\theta$. Hence a non-Bayesian mean predictor of r would be $E(r|\theta, m) = m\theta$. Thus, a naive prediction (NP) of r is given as $\widetilde{r} = m\frac{p}{n}$ where θ is replaced by its maximum likelihood (ML) estimate obtained from the observed data. Though, $\widehat{\theta} = \frac{p}{n}$ is the maximum likelihood estimator of θ, $\widetilde{r}$ is not a maximum likelihood predictor. In fact, classical likelihood theory does not allow its application as a predictive criterion (Hinkley, 1979). A likelihoodist sees the above problem as the one dealing with two unknowns, θ and r where r is of inferential interest and θ is considered as a nuisance parameter. The above line of thinking leads the likelihoodists to construct a joint likelihood function (Bjørnstad, 1990) of θ and r as

$$\begin{aligned} L(r, \theta|p, m, n) &= L(r|m, n, p, \theta) L(\theta|p, m, n) \\ &= \binom{m}{r}\binom{n}{p} \theta^{p+r} (1-\theta)^{n+m-p-r} \end{aligned}$$

Alhough $L(r, \theta|p, m, n)$ is justified as a likelihood for prediction, the likelihood principle does not state clearly what one should do with θ and how the information about r contained in $L(r, \theta|p, m, n)$ should be extracted (Berger and Wolpert, 1989). At this point the likelihoodists introduce the method of profile likelihood (Pawitan, 2001) which essentially maximizes the likelihood with respect to a subset of parameters treating the remaining parameters as constants (known). For Example 1, we have the following profile likelihood.

$$L_p\left(r|p,m,n\right) = \underset{\theta}{\sup} L\left(r,\theta|p,m,n\right)$$

$$\Rightarrow L_p\left(r|p,m,n\right) \propto \binom{m}{r}\binom{n}{p}\left(p+r\right)^{p+r}\left(m+n-p-r\right)^{m+n-p-r}$$

The likelihoodists treat L_p differently from the formal (or estimative) likelihood in the sense that L_p is often normalized to mimic a Bayesian posterior density for r. Such a normalization is justified since r, unlike the fixed parameters θ, has a probability distribution. Using Stirling's approximation to $L_p\left(r|p,m,n\right)$ it can be shown that

$$L_p\left(r|p,m,n\right) \propto \frac{p\left(r|p,n\right)}{\sqrt{\widehat{\theta}^*\left(1-\widehat{\theta}^*\right)}} \tag{2}$$

where, $p\left(r|p,n\right)$ is the Bayesian posterior predictive density of r under a flat prior and $\widehat{\theta}^* = \frac{p+r}{m+n}$ is obtained from maximizing $L\left(r,\theta|p,m,n\right)$ w.r.t. θ. A critical drawback of $L_p\left(r|p,m,n\right)$ is that it replaces the nuisance parameter with its MLE thereby introducing an additional uncertainty in the predictive distribution which in turn calls for some adjustment. We also see that a multiplicative adjustment term of $\sqrt{\widehat{\theta}^*\left(1-\widehat{\theta}^*\right)}$ makes $L_p^{(1)}\left(r|p,m,n\right) = p\left(r|p,n\right)$ where $L_p^{(1)} = \sqrt{\widehat{\theta}^*\left(1-\widehat{\theta}^*\right)}L_p$ is the profile adjusted predictive likelihood.

Further note that the adjustment term has the form $\sqrt{\widehat{\theta}^*\left(1-\widehat{\theta}^*\right)} \propto \mathcal{I}^{-1/2}_{\theta=\widehat{\theta}^*}$ where $\mathcal{I}_{\theta=\widehat{\theta}^*}$ is the observed Fisher's information of θ obtained from $log\left(L\left(r,\theta|p,m,n\right)\right)$ i.e. $\mathcal{I} = -\frac{\partial^2 log\left(L\left(r,\theta|p,m,n\right)\right)}{\partial\theta^2}$. In matter of fact, the adjustment can always make $L_p^{(1)}\left(z|y\right) \propto p\left(z|y\right)$ up to an order $O\left(n^{-1}\right)$ (Davison, 1986). Thus we treat $L_p^{(1)}$ as equivalent to the Bayesian posterior prediction (PP) with a flat prior.

The equivalence of the predictive likelihood and the posterior predictive density with flat prior is easy to understand. The Bayesian posterior with flat prior is mathematically equivalent to the (estimative) likelihood function and herefore if there exist any predictive likelihood, then the latter should be equivalent to the posterior predictive distribution with a flat prior.

Predictive statistics for Example 1 and Example 2, below, are given in Table 1. For $m=1$, $L_p^{(1)}$ (or PP) gives $E\left(r|p,m=1,n\right) = P\left(r=1|p,m=1,n\right) = \frac{p+1}{n+2}$ which is different from the NP which is $\widetilde{r} = \frac{p}{n}$ (see Table 1). Thus the difference between NP and PP matters in cases with small n and extreme observed p.

Example 1 is a nice example of statistical prediction with independently and identically distributed (iid) variables. Next we illustrate the problem for a situation with non-identical distribution by using an example of a linear regression model.

Example 2: Assume a regression model, $y_i = \alpha + \beta x_i + \varepsilon_i$ where, $\varepsilon_i \overset{iid}{\sim} N\left(0,\sigma^2\right)$ with σ being known. We observe the paired sequence $\{y_i, x_i\}$, also x_j are known but we do not observe y_j^*. The problem here is to predict those unobserved y_j^*'s which are observable in the future.

In Example 2, we have observed data, $\mathbf{y} = \{y_i\}$ and $X = \{x_i\}$, unobserved future values $\mathbf{y}^* = (y_1^*, y_2^*, ..., y_m^*)^T$, known future covariates, $\dot{X} = \{x_j\}$ and nuisance parameters $\theta = (\alpha, \beta)$. A naive prediction of y_j^* is given as $\widetilde{y}_j^* = \widehat{\alpha} + \widehat{\beta} x_j$ where, $\widehat{\alpha}$ and $\widehat{\beta}$ are the ordinary least square estimates (also the MLEs in this case) of α and β obtained from the observed data. A naive variance estimator for y_j^* is given as $Var\left(\widetilde{y}_j^*\right) = Var\left(\widehat{\alpha}\right) + x_j^2 Var\left(\widehat{\beta}\right) + 2x_j Cov\left(\widehat{\alpha}, \widehat{\beta}\right)$ which does not account for the uncertainty in y_j^* itself. A reasonable measure of uncertainty in $\widetilde{y}_j^*$ is easily computed in this case and is given by $Var\left(\widetilde{y}_j^*\right) = \sigma^2 \left(1 + \dot{\mathbf{x}}_j \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \dot{\mathbf{x}}_j^T\right)$ where $\dot{\mathbf{x}}_j$ is the j^{th} row of the design matrix, $\dot{\mathbf{X}}$. $\widetilde{y}_j^*$ is known as the best (having minimum mean squared prediction error) linear unbiased predictor (BLUP).

In cases where σ is unknown it is replaced by its unbiased estimate, i.e. $\widetilde{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i\right)^2$. The profile adjusted predictive likelihood for Example 2 is given as

$$L_P^{(1)}\left(Y^*|y, \sigma\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}_F^* - \mathbf{X}_F\widehat{\boldsymbol{\beta}}^*\right)^T \left(\mathbf{y}_F^* - \mathbf{X}_F\widehat{\boldsymbol{\beta}}^*\right)\right] |\sigma^2 \left(\mathbf{X}_F^T\mathbf{X}_F\right)^{-1}|^{-\frac{1}{2}}$$

$$\Rightarrow L_P^{(1)}\left(Y^*|y, \sigma\right) \propto \exp\left[-\frac{1}{2}\left(\mathbf{y}^* - \dot{\mathbf{X}}\widehat{\boldsymbol{\beta}}\right)^T \left(\sigma^2\left(\mathbf{I} + \dot{\mathbf{X}}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\dot{\mathbf{X}}^T\right)\right)^{-1} \left(\mathbf{y}^* - \dot{\mathbf{X}}\widehat{\boldsymbol{\beta}}\right)\right] \qquad (3)$$

where, $\mathbf{y}_F^{*T} = \left(\mathbf{y}^T, \mathbf{y}^{*T}\right)$ is the full response vector, $\widehat{\boldsymbol{\beta}}^*$ is the MLE of $\boldsymbol{\beta} = (\alpha, \beta)^T$ based on the full data, $\mathbf{X}_F^T = \left(\mathbf{X}^T, \dot{\mathbf{X}}^T\right)$ and $\widehat{\boldsymbol{\beta}} = \left(\widehat{\alpha}, \widehat{\beta}\right)^T$ is the MLE based on the observed data. The detailed mathematical derivation of (3) is given by Eaton and Sudderth (1998).

The predictive likelihood in (3) is the kernel of a multivariate normal distribution, i.e. $L_P^{(1)}\left(\mathbf{y}^*|\mathbf{y}, \sigma\right) \sim N_{(N-n)}\left(\dot{\mathbf{X}}\widehat{\boldsymbol{\beta}}, \sigma^2\mathbf{V}\right)$ where, $\mathbf{V} = \mathbf{I} + \dot{\mathbf{X}}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\dot{\mathbf{X}}^T$. Hence, in this example the naive prediction coincides with the mean of the predictive likelihood[1]. The predictive statistics for Example 2 are presented in Table 1.

Table 1 Predictive statistics for Examples 1 and 2 according to NP, LP and BLUP

Example	Methods	Point Predictor	Predictive Variance	Predictive distribution
Example 1	NP	$E\left(r\right) = m\frac{p}{n}$	$m\frac{p}{n}\left(1 - \frac{p}{n}\right)$	$Binomial(m, \frac{p}{n})$
	$L_P^{(1)}$	$E\left(r\right) = \frac{m(p+1)}{n+2}$	$\frac{m(p+1)(n-p+1)}{(n+2)^2(n+3)}$	$\frac{\binom{m}{r}\binom{n}{p}}{\binom{m+n}{r+p}}\frac{n+1}{n+m+1}, r = 0, 1, ..., m$
	BLUP	$E\left(r\right) = m\frac{p}{n}$	$m\frac{p}{n}\left(1 - \frac{p}{n}\right)\frac{m+n-1}{n}$	NA
Example 2	NP	$E\left(Y^*\right) = \dot{\mathbf{X}}\widehat{\beta}$	$\sigma^2\dot{\mathbf{X}}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\dot{\mathbf{X}}^T$	$N\left(\dot{\mathbf{X}}\widehat{\beta}, \sigma^2\dot{\mathbf{X}}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\dot{\mathbf{X}}^T\right)$
	$L_P^{(1)}$	$E\left(Y^*\right) = \dot{\mathbf{X}}\widehat{\beta}$	$\sigma^2\mathbf{V}$	$N\left(\dot{\mathbf{X}}\widehat{\beta}, \sigma^2\mathbf{V}\right)$
	BLUP	$E\left(Y^*\right) = \dot{\mathbf{X}}\widehat{\beta}$	$\sigma^2\mathbf{V}$	$N\left(\dot{\mathbf{X}}\widehat{\beta}, \sigma^2\mathbf{V}\right)$

Note: $\mathbf{V}= \mathbf{I}+\dot{\mathbf{X}}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\dot{\mathbf{X}}^T$

Example 2 is still a simple one. To introduce more difficult situations, we next present prediction with generalized generalized linear mixed models.

[1] If σ^2 is unknown, the above mathematical derivation becomes very tedious therefore, we skip the latter case. Interested readers are referred to Bjørnstad (1990) for further results .

3 Prediction with GLMM

For observed Y and $\mathbf{X}$, a generalized linear mixed model can be presented through the following five assumptions: i) $Y = \{y_{ikt}\}$; $i = 1, 2, ..., n_{kt}$; $k = 1, 2, ..., K$; $t = 1, 2, ..., T$; is observed independently at a given value of the covariate $\mathbf{X} = \{\mathbf{x}_{ikt}\}$, and a given realization of the random effect u_{kt}, ii) $\mathbf{x}_{ikt}$ and u_{kt} influence the distribution of y_{ikt} via a linear function $\eta_{ikt} = \mathbf{x}_{ikt}\boldsymbol{\beta} + u_{kt}$ which is called the linear predictor, iii) conditional on u_{kt}, $\mu_{ki} = E(y_{ki}|u_{ki})$ satisfies $g(\mu) = \eta$ for some function g which is called a link function, iv) conditional on $\mathbf{u}_t = (u_{t1}, u_{t2}, ..., u_{tK})^T$, the distribution of y_{ikt} belongs to the exponential family of distributions and v) $\mathbf{u}_t$ follows a marginal distribution, $h(\mathbf{u})$. Often, $\mathbf{u}_t$ is assumed to have an independent multivariate normal distribution *i.e.* $\mathbf{u}_t \backsim \mathbf{N}_K(\mathbf{0}, \mathbf{D})$.

An example of GLMM can be given from a spatial data example, e.g. the analysis of Pittsburgh air particulate matter (PM) data (Lee et. al, 2006; section 8.6.3). Assume that y_{ikt} represent the i^{th} measure (replication) on PM at the k^{th} site on the t^{th} day. The covariate matrix, $\mathbf{X}$, includes seasonal indicators and the measures on daily weather conditions. The random effects, $\mathbf{u}_t$, represent random site effect which can be explained as the daily random cite-specific fluctuations, where a non-diagonal $\mathbf{D}$ implies that the observations from the different sites are correlated. Unlike the analysis in Lee et al. (2006), the aim in this application is to predict future PM, $Y^* = \left\{y^*_{kt'}\right\}$, or some function of it, $S = s(Y^*)$ where $t' > T$ but the number of sites (K) is fixed. Further assume that the design matrix, associated with $\mathbf{y}^*$, can be partitioned as $\dot{\mathbf{X}}^* = \left(\dot{\mathbf{X}}_C|\dot{\mathbf{X}}^*_S\right)$ where $\dot{\mathbf{X}}_C$ is currently known, e.g. the seasonal indicator, and $\dot{\mathbf{X}}^*_S$ is currently unknown, e.g. the precipitation, wind speed etc., and can only be observed in the future.

The above prediction problem fits well under the framework of unobservable variables, nuisance variable and parameters' likelihood presented in Berger and Wolpert (1988; sections 3.5.2 and 3.5.3). In this case $\xi = \left(Y^*, \dot{X}^*_S\right)$ is the unobserved variable, with Y^* being of interest, the random effects $\mathbf{u}$ is the nuisance variable and any parameter in the distributions of Y, ξ and $\mathbf{u}$ is a nuisance parameter. For further derivation of the predictive criteria we can use the "nuisance variables likelihood principles" (Berger and Wolpert, 1988).

3.1 Derivation of the predictive likelihood for GLMM

In general, with GLMM, we have observed data, $X = (X_C, X_S)$ where X_C consists of non-stochastic and X_S consists of stochastic covriates and $\mathbf{y} = \{y_{ikt}\}$ $(i = 1, 2, ..., n_{kt}; k = 1, 2, ..., K;\ t = 1, 2, ..., T)$, future covariates $\dot{X}^* = \left\{\dot{X}_{jkt',C}, \dot{X}^*_{jkt',S}\right\}$ $(t' \in (1, 2, ..., \max(t', T)))$ of which $\dot{X}_{jkt',C}$ is currently known, future response, $\mathbf{y}^* = \left\{y^*_{jkt'}\right\}$, which we want to predict and $\mathbf{u}_t$ and $\mathbf{u}_{t'}$ are the random effects which are independently distributed as $N(0, \mathbf{D})$ where $\mathbf{D}$ is an unknown but fixed positive definitive matrix. Denote $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, vech(\mathbf{D}))$ and the parameter vector in the distribution of $\dot{X}^*_S$ as $\boldsymbol{\kappa} = (\kappa_1, ..., \kappa_F)$. Assuming no overlap between $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$, i.e. $\boldsymbol{\theta} \cap \boldsymbol{\kappa} = \varnothing$, the joint likelihood function for this case is given by

$$\begin{aligned} L(\boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\kappa}, \xi|\mathbf{y}, X, X_C) &= f\left(\mathbf{y}, \boldsymbol{\xi}, \mathbf{u}, \dot{X}^*_S|\boldsymbol{\theta}, \boldsymbol{\kappa}, X, \dot{X}_C\right) & (4) \\ &= f\left(\mathbf{y}, \mathbf{y}^*|X, \dot{X}^*, \mathbf{u}, \boldsymbol{\theta}\right) f(X_S, X^*_S|\boldsymbol{\kappa}) f(\mathbf{u}|\mathbf{D}) & (5) \end{aligned}$$

The principle of marginal likelihood (Berger and Wolpert, 1988) says that any nuisance variable should be integrated out from the likelihood at the first hand. Without loss of generality we can denote the clusters (k dimension) in observed data with 1 to K and any cluster appears in the predictive space but not in observed data with $(k+1)$, $(k+2)$ and so on up to K' and do the same for t which goes up to T'. Therefore, the joint likelihood of $\boldsymbol{\theta}, \xi, \boldsymbol{\kappa}$ is given by

$$L\left(\boldsymbol{\theta}, \xi, \boldsymbol{\kappa}|Y\right) = \int \cdots \int_{-\infty}^{\infty} \prod_{t=1}^{\max(T,T')} f\left(\mathbf{y}, \mathbf{y}^*|X, \dot{X}^*, \mathbf{u}, \boldsymbol{\theta}\right) f\left(X_S, \dot{X}_S^*|X, \dot{X}_C^*, \boldsymbol{\kappa}\right) f\left(\mathbf{u}_t|\mathbf{D}\right) d\left(\mathbf{u}_t\right) \tag{6}$$

The integration in (6) is generally analytically intractable even for the observed data likelihood (Lee et al. 2006). For the GLMM, equation (6) can be presented in matrix notations as

$$L\left(\boldsymbol{\theta}, \xi, \boldsymbol{\kappa}|Y\right) = \left(\int \cdots \int_{-\infty}^{\infty} \prod_{t=1}^{\max(T,T^*)} \exp\left[\frac{\mathbf{y}_{F,t}^{*T}\zeta_t - \mathbf{1}^T b\left(\zeta_t\right)}{\phi} + \mathbf{1}^T c\left(\mathbf{y}_{F,t}^*, \phi\right)\right] f\left(\mathbf{u}_t\right) d\left(\mathbf{u}_t\right)\right) L\left(\dot{X}_S^*, \boldsymbol{\kappa}\right) \tag{7}$$

where $\mathbf{y}_{F,t}^{*T}{=}\left(\mathbf{y}_t^T, \mathbf{y}_t^{*T}\right)$ is the vector of observed and unobserved responses, $\zeta = \{\zeta_{ikt}\}$ is the vector of canonical parameters such that with canonical link $\zeta_t = \eta_t = \mathbf{X}_{F,t}\boldsymbol{\beta} + \mathbf{Z}_t\mathbf{u}_t$ where $\mathbf{X}_F = \left(\mathbf{X}_t^T, \dot{\mathbf{X}}_t^{*T}\right)^T$ is the design matrix associated with $\boldsymbol{\beta}$ for the data set at t (quarter) and $\mathbf{Z}_t$ is the design matrix associated with $\mathbf{u}_t = \left(u_{1t}, u_{2t}, ..., u_{K't}\right)^T$, $b\left(\right)$ is called the cumulant function and it is a function in "S" convention i.e. $b\left(\zeta_1, \zeta_2\right) = \left(b\left(\zeta_1\right), b\left(\zeta_2\right)\right)$, ϕ is the dispersion parameter of the conditional mean model and $L\left(\dot{X}_S^*, \boldsymbol{\kappa}\right) = f\left(X_S, \dot{X}_S^*|X_C, \dot{X}_C, \boldsymbol{\kappa}\right)$. For binomial and Poisson GLMM, $\phi = 1$.

Applying Laplace approximation to (7) the joint likelihood is simplified , after ignoring terms having zero expectation (see Breslow and Clayton, 1993; section 2.1), as

$$L\left(\boldsymbol{\theta}, \xi, \boldsymbol{\kappa}|\mathbf{y}\right) \approx |\mathbf{I}{+}\mathbf{D}^{-1}\mathbf{Z}\mathbf{W}\mathbf{Z}^T|^{-\frac{1}{2}} \exp\left[-\frac{\mathbf{y}_F^{*T}\zeta - \mathbf{1}^T b\left(\zeta\right)}{\phi} - \frac{1}{2}tr\left(\mathbf{D}^{-1}\mathbf{u}^T\mathbf{u}\right) - \mathbf{1}^T c\left(\mathbf{y}_F^*, \phi\right)\right]_{|\mathbf{u}=\widetilde{\mathbf{u}}} L\left(\dot{X}_S^*, \boldsymbol{\kappa}\right) \tag{8}$$

where $\mathbf{W}$ is the GLM weight matrix (McCullagh and Nelder, 1989) and $\widetilde{\mathbf{u}} = \{\mathbf{u}_t\}_{T'\times K'}$ is the maxima of the integrand function in (7) w.r.t. $\mathbf{u}$. For the detailed derivation of (8) readers are referred to Breslow and Clayton (1993) and Wand (2002).

The remaining task is to eliminate the nuisance parameter $\boldsymbol{\theta}$, $\boldsymbol{\kappa}$ and $\dot{X}_S^*$ from the model. Since $\dot{X}_S^*$ has probability distribution it can either be integrated out or profiled out while $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$ can only be profiled out. Since, adjusted profile likelihood is same as integrating the nuisance parameter out using Laplace approximation, we can profile out $\boldsymbol{\omega} = \left(\boldsymbol{\theta}, \boldsymbol{\kappa}, \dot{X}_S^*\right)$ altogether from (8). Thus we obtain the profile adjusted predictive likelihood for Z as

$$L_P^{(1)}\left(\mathbf{y}^*|\mathbf{y}, \mathbf{X}, \dot{\mathbf{X}}_C\right) = L_P\left(\mathbf{y}^*|\mathbf{y}, \mathbf{X}, \dot{\mathbf{X}}_C\right) |\mathcal{I}^*\left(\widehat{\boldsymbol{\omega}}^*\right)|^{-1/2} \tag{9}$$

where, $L_P\left(\mathbf{y}^*|\mathbf{y}, \mathbf{X}, \dot{\mathbf{X}}_C\right) = \sup_{\boldsymbol{\omega}}\left\{l\left(\boldsymbol{\omega}, \mathbf{y}^*|\mathbf{y}, \mathbf{X}, \dot{\mathbf{X}}_C\right)\right\}$ with $l = \log L\left(\boldsymbol{\omega}, \mathbf{y}^*|\mathbf{y}, \mathbf{X}, \dot{\mathbf{X}}_C\right)$ and $\mathcal{I}^*\left(\widehat{\boldsymbol{\omega}}^*\right) = \{\mathcal{I}_{vw}^*\left(\widehat{\boldsymbol{\omega}}^*\right)\}$ with $\mathcal{I}_{vw}^*\left(\widehat{\boldsymbol{\omega}}^*\right) = -\frac{\partial^2 l}{\partial\boldsymbol{\omega}_v\partial\boldsymbol{\omega}w}_{|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}^*}$ is the observed information matrix for $\boldsymbol{\omega}$ with fixed $\mathbf{y}^*$. Although equation (9) looks simple, its exact analytical derivation may be challenging, depending on (8). After $L_P^{(1)}$ has been computed one can predict Z from (9) in the following two ways (Bjørnstad, 1996)

a) mean prediction: normalize $L_P^{(1)}\left(y^*|y\right)$ to make it a pdf (pmf) and predict $\hat{y}^* = E_P\left(y^*|y\right)$.

Also base any statistical inference on the normalized $L_P^{(1)}(y^*|y)$ and

b) ML prediction: predict $\widehat{y}^*$ that maximizes $L_P^{(1)}(y^*|y)$, for continuous Y, and treat $L_P^{(1)}(y^*|y)$ as a likelihood function to make inference on y^*.

Bjørnstad (1996) prefers the mean prediction over the ML prediction considering the shortcomings of ML for the correlated data e.g. $\widehat{\boldsymbol{\theta}}^*$ and y^* are not, in general, invariant under one–to–one parameter transformation. Since $L_P^{(1)}$ is the approximate Bayesian posterior predictive density with flat prior, we may use the available Bayesian MCMC procedures (Gelman et al., 2004) to facilitate the computation of $\hat{\mathbf{y}}^*$ as the posterior mode or the posterior mean.

The prediction problem and its approximate likelihood solution presented in (4)–(8) are quite general. The above technique is also applicable to the prediction of credit defaults under the modeling framework of Carling et al. (2004) and Alam (2008). For further simplification of predictive density (9) we require specific model for Y and X_S. In the following section some special cases and their respective simplifications of (9) are presented.

4 Examples of likelihood prediction under covariate uncertainty

Prediction problems with GLM and GLMM appear in may applications and they are dealt with a variety of ways, some of which are mentioned in Section 1. We pick some examples from the existing literature and give their solutions via the predictive likelihood approach. The examples are not purposively selected; they were the only articles on prediction with GLM and GLMM under uncertainty in the response or the covariates found in the existing literature. Example 3 is related to survey sampling and arises because of an error-in-variable super population model as presented in Bolfarine (1991).

Example 3 Assume a finite population denoted by $\mathcal{P} = (1, 2, ..., N)$, where N is known and we draw a random sample of size n from $\mathcal{P}$. We denote the sample observations by $\mathbf{y} = \{y_i\}_{i=1}^n$ and the unobserved part of the population by $\mathbf{y}^* = \left\{y_j^*\right\}_{j=n+1}^N$. After observing the sample, the target is to predict the finite population total, i.e. $T = \sum_i y_i + \sum_j y_j^*$ and to provide a measure of uncertainty about the prediction. However, the y_i's are not directly observable, instead we have to use some instrument to measure y_i which gives the observation X_i such that $X_i = y_i + \delta_i$ where δ_i is a random error which is independent of y_i.

We assume that y_i's are realizations of Y_i from a super-population following a normal distribution with some constant mean and variance. We also assume that δ_i's follow the normal distribution with mean 0 and a constant variance. Under these assumptions a naive predictor of T is $\widetilde{T} = N\overline{X}$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and the variance of $\widetilde{T}$ readily found with the NP becomes BLUP (Bofarine, 1991).

The above assumptions imply that $X_i|Y_i = y_i \backsim N\left(y_i, \sigma_\delta^2\right)$. Let, $\theta = (\mu, \sigma, \sigma_\delta)$ and $\xi = (y_1, ..., y_N^*)$, $X = (X_1, X_2, ..., X_n)$, $X^* = \left(X_{n+1}^*, ..., X_N^*\right)$ with $T = \sum_i y_i + \sum_j y_j^*$ being of interest. For simplicity, we assume σ and σ_δ are known[2].

The above normality gives the following likelihood

$$L_{\theta,\xi,X^*} = \prod_{i=1}^n \left(f\left(X_i|y_i,\theta\right) f\left(y_i|\theta\right)\right) \prod_{j=n+1}^N \left(f\left(X_j^*|y_j^*,\theta\right) f\left(y_j^*|\theta\right)\right)$$

[2] See Bolfarine (1991) and Buzas and Stefanski (1996) for further discussion on the problems induces by unknown σ and σ_δ

$$
\begin{aligned}
\Rightarrow \quad L_{\theta,\xi} &= \prod_{i=1}^{n}\left(f\left(X_i|y_i,\theta\right)f\left(y_i|\theta\right)\right)\prod_{j=n+1}^{N}\int_{X_j^*} f\left(X_j^*|y_j^*,\theta\right)f\left(y_j^*|\theta\right)dX_j^* \\
&= \prod_{i=1}^{n}\left(f\left(X_i|y_i,\theta\right)f\left(y_i|\theta\right)\right)\prod_{i=n+1}^{N} f\left(y_i^*|\theta\right) \\
\therefore \quad L_{\theta,\xi} &\propto \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}\left(\frac{y_i-X_i}{\sigma_\delta}\right)^2+\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2+\sum_{j=n+1}^{N}\left(\frac{y_j^*-\mu}{\sigma}\right)^2\right\}\right]
\end{aligned}
$$

Denoting, $\overline{Y}^* = \frac{1}{N}\left(\sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} y_i^*\right)$ we have

$$
L_P^{(1)}\left(\xi|X,\sigma_\delta,\sigma\right) \propto \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}\left(\frac{y_i-X_i}{\sigma_\delta}\right)^2+\sum_{i=1}^{n}\left(\frac{y_i-\overline{Y}^*}{\sigma}\right)^2+\sum_{j=n+1}^{N}\left(\frac{y_j^*-\overline{Y}^*}{\sigma}\right)^2\right\}\right] \tag{10}
$$

Differentiating (10) w.r.t. y_i and setting them to zero gives $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} X_i$. Again doing the same for y_j^* we have $\sum_{j=n+1}^{N} y_j^* = \frac{N-n}{n}\sum_{i=1}^{n} X_i$. Adding the above two results we obtain $\widehat{T} = \frac{N}{n}\sum_{i=1}^{n} X_i = N\overline{X}$. The above $\widehat{T}$ is an unbiased estimator for T and its variance can be calculated as

$$
\begin{aligned}
Var\left(\widehat{T}-T\right) &= Var\left(N\overline{X}_s - \sum_{i=1}^{N} Y_i\right) \\
&= N\frac{1-f}{f}\sigma^2 + \frac{N}{f}\sigma_\delta^2
\end{aligned}
$$

where, $f = \frac{n}{N}$. From equation (**??**) we see that $L_{\theta,\xi}$ is proportional to the Bayesian posterior with flat prior. The Laplace approximation applied to $L_{\theta,\xi}$ in order to obtain $L_P^{(1)}\left(\xi|X,\sigma_\delta,\sigma\right)$ is exact since the log-posterior is a quadratic function. Thus the Bayesian solutions presented in Bolfarine (1991) are identical to the predictive likelihood solutions. The above $\widehat{T}$ is also the BLUP (Bolfarine, 1991).

Example 3 deals with measurement uncertainty in the response but not in the covariates. A theoretical example of dealing with uncertainties both in the $\mathbf{Y}$ and the $\mathbf{X}$ space under the linear model's framework is also presented in Bolfarine (1991). Next, Example 4 gives a prediction problem with GLM under covariate uncertainty. We consider a Poisson GLM with one covariate which is measured with error. This example is originally presented in Huwang and Hwang (2002) but their method of solution was different.

Example 4: Consider a Poisson model, $Y_i|U_i \backsim Poisson\left(\mu_i\right)$, $\log\left(\mu_i\right) = \eta_i = \beta_0 + \beta_1 U_i$ and $X_i = U_i + \delta_i \ \forall i = 1,2,...,n$. We also assume that $U_i \backsim N\left(\mu_u, \sigma_u^2\right)$, $\delta_i \backsim N\left(0, \sigma_\delta^2\right)$ and $U_i \perp \delta_j \ \forall i,j$. Our target is to predict $Y_{n+1} = y_{n+1}^*$ when X_i, $i = 1,2,...n+1$, and y_i ,$i = 1,2,...,n$, are observed but U_i's are not observable.

From the virtue of the normality and independence of U and δ we have $V_i = \left(U_i, X_i\right)^T = N_2\left(\mathbf{1}_2\mu_u, \mathbf{\Lambda}\right)$ where, $\mathbf{1}_2$ is a 2×1 column vector of 1's and $\mathbf{\Lambda} = \begin{pmatrix} \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2+\sigma_\delta^2 \end{pmatrix}$. Denote, $\theta = \left(\beta_0, \beta_1, \mu_u, \sigma_u^2, \sigma_\delta^2\right)$ and $\xi = \left(Y_{n+1}, U_1, ...U_{n+1}\right)$. Using the independence assumption we can construct the following joint likelihood

$$L_{\theta,\xi} = f\left(y^*_{n+1}|\theta, U_{n+1}, X_{n+1}\right) f\left(U_{n+1}, X_{n+1}|\theta\right) \prod_{i=1}^{n} f\left(y_i|\theta, U_i, X_i\right) f\left(U_i, X_i|\theta\right) \qquad (11)$$

The second term in the right-hand-side of equation (11) is the pdf of a bivariate normal distribution. Therefore, the joint distribution of $f\left(U_i, X_i|\theta\right)$ in the likelihood can be factored as

$$f\left(U_i, X_i|\theta\right) = f\left(U_i|X_i\theta\right) f\left(X_i|\theta\right)$$

Defining, $E\left(U_i|X_i\right) = \gamma_0 + \gamma_1 X_i$ and $\tau^2 = Var\left(U_i|X_i\right)$ where, $\gamma_0 = \left(1-\gamma_1\right)\mu_u$, $\gamma_1 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\delta^2}$ and $\tau^2 = \sigma_u^2\left(1 - \frac{\sigma_u^2}{\sigma_u^2+\sigma_\delta^2}\right)$. Now, using the usual tricks for bivariate normal distribution (see Berger and Wolpert (1988), pp–41.4) it can be shown that $X = \left(X_1, X_2, ...X_n\right)$ is ancillary for γ_0, γ_1, τ and U. Hence, $f\left(X_i|\theta\right)$ carries no information about the parameters needed for prediction and it can therefore be ignored in the construction of the predictive likelihood. Thus, the joint likelihood (11) reduces to

$$L_{\theta,\xi} \propto \exp\left(\mathbf{y}_F^T\eta - 1^T b\left(\eta\right) - c\left(\mathbf{y}_F\right)\right)\frac{1}{\tau^{n+1}}\exp\left[-\frac{1}{2\tau^2}\sum_{i=1}^{n+1}\left(u_i - \gamma_0 - \gamma_1 X_i\right)^2\right]$$

$$\Rightarrow L_{\theta,\xi} \propto \exp\left(\mathbf{y}_F^T\eta' - 1^T b\left(\eta'\right) - c\left(\mathbf{y}_F\right)\right)\frac{1}{\tau'^{n+1}}\exp\left[-\frac{1}{2\tau'^2}\sum_{i=1}^{n+1} u_i'^2\right] \qquad (12)$$

where $\mathbf{y}_F^T = \left(y_1, ..., y_n, y^*_{n+1}\right)$, $\eta' = \beta_0' + \beta_1' x_i + u_i'$, $\beta_0' = \beta_0 + \beta_1\gamma_0$, $\beta_1' = \beta_1\gamma_1$, $\tau' = \tau\beta_1$ and $u_i' = \beta_1\left(u_i - \gamma_0 - \gamma_1 X_i\right)$. Note that the equation (12) is the joint likelihood of a Poisson-Normal mixed model. Thus we conclude that the prediction problem under the measurement error in GLM reduces to the prediction problem with its GLMM analogue. However, an exact analytical solution of the problem is not possible. In absence of an exact analytical solution we may obtain $L_P^{(1)}$ through Bayesian posterior simulation.

As a competing approach, Huwang and Hwang (2002) suggested a pseudo likelihood (PsL) method for the prediction with Poisson error-in-variable model. We consider PsL as the benchmark to compare with $L_P^{(1)}$. In order to compare the performance of $L_P^{(1)}$ with the PsL, we conduct a simulation study with $\beta_0 = \beta_1 = 1$, $\mu_u = 0$, $\sigma_u^2 = 0.25$ and $\sigma_\delta^2 = 0.1$ and 0.25. We consider the sample sizes to be $n = 30$, 50, and 100 and predict one out of sample response (y_{n+1}) based on the observed data and X_{n+1}. The choice of the parameter value and sample size matches Huwang and Hwang (2002). The computation of the $L_P^{(1)}$ is carried out through Bayesian posterior simulation implemented in OpenBugs (Spigelhalter, 2007). A flat prior, $Uniform(0, 100)$ for τ' and $N\left(0, 10000\right)$ for β_0' and β_1' was used for the Bayesian model. We compare the performances of $L_P^{(1)}$ and PsL in terms of the coverage interval and the average length of prediction intervals for a nominal level, 0.95. We use 1000 Monte-Carlo replication to obtain the results which are presented in Table 2.

Table 2 Coverage Probabilities and the average length of prediction intervals for the Poisson error-in-variable prediction (Example 4) with nominal probability 0.95.

Sample Size	$Var\left(\delta_i\right)$	Coverage probability		Length of prediction interval	
n	σ_δ^2	$L_P^{(1)}$	PsL	$L_P^{(1)}$	PsL
30	0.25	0.982	0.945	8.733	8.825
50		0.969	0.958	8.355	8.362
100		0.976	0.961	7.909	8.231
30	0.1	0.984	0.954	8.405	8.386
50		0.984	0.964	7.737	7.785
100		0.986	0.943	7.518	7.642

Note: The results of the PsL are quoted from Huwang and Hwang (2002)

Though the coverage probability for $L_p^{(1)}$ exceeds the nominal level by a big margin (Table 2), it may not be a problem of $L_P^{(1)}$, rather it may be due to discrete predictive distribution for which an exact 95% prediction interval may not be possible to construct. However, $L_p^{(1)}$ guarantees that the coverage probability is not less than the nominal level while keeping the average length of the prediction interval shorter than PsL. The average length of the $L_P^{(1)}$ decreases at a rate faster than PsL as the sample size increase.

In the simulation, σ_u^2 and σ_δ^2 are quite small and therefore a naive prediction implemented through a simple Poisson GLM of y on X does not perform bad. For example, with $n = 30, \sigma_u^2 = 0.25$ and $\sigma_\delta^2 = 0.25$ a 95% prediction interval of a simple GLM gives 94% coverage probability. However, as we increase the variance parameters to $\sigma_u^2 = 1.25$ and $\sigma_\delta^2 = 1.25$ and set $\beta_0 = 0.5$ and $\beta_1 = 1.5$ with $n = 30$, the simulation results for 95% prediction interval in $L_P^{(1)}$ still having a 98% coverage probability whereas a naive GLM prediction interval covers the true future values only in 77% cases.

The final example of prediction with GLMM under covariate uncertainty is a hypothetical model for credit risk prediction.

Example 5: Let us assume that a portfolio of loans consists of n_{kt} loans in industry k, $k = 1, 2, ..., K$ at time t, $t = 1, 2, ..., T$. The event that the i^{th} loan in industry k is default at time t is given by y_{ikt} which takes the value 1 if the loan defaults and 0 otherwise. Further, assume that the default probability is modeled as a binomial GLMM. In predicting a default event at time $t+1$, the information path at t is observed while some of the future covariates are unknown at the time. For simplicity we assume that there are $p+1$ covariates and only the value of the last covariate, $\dot{X}^*_{jk(T+1)(p+1)}$ is unknown at time T although the covariate, $X_{ikt(p+1)}$, is known to follow an $AR\left(1\right)$ process.

For simplicity we set $p = 3$ and assume that the random time effects in cluster k at each time t is distributed as $u_{kt} \backsim N\left(0, \sigma_k^2\right)$, $u_{kt} \perp u_{k't'} \forall k \neq k'$ & $t \neq t'$. Denote, the future $\dot{X}^*_{jk(T+1)(p+1)} = x^*$ and we want to predict $E\left(y^*_{jk(t+1)}\right) = \mu^*_{jk(t+1)}$. A naive approach would suggest predicting x^* from the historical data on X and then predict $\mu^*_{jk(t+1)}$ as if x^* were known and that the other model parameters also were known and equal to the MLE obtained from the observed data up to time T. However, for the likelihood principle, the joint likelihood, considering all the uncertainties, is given as

$$l\left(\theta,\mathbf{y}^{*},\kappa|y,x\right)=\int\exp\left[\mathbf{y}_{F}^{*T}\eta-1^{T}diag\left\{b\left(\eta\right)\right\}\right]f\left(u\right)f\left(x^{*}|x\right)f\left(x\right)dudx^{*}$$

$$\Rightarrow\quad l\left(\theta,\mathbf{y}^{*},\kappa|\mathbf{y},x\right)=\int\exp\left[\sum_{t}\sum_{k}\left(\sum_{i}y_{ikt}\eta_{ikt}-b\left(\eta_{ikt}\right)\right)\right]$$
$$\frac{1}{\sqrt{2\pi\sigma_{k}^{2}}}\exp\left[-\frac{u_{kt}^{2}}{2\sigma_{k}^{2}}\right]f\left(x^{*}|x,\kappa\right)f\left(x|\kappa\right)dudx^{*}$$

where $\eta_{ikt}=\beta_0+\beta_1x_{1ikt}+\beta_2x_{2ikt}+u_{kt}$ and κ represents the parameter vector required to model X. Assuming, X_{2ikt} varies only over t an AR(1) precess on X is defined as $X_{t+1}=\mu+\rho X_t+e_t;\ |\rho|<1$ and $e_t\backsim iid\ N\left(0,\sigma_e^2\right)$ giving $\kappa=\left(\mu,\rho,\sigma_e^2\right)$. The assumptions lead to a simplification of the joint likelihood

$$\begin{aligned}l\left(\theta,y^{*},\kappa|y,x\right)\quad&=\quad\int\exp\left[\sum_{t}\sum_{k}\left(\sum_{i}y_{ikt}\eta_{ikt}-b\left(\eta_{ikt}\right)\right)\right]\frac{1}{\sqrt{2\pi\sigma_{k}^{2}}}\exp\left[-\frac{u_{kt}^{2}}{2\sigma_{k}^{2}}\right].\\&\qquad f\left(x_{T+1}^{*}|x_{T},\kappa\right)f\left(x_{1}|\kappa\right)\prod_{t=2}^{T+1}f\left(x_{t}|x_{t-1},\kappa\right)dudx^{*}\end{aligned}\tag{13}$$

Thus, in order to estimate κ parameters we only need to maximize the second line of (13). However, under likelihood principle we consider the full likelihood (13) for the prediction of x^* while in a formal time series prediction (forecasting) one would predict x^* only on the basis of the the second line in (13).

5 Motivations of $L_P^{(1)}$

We provide likelihood solution of the selected examples through profile adjusted predictive likelihood, $L_P^{(1)}$. However, $L_P^{(1)}$ is not the the only choice to carry out likelihood prediction. Initially $L_P^{(1)}$ was motivated through its approximate equivalence of Bayesian posterior with flat prior (Davison, 1986). In this section we show that, apart from the Bayesian justification, $L_P^{(1)}$ does have other attractive explanations.

Bjørnstad (1990) surveyed 14 different types of predictive likelihoods. Many of them are equivalent but not all of them comply with the likelihood principle. Bjørnstad (1996) presented a definition of the proper predictive likelihood based on the likelihood principle. A predictive likelihood $L\left(y^*|Y\right)$ is said to be proper if, given two experiments E_1 and E_2, $L_\theta\left(y,y^*|E_1\right)\propto L_\theta\left(y,y^*|E_2\right)$ implies $L\left(y^*|y,E_1\right)\propto L\left(y^*|y,E_2\right)$. According to the above definition, only 5 out 14 predictive likelihoods surveyed in Bjørnstad (1990) qualify as the proper predictive likelihoods. Denoting $\widehat{\boldsymbol{\theta}}$ as the MLE of $\boldsymbol{\theta}$ based on observed data only and $\widehat{\boldsymbol{\theta}}^*$ as the MLE of $\boldsymbol{\theta}$ based on both observed and unobserved data the proper predictive likelihoods are given as

1. $L_e=L\left(y^*|y,\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_y\right)$ where, . L_e is called the estimative likelihood.

2. $L_P=L\left(y^*|y,\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^*\right)$ where, . L_P is called the profile likelihood.

3. $L_P^{(1)} = L\left(y^*|y, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^*\right) |\mathcal{I}^*\left(\widehat{\boldsymbol{\theta}}^*\right)|^{-1/2}$ where, $\mathcal{I}^* = -\frac{\partial^2 \log(L_\theta(z,y))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^*}$ which is called profile adjusted predictive likelihood.

4. $L_P^{(2)} = L_P^{(1)} \left\|\frac{\partial\widehat{\theta}}{\partial\widehat{\theta}^*}\right\|$ which is a transformation invariant version of $L_P^{(1)}$.

5. $L_P^{(3)} = \sup_\theta \left\{\frac{L_\theta(y,y^*)}{\sup_{y^*}\{f_\theta(y^*|y)\}}\right\}$

Bjørnstad (1996) did not offer any discussion as to whether all of the above 5 predictive likelihoods are equally as good. However, a careful inspection of the above 5 predictive likelihoods reveals that all of them are based on the joint likelihood and they differ only in the way they profile the nuisance parameters out of the joint likelihood. Like the naive approach, L_e does not take into account the fact that the parameter $\widehat{\boldsymbol{\theta}}_y$ is estimated. Hence, L_e undermines the uncertainty associated with the prediction. L_P can be recognized as the first order Taylor's approximation to the joint likelihood around $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^*$ while the second order Taylor's approximation to $\log(L_\theta\left(y, y^*\right))$ around $\widehat{\boldsymbol{\theta}}^*$ gives

$$L_\theta\left(y, z\right) \approx L\left(z|y, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^*\right) \exp\left[\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^*\right)^T \mathcal{I}^*\left(\widehat{\boldsymbol{\theta}}^*\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^*\right)\right]$$

Assuming normality of $\widehat{\boldsymbol{\theta}}^*$ i.e. $g\left(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{\theta}\right) = N\left(\boldsymbol{\theta}, \left(\mathcal{I}^*\left(\widehat{\boldsymbol{\theta}}^*\right)\right)^{-1}\right)$ we have

$$L_\theta\left(y, y^*\right) \approx \frac{L\left(y^*|y, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^*\right) \exp\left[\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^*\right)^T \mathcal{I}^*\left(\widehat{\boldsymbol{\theta}}^*\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^*\right)\right]}{g\left(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{\theta}\right)} g\left(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{\theta}\right)$$

$$\Rightarrow L_\theta\left(y, y^*\right) \approx L\left(y^*|y, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_z\right) |\mathcal{I}^*\left(\widehat{\boldsymbol{\theta}}^*\right)|^{-1/2} g\left(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{\theta}\right) \tag{14}$$

From (14), we see that

$$L_\theta\left(y, y^*\right) \approx L_P^{(1)} g\left(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{\theta}\right)$$

where $L_P^{(1)}$ contains information only on y^* and $g\left(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{\theta}\right)$ contains all the information on $\boldsymbol{\theta}$ in addition to partial information on y^*. Therefore, the amount of information on y^* contained in $g\left(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{\theta}\right)$ is likely to be small compared to that contained in $L_P^{(1)}$ and may be negligible. Under the above, assumption, $L_P^{(1)}$ is also the partial likelihood of y^*. Again, $L_\theta\left(y, y^*\right) = f\left(y, y^*|\boldsymbol{\theta}\right)$ and $f\left(y, y^*, \widehat{\theta}|\boldsymbol{\theta}\right) = f\left(y, y^*|\boldsymbol{\theta}\right)$ implies that $L_P^{(1)}$ is the approximate conditional distribution of y and y^* given $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^*$ i.e., $L_P^{(1)} \approx f\left(y, y^*|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^*\right)$. Thus, $L_P^{(1)}$ does not have to be motivated through the Bayesian argument rather it has its own frequentist interpretation which is missing for the other proper predictive likelihoods. $L_P^{(2)}$ is applicable only if $\widehat{\theta}^*$ can be expressed as a function of $\widehat{\theta}$ which is not possible while we need to use numerical method to obtain the maximum likelihood estimator. $L_P^{(3)}$ is also a first order Taylor's approximation around a different estimate of $\boldsymbol{\theta}$ than $\widehat{\boldsymbol{\theta}}_y$ and $\widehat{\boldsymbol{\theta}}^*$.

6 Concluding discussion

This paper demonstrate that the likelihood principle gives a unified analytic framework for predictive inference. For a particular problem in hand, one might be able to find a technique e.g. BLUP for linear models, which enjoy some nice frequentist properties. However, a generalization of those techniques may be challenging. In contrast, profile predictive likelihood method provides a general and unified principle and method. The exact computation of the profile likelihood may be problematic. Moreover, the lack of computational procedures for profile predictive likelihood is also a hindrance in implementation. We leave the last two issues for possible future work.

Though there are many predictive likelihoods in the literature we prefer profile adjusted predictive likelihood, $L_P^{(1)}$, for the following reasons. First, it has nice frequentist explanation (see section 5) and second, due to its equivalence of Bayesian posterior distribution (Davison, 1986), the computation of it can be carried out by using existing Bayesian computational procedures such as by using WinBugs. For a Poisson error-in-variable GLM (example 4), we carry out predictive inference through Bayesian posterior simulation by using OpenBugs. Simulation results show that $L_P^{(1)}$ performs better than the pseudo likelihood approach and the naive approach.

References

[1] Alam, M. and Carling, K. (2008), Computationally feasible estimation of the covariance structure in Generalized Linear Mixed Models (GLMM), *Journal of Statistical Computation and Simulation,* **78**(12), 1227-1237.

[2] Alam, M. (2008), An efficient estimation of the generalized linear mixed models with correlated random effects, in P. Brito (edt.), *Proceedings of COMPSTAT'2008* (Vol. II: Contributed Papers), 853-861, Physica-Verlag, Heidelberg.

[3] Berger, J. O. and Wolpert, R. L. (1988), The Likelihood Principle, *The IMS Lecture Notes-monograph Series* **6**, IMS, Hayward.

[4] Bjørnstad, J. F. and Sommervoll, D. E. (2001), Modelling binary panel data with nonresponse, Discussion Papers No. 297, Statistics Norway, Research Department.

[5] 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.

[6] Bjørnstad J. F. (1990), Predictive likelihood: a review (with discussion), *Statistical Science*, **5**(1), 242-256.

[7] Bolfarine, H. (1991), Finite population prediction under error-in-variables subpopulation models, *The Canadian Journal of Statistics,* **19**(2), 191–207.

[8] Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association,* **88**, 9-25.

[9] Buzas, J. S. and Stefanski, L. A. (1996), Instrumental variable estimation in generalized linear measurement error models, *Journal of the American Statistical Association,* **91**(435), 999-1006.

[10] Carling, K., Rönnegård, L. and Roszbach, K. (2004), Is firm interdependence within industries important for portfolio credit risk?, Sverige Riksbank Working Paper Series No. 168.

[11] Davison, A. C. (1986), Approximate predictive likelihood, *Biometrika,* **73**(2), 323-332.

[12] Duffie, D., Saita, L. and Wang, K. (2007), Multi-period corporate default prediction with stochastic covariates, *Journal of Financial Economics,* **83**, 635-665.

[13] Eaton, M. L. and Sudderth, W. D. (1998), A new predictive distribution for normal multivariate linear models, *Sankhyā: The Indian Journal of Statistics (A),* **60**(3), 363-382.

[14] Gelman, A., Carlin, J. B., Stern. H. S. and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman and Hall/CRC.

[15] Hinkley, D. V. (1979), Predictive likelihood, *The Annals of Statistics,* **7**(4), 718-728.

[16] Honoré, B. E. and Kyriazidou, E. (2000), Panel data discrete choice models with lagged dependent variables, *Econometrica,* **68**, 839-874.

[17] Huwang, L. and Hwang, J. T. G. (2002), Prediction and confidence intervals for nonlinear measurement error models without identifiability information, *Statistics and Probability Letters,* **58**, 355–362.

[18] Laplace, P. S. (1774), Mémoire sur la probabilité des causes par les évènemens, par M. de la Place, in *Mémoires de Mathématique et dePhysique, Présentés à l'Académie Royale des Sciences, par divers Savans & lûs dans ses Assemblées* **T. vi.,** 621-656 (Memories on the probability of causes of events, English translation by S. M. Stigler (1986), *Statistical Science,* **1**(3), 364-378).

[19] Lee. Y., Nelder. J. A. and Pawitan, Y. (2006), *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Chapman and Hall/CRC.

[20] Lauritzen, S. L. (1974), Sufficiency prediction and extreme models, *Scandinavian Journal of Statistics,* **1**, 128-134.

[21] Patel, J. K. (1989), Prediction intervals - a review, *Communications in Statistics: Theory and Methods,* **18**(7), 2393-2465.

[22] Pawitan, Y. (2001), *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press.

[23] Pearson, K. (1920), The fundamental problem of practical statistics, *Biometrika,* **13**(1), 1-16.

[24] Pesaran, M. H., Scheurman, T., Treutler, B.-J. and Weiner, S. M. (2006), Macroeconomic dynamics and credit risk: a global perspective, *Journal of Money Credit and Banking,* **38**, 1211-1262.

[25] Slud, E. and Kedem, B. (1994), Partial likelihood analysis of logistic regression and autoregression, *Statistica Sinica,* **4**, 89-106.

[26] Spiegelhalter, D., Thomas, A., Best., N. and Lunn, D. (2007), OpenBUGS User Manual (Version 3.0.2), URL:http://mathstat.helsinki.fi/openbugs/Manuals/Manual.html (last accessed June 09, 2010).

[27] Startz, R. (2008), Binomial autoregressive moving average models with an application to US recessions, *Journal of Business and Economic Statistics,* **26**, 1–8.

[28] Stigler, S. M. (1986), Laplace's 1774 memoir on inverse probability, *Statistical Science,* **1**(3), 359-378.

[29] Wand, M. P. (2002), Vector differential calculus in statistics, *The American Statistician,* **56**(1), 1-8.