Risk performance of classification decisions: a framework for posterior inference based on empirical likelihood

Yuan Liao Wenxin Jiang

Princeton University and Northwestern University
August 22, 2011

Abstract

We consider an approximate posterior approach to making joint probabilistic inference on the action and the associated risk in classification. The posterior probability is based on an empirical likelihood (EL), which imposes a moment restriction relating the action to the resulting risk, but does not otherwise require a probability model for the underlying data generating process. We illustrate with examples how this framework can be used to describe the EL-posterior distribution of actions to take in order to achieve a low risk, or conversely, to describe the posterior distribution of the resulting risk for a given action. A theoretical study on the frequentist properties reveals that the EL-posterior concentrates around the true risk-action relation with high probability for large data size, and that the actions can be generated from this posterior to reliably control the true resulting risk. Finally, an application to the German credit data is presented.

Key words: classification, posterior consistency, EL- posterior, Bayesian empirical likelihood, moment condition, partially identified models, risk

1 Introduction

One of the classical problems in data mining is to predict the unknown nature of a feature, by classifying the data into subgroups. Suppose $Y \in \{0, 1\}$ is a binary random variable to be predicted, which is associated with a vector of random predictors X. The classification and prediction are made through a classification rule $C(X, \theta) \in \{0, 1\}$, which is a function of X indexed by θ, for example, $C(X, \theta) = I(X^T\theta > 0)$.

One way of assessing the classification accuracy is to introduce a loss function $l(Y, C(X, \theta))$, and define its expectation as the *risk* $r = E(l(Y, C(X, \theta))|\theta)$, which is a function of θ. We will mostly focus on the *classification risk* r associated with the common 0-1 loss $l = |Y - C(X, \theta)|$. However, we will also briefly describe how our approach can handle more general risks in Section 4. The parameter θ now describes an $action$ related to the classification decision $C(X, \theta) = I(X^T\theta > 0)$. This paper provides a new framework to describe the relation between actions indexed by θ and their associated classification risks, which enables us to answer the following two questions without knowing the underlying distribution of (Y, X): (i) what kind of θ should be chosen in order to control the resulting risk, so that $r \leq r_0$ for a pre-determined desired level r_0? (ii) What is the range of plausible values for the classification risk r associated with any chosen classification rule θ?

Our new framework has a Bayesian flavor, and will involve deriving formally a joint posterior distribution for the action-risk pair (θ, r), based on an observed data set $D = \{(Y_1, X_1), ..., (Y_n, X_n)\}$ consisting of n i.i.d. (independent and identically distributed) copies of (Y, X). However, there are three distinctive features that make our approach very different from the traditional Bayesian approach. We only briefly explain these three differences here, and will leave the detailed discussions in Section 2.2, after we describe the formal mathematical structure of the method in

Section 2.1.

A. We use the Empirical Likelihood (EL, see, e.g., Owen 1990), instead of the true likelihood, when constructing the posterior (which will be consequently called the *EL-posterior*). This way, our method inherits a major advantage of EL, which requires only the specification of a moment condition, rather than a full probability model for (Y, X).

B. The action parameter θ in this paper is not a parameter in the usual sense, since it is not a functional of the data generating process, but only indexes a decision rule to be studied by the user. This will be discussed in more detail in Section 2.2.

C. The EL-posterior is based on a moment condition that defines the risk-action relationship:

$$E(l(Y, C(X, \theta))|\theta) = r, \tag{1.1}$$

which does not identify (r, θ) to a point, but to a curve instead. Consequently, we will show an unusual mode of posterior consistency, that with increasing sample sizes, the posterior probability mass will not converge to a point, but to the curve $r = E(l(Y, C(X, \theta))|\theta)$ instead. This feature is similar to the findings in the literature of Bayesian partial identification (e.g., Poirier (1998), Moon and Schrofheide (2010), Liao and Jiang (2010), Gustafson (2010)). But as far as we know, the current paper is the first one that proves this kind of posterior consistency result when EL is used (instead of the usual likelihood).

Note that when an EL-posterior, which we will denote by $P_{EL}(\theta, r|D)$, is formally derived for the action-risk pair jointly, we will be able to obtain useful conditional distributions. For example, Corollary 3.1 shows the following implication of the consistency of the joint posterior $P_{EL}(\theta, r|D)$: With a large data set D, when

the action parameter θ is generated from the conditional posterior $P_{EL}(\theta|r \leq r_0, D)$ (given that risk r is controlled to be at a certain nominal level $r \leq r_0$), the true expected risk $El(Y, C(X, \theta))$ will indeed often be controlled to nearly r_0 (or lower), regardless of the true distribution of the data generating process.

The remainder of this paper is organized as follows. Section 2.1 introduces the basic framework of the proposed EL-posterior. Section 2.2 discusses in detail the three distinctive features of the proposed method. Section 3 presents the main theoretical results related to the frequentist properties of the proposed method. Section 4 comments on the possible extension to more general risk functions in data mining. Section 5 provides a simulation example to illustrate the main results and demonstrates how they are used in practice. Section 6 illustrates an empirical application using the German Credit Benchmark data. Finally, Section 7 concludes with discussions.

2 Empirical Likelihood Posterior Distribution

2.1 Construction of the EL-posterior

Consider the following equation

$$r = E[\rho(W, \theta)|\theta] = \int \rho(w, \theta) dP_W(w), \tag{2.1}$$

where θ is an auxiliary parameter which indexes a decision rule, r is the resulting risk, and $W = (Y, X)$ with $Y \in \{0, 1\}$ being the label to be predicted and X being a predictor. Here $\rho(W, \theta)$ denotes the loss function. To simplify the technical derivation and illustrate the main idea, this paper mainly considers the absolute loss function (also known as the l_1 loss) $\rho(W, \theta) = |Y - C(X, \theta)|$, where $C(X, \theta) \in \{0, 1\}$ is a classification rule that is indexed by θ. The results obtained can be naturally

generated to general loss functions. The probability measure P_W is based on the true distribution of W, which is unknown. We assume that θ belongs to an action space Θ, and $r \in [0,1]$. Therefore the parameter space for (θ, r) is $\Theta \times [0,1]$.

Instead of assuming a full probability model for W, we construct an approximate posterior distribution based on the empirical likelihood, jointly for the action-risk pair (θ, r), using only the moment condition (2.1). The empirical likelihood (EL) posterior is constructed as follows. Suppose we observe data $D = (W_1, ..., W_n)$, which are assumed to be independent and identically distributed realizations of W. The profile EL based on (2.1) is defined by (Owen (1990) and Qin and Lawless (1994)):

$$\begin{aligned} L_n(\theta, r) &= \sup_{p_1,...,p_n} \{\prod_{i=1}^{n} p_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i[\rho(W_i, \theta) - r] = 0\} \\ &= \exp\left(-\sum_{i=1}^{n} \log\{1 + \mu(\theta, r)[\rho(W_i, \theta) - r]\} - n \log n \right), \end{aligned} \quad (2.2)$$

where

$$\mu(\theta, r) = \arg\max_{\mu} \sum_{i=1}^{n} \log\{1 + \mu[\rho(W_i, \theta) - r]\}.$$

We combine $L_n(\theta, r)$ with a joint prior $\pi(\theta, r)$ via the Bayes' formula to obtain a pseudo posterior (called the "EL posterior", up to a normalization factor):

$$P_{EL}(\theta, r|D) \propto \exp\left(-\sum_{i=1}^{n} \log\{1 + \mu(\theta, r)[\rho(W_i, \theta) - r]\} \right) \pi(\theta, r). \quad (2.3)$$

This formalism of EL-posterior provides a useful starting point that allows us to derive a number of interesting results, which describe the relationship between the action and the resulting risk. For example:

1. $P_{EL}(r|\theta, D)$, which is the EL-posterior of the resulting risk achieved by a given action θ;

2. $P_{EL}(\theta|r = r_0, D)$ or $P_{EL}(\theta|r \leq r_0, D)$, which is the EL-posterior of the action θ needed to achieve a risk either being r_0 or at most r_0;

3. $P_{EL}(\theta \in A_1|r \leq r_0, D)/P_{EL}(\theta \in A_2|r \leq r_0, D)$, which compares the posterior probabilities of two models A_1 and A_2 (or two sets of actions) given that $r \leq r_0$, to see which model is "more likely" (corresponding to higher EL-posterior probability) given that r_0 is achieved.

4. $P_{EL}(r|D)$, which is the EL-posterior of the achievable risk by all possible actions θ proposed by $\pi(\theta)$.

2.2 Three distinctive features of the proposed approach

It is easy to notice that the approach proposed above has three features that are different from most other works. First, the EL-posterior is different from common Bayesian posteriors, since it now uses EL. Second, the parameter structure is unusual, since the action parameter is not a functional of the data generating process. Third, the posterior distribution behaves differently from common situations of point-identification, since it does not converge to a single point asymptotically. We will discuss these three differences below in detail.

2.2.1 The use of EL in posterior construction

Exactly for the same reasons that EL is a valuable alternative to the usual likelihood for a frequentist approach, the EL-posterior is also a valuable alternative to the usual posterior. This point is realized much less in the Bayesian world than in the frequentist counterpart, although we believe that this situation is now changing, largely due to recent works of, e.g., Chernozhukov and Hong (2003), Lazar (2003),

Ragusa (2007). When making inference of a parameter of interest appearing in a moment condition/estimating equation, the EL-based method (whether frequentist or Bayesian in style) has a key advantage, that its validity is roughly equivalent to that of a semiparametric likelihood approach, but the latter would require estimating infinitely many more nuisance parameters which are not directly relevant to the moment condition.

Recent theoretic and numeric studies suggest that the EL-posterior probability may be regarded as an approximation to the true posterior probability, rather than just a common-sense construction. For example, Kitamura (2001) understands EL as an approximate likelihood based on a probability model that is closet to the true model (in the Kullback-Leibler divergence), among all the probability models satisfying the same defining moment condition. Lazar (2003) provides numerical evidence of the validity of the use of EL-posterior for Bayesian inference, under the definition of "posterior validity" proposed by Monahan and Boos (1992).

In addition, in the classification problem studied in the current paper, we will show later that the EL-posterior (based on the same moment condition) is equivalent to the so-called "Bayesian exponentially tilted empirical likelihood (BETEL) posterior" proposed by Schennach (2005). It is shown in Schennach (2005) that the BETEL-posterior arises naturally from a nonparametric Bayesian procedure, with a type of noninformative prior placed on the space of distributions. Since the EL-posterior is exactly the same as the BETEL-posterior when the moment condition is given by $E(|Y - C(X, \theta)||\theta) = r$, Schennach (2005)'s argument provides an additional probabilistic interpretation of the EL-posterior used in our paper [1].

In practice, the EL-posterior has been used by a number of authors in applied

[1] See Appendix A.3 for the definition of BETEL-posterior proposed in Schennach (2005), as well as the proof of its equivalence to our EL-posterior in the classification problem.

problems., e.g., Chaudhuri and Ghosh (2010) and Rao and Wu (2010).

2.2.2 The action parameter and the prior

The action parameter θ is not a usual parameter, since it is not a functional of the data generating process. Due to the quasi-Bayesian framework, we have the flexibility of placing a prior $\pi(\theta, r) = \pi(r|\theta)\pi(\theta)$, in which $\pi(\theta)$ describes a set of θ's whose risk performances the decision maker would like to study, and $\pi(r|\theta)$ is the prior guess of the risk once an action θ is taken. For example, suppose X is one dimensional and a linear classification rule $C(X, \theta) = I(X > \theta)$ is applied, in which θ can be any threshold value a decision maker chooses from $\mathbb{R}$. Placing a standard normal prior on θ means that the decision maker is mostly interested in studying the risk behavior of those θ's that are close to zero.

The result of the proposed procedure will be relative to the given prior $\pi(\theta)$, supplied by the decision maker. Given this choice $\pi(\theta)$, we will get an idea about how low the risk r can be realistically achieved by this group of proposed θ's, by looking at $P(r|D)$. We will also be able to know which part of the θ's proposed by the prior $\pi(\theta)$ will achieve a risk lower than some threshold value r_0, by looking at $P(\theta|r < r_0, D)$. The result will be valid for any r_0 that is not too much lower than the lowest point of the support of $P(r|D)$ (see Corollary 3.1 below).

2.2.3 Partial identification

The defining moment condition (1.1) that relates the risk r to the action θ clearly does not identify the pair (θ, r) uniquely. It is the functional relationship $r = E(l(Y, C(X, \theta))|\theta)$ itself, rather than any single point satisfying this relation, that is identifiable. Our theoretical result later shows that the posterior distribution will

not converge to a point, no matter how large the data set is. The limiting posterior behavior is therefore more similar to the findings in the literature of partial identification.

In recent years, partially identified models are receiving rapidly growing attentions in both statistics and econometrics literatures. Manski (2007) and Tamer (2010) both give excellent reviews and discussions of the applications of these models in social sciences. In these models, the identified region becomes the object of interest. (See for example, Chernozhukov, Hong and Tamer (2007).) In our paper, the identified region corresponds to the set of parameters that satisfy the moment restriction (1.1). More recently, Liao and Jiang (2010) have studied the properties of the posterior distribution of the parameters in a similar setting of the moment restriction (1.1), where they used the limited information likelihood idea (Kim 2002) to construct the likelihood function. In this paper, we show that the EL-posterior has similar asymptotic properties to those described in Liao and Jiang (2010). To be specific, Theorem 3.2 implies that the joint EL-posterior distribution for (θ, r), denoted by $P_{EL}(\theta, r|D)$, will be asymptotically supported on an arbitrarily small neighborhood of the part of the curve $\{(\theta, r) : El(Y, C(X, \theta)) = r\}$ that lies within the support of the prior $\pi(\theta, r)$. So far, the consistency of the EL-posterior for partially identified models has not been formally established, while the point identified case was previously studied by Chernozhukov and Hong (2003). Therefore one of the contributions of the current paper is the EL-posterior consistency for the parameters that are only partially identified by the moment condition of the form (1.1).

3 Main Results

In the classification problem when $\rho = |Y - C(X, \theta)|$ and Y and $C(X, \theta) \in \{0, 1\}$, it is straightforward to verify that the empirical likelihood and the corresponding posterior distribution have explicit analytic expressions. The following theorem states that, in the classification problem framework, the log-empirical likelihood function is proportional (up to the scale $-n$) to the Kullback-Leibler distance between two Bernoulli distributions with success probabilities, $\hat{R}(\theta)$ and r, respectively.

Theorem 3.1. *Suppose $\rho = |Y - C(X, \theta)|$, Y and $C(X, \theta) \in \{0, 1\}$, $\hat{R}(\theta)$ and $r \in [0, 1]$, where $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - C(X_i, \theta)|$. Then the EL-posterior in (2.3) is given by*

$$P_{EL}(\theta, r|D) \propto \pi(\theta, r) \exp(-nK(\hat{R}(\theta), r)), \tag{3.1}$$

where

$$K(p, q) = \begin{cases} p \ln(p/q) + (1-p) \ln\{(1-p)/(1-q)\}, & \text{if } p, q \in (0, 1) \\ +\infty, \quad \text{if } p \in (0, 1], q = 0, \text{ or } p \in [0, 1), q = 1 \\ 0 \quad \text{if } q \in [0, 1), p = 0, \text{ or } q \in (0, 1], p = 1. \end{cases} \tag{3.2}$$

Proof. See the Appendix.

As described in Section 2.2, (θ, r) is not point identified. As a result the posterior does not degenerate to any single point asymptotically. Based on a treatment similar to Liao and Jiang (2010), however, we can establish the posterior consistency under partial identification, which is, as $n \to \infty$, the EL-posterior concentrates around the region of (θ, r) that satisfy the moment condition $E|Y - C(X, \theta)| = r$.

Suppose a random variable $Z_n = Z_n(D)$ is a function of D. In what follows, we write $Z_n \overset{P_W^n}{\to} 0$ if $\forall \epsilon > 0$, $P_W(|Z_n(D)| > \epsilon) \to 0$.

Theorem 3.2. *Consider the classification case, when* $\rho(W, \theta) = |Y - C(X, \theta)|$ *and* $Y, C(X, \theta) \in \{0, 1\}$. *Denote* $R(\theta) = E[\rho(W, \theta)|\theta]$, $\hat{R}(\theta) = n^{-1} \sum_{i=1}^{n} \rho(W_i, \theta)$, *and* $\eta(\theta, r) = \min\{R(\theta), 1 - R(\theta), r, 1 - r\}$. *Assume the following:*

(i) The prior $\pi(|R(\theta) - r| \leq \delta, \eta \geq \tau) > 0$ *for any constants* $\delta > 0, \tau \in (0, 1)$*;*

(ii) $\sup_{\theta \in \Theta} |\hat{R} - R| \overset{P_W^n}{\to} 0$ *as* $n \to \infty$.

Then for any $\epsilon > 0$, *we have: as* $n \to \infty$

$$P_{EL}(R(\theta) - \epsilon \leq r \leq R(\theta) + \epsilon | D) \overset{P_W^n}{\to} 1.$$

Proof. See the Appendix.

Condition (i) imposes a regularity assumption on the support of prior. Condition (ii) requires the uniform convergence in probability of the empirical risk $\hat{R}(\theta)$ to the true risk $R(\theta)$. This condition holds, for example, if the class of decision rules $\mathcal{C} = \{C(X, \theta) : \theta \in \Theta\}$ has finite Vapnik-Chervonenkis dimension V_C, in which case $P(\sup_{\theta \in \Theta} |\hat{R}(\theta) - R(\theta)| > \epsilon) \leq 8 n^{V_C} e^{-n\epsilon^2/32}$ (see Vapnik and Chervonenkis (1971), and Devroye et al. (1996) chapter 12).

We give an intuitive explanation of Theorem 3.2. A generalized posterior consistency theory implies that the posterior distribution should be asymptotically supported on the set of maximizers of the log-likelihood (see Chernozhukov and Hong (2003)), which in our case, is the set of minimizers of $K(\hat{R}(\theta), r)$. Since $K(\hat{R}(\theta), r)$ is the Kullback-Leibler distance between two Bernoulli distributions with success probabilities $\hat{R}(\theta)$ and r respectively, $K(\hat{R}(\theta), r)$ is thus minimized when $\hat{R}(\theta)$ and r are close to each other. By the uniform convergence of $\hat{R}(\theta) - R(\theta)$, therefore, the posterior should asymptotically concentrate around the region where $R(\theta)$ is close

to r. This theorem indicates that, even though the EL-posterior of the true risk $R(\theta)$ is unknown (since it also depends on the unknown distribution P_W), we can make inference of $R(\theta)$ based on the EL-posterior of r.

The following corollaries describe two useful implications. If we would like to find actions to control the risk to be under r_0, we can use actions generated from the conditional posterior $P_{EL}(\theta|r \leq r_0, D)$, which tends to generate actions that result the risk $R(\theta)$ to be at most slightly worse than the desired level r_0.

Corollary 3.1. *Suppose that $P_{EL}(r \leq r_0|D) > \xi$ for some constant $\xi > 0$, then under the regularity conditions in Theorem (3.2), for any $\epsilon > 0$,*

$$P_{EL}(E[\rho(W,\theta)|\theta] \leq r_0 + \epsilon|D, r \leq r_0) \rightarrow^{P_W^n} 1$$

as $n \rightarrow \infty$.

Proof. Denote $R = E[\rho(W,\theta)|\theta]$. Then $P_{EL}(R > r_0 + \epsilon|D, r \leq r_0) = P_{EL}(R - r_0 > \epsilon, r \leq r_0|D)/P_{EL}(r \leq r_0|D) \leq P_{EL}(|R - r| > \epsilon|D)\xi^{-1} \rightarrow^{P_W^n} 0$ due to Theorem (3.2). Q.E.D.

Corollary 3.1 indicates the usefulness of $P_{EL}(\theta|r \leq r_0, D)$ in controlling the classification risk, which shows that, if the action parameters are generated from $P_{EL}(\theta|D, r \leq r_0)$, then with high EL-probability, the true risk $E[\rho(W,\theta|)\theta]$ will be less than r_0 plus an arbitrarily small constant, where r_0 is the level of the risk that the decision makers can tolerate. In practice, once a sequence of $\{\theta_i\}$ is generated from the distribution $P_{EL}(\theta|D, r \leq r_0)$, we can take an action as the average of these posterior draws. See Section 6 for the implementation in a real data example.

Define $r^* = \inf_{\theta \in \Theta} E[\rho(W,\theta)|\theta]$, the minimal expected risk over all the actions in Θ. The next corollary states that asymptotically, the EL-posterior distribution of r has no support below r^*.

Corollary 3.2. *Under the regularity conditions in Theorem (3.2), for any* $\epsilon > 0$*,* $P_{EL}(r < r^* - \epsilon|D) \to^{P_W^n} 0$ *as* $n \to \infty$*.*

Proof. $P_{EL}(r < r^* - \epsilon|D) = P_{EL}(r + \epsilon < \inf_\theta R|D) \to^{P_W^n} 0$ due to Theorem (3.2).

For numerical computation, we point out that the Beta distribution is a conjugate prior for the conditional posterior distribution of $r|\theta, D$. Specifically, if the priors for θ and r are independent, and $\pi(r)$ is Beta(a, b), then straightforward calculation yields:

$$\begin{aligned} P_{EL}(r|\theta, D) &\sim \text{Beta}(n\hat{R}(\theta) + a, n(1 - \hat{R}(\theta)) + b) \\ P_{EL}(\theta|D) &\propto \pi(\theta)\hat{R}(\theta)^{-n\hat{R}(\theta)}(1 - \hat{R}(\theta))^{-n(1-\hat{R}(\theta))}B(n\hat{R}(\theta) + a, n(1 - \hat{R}(\theta)) + b), \end{aligned}$$

where $B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1}dx$. It is recommended in practice to generate (θ, r) jointly from $P_{EL}(\theta, r|D)$, as it allows us to investigate $P(\theta|r \le r_0, D)$ for various levels of r_0 from a single set of MCMC draws (see Section 6.3). On the other hand, if one is particularly interested in a specific r_0 that the decision maker can tolerate, the procedure can be simplified to a single MCMC of θ from $P_{EL}(\theta|r \le r_0, D)$, as this distribution also has a simple expression after marginalizing onto θ:

$$P_{EL}(\theta|r \le r_0, D) \propto P_{EL}(\theta|D)F_{\text{Beta}}(r_0, n\hat{R}(\theta) + a, n(1 - \hat{R}(\theta)) + b), \qquad (3.3)$$

where $F_{\text{Beta}}(x, n\hat{R}(\theta) + a, n(1 - \hat{R}(\theta)) + b)$ denotes the cumulative distribution function of Beta$(n\hat{R}(\theta) + a, n(1 - \hat{R}(\theta)) + b)$, which is an incomplete Beta function.

4 More General Risk Functions

It is noted that while we have been focusing on the classification risk, the current method and theoretical results can be easily generalized to other risk functions of the form $R(\theta) = E[\rho(W, \theta)|\theta]$, where $\rho(W, \theta) = f(Y, C)$ with $Y \in \{0, 1\}$ and $C =$

$C(X,\theta) \in \{0,1\}$. (One simple example is a linear rule $C = I(X^T\theta > 0)$ indexed by θ.) For one example in a data mining context: A marketing effort $C = I$[mail] of mailing out an advertisement with cost $c = 1$ will be based on X (including, e.g., gender, age, ethnic group, education, etc). The outcome will be $Y = I$[purchase] where a purchase will lead to net income $g = 100$. Then one would like to maximize the expected profit $E[(gY - c)C]$ or minimize a risk $R = \text{constant} - E[(gY - c)C]$. Here up to a constant, $f(Y,C) = -(gY - c)C$, so that $f(0,0) = f(1,0) = 0$, $f(0,1) = c = 1$, $f(1,1) = c - g = -99$. Such profit-and-loss decision matrices are included in popular data mining software such as the SAS Enterprise Miner. We can apply the proposed method to construct the EL-posterior distribution 2.3 jointly for the action parameter θ and the resulting risk r.

5 A Simulated Example

Let the data be generated from the following design

$$Y = I_{(3X-\epsilon>0)},$$
$$X \sim N(0,1), \epsilon \sim N(0,3),$$

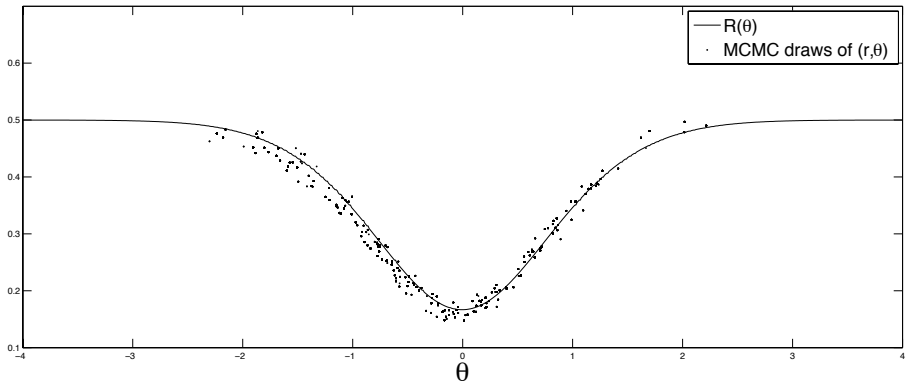
where X and ϵ are independent. We apply the classification rule $C(X,\theta) = I_{(X-\theta>0)}$. Let $\rho(W,\theta) = |Y - C(X,\theta)|$. One can then show that the expected risk is given by

$$R(\theta) = E[\rho(W,\theta)|\theta] = E_X\{[1 - \Phi(\sqrt{3}X)]I_{(X>\theta)} + \Phi(\sqrt{3}X)I_{(X\leq\theta)}\} \qquad (5.1)$$

where the expectation E_X is taken with respect to the distribution of X, which is standard normal, and $\Phi(.)$ denotes the cumulative distribution function of $N(0,1)$. We generated $n = 2,000$ data points $(Y_1,X_1),...,(Y_n,X_n)$. The EL-posterior for (θ,r) was constructed based on (3.1), with priors $\pi(\theta) \sim N(0,1)$, $\pi(r) \sim$ Uniform$[0,1]$, and $\pi(\theta,r) = \pi(\theta)\pi(r)$. According to Theorem 3.2, the posterior distribution should

concentrate around the risk curve $\{(\theta, R(\theta))\}$. To illustrate this, $B = 10,000$ MCMC draws were then generated from the EL-posterior. In each step of the Metropolis algorithm, we used proposal density $\theta^t \sim N(\theta^{t-1}, 0.5)$, and $r^t \sim U[0,1]$. The first quarter of the draws were treated as the "burn-in" period and were discarded.

Figure 1: Plot of $R(\theta)$ and MCMC draws

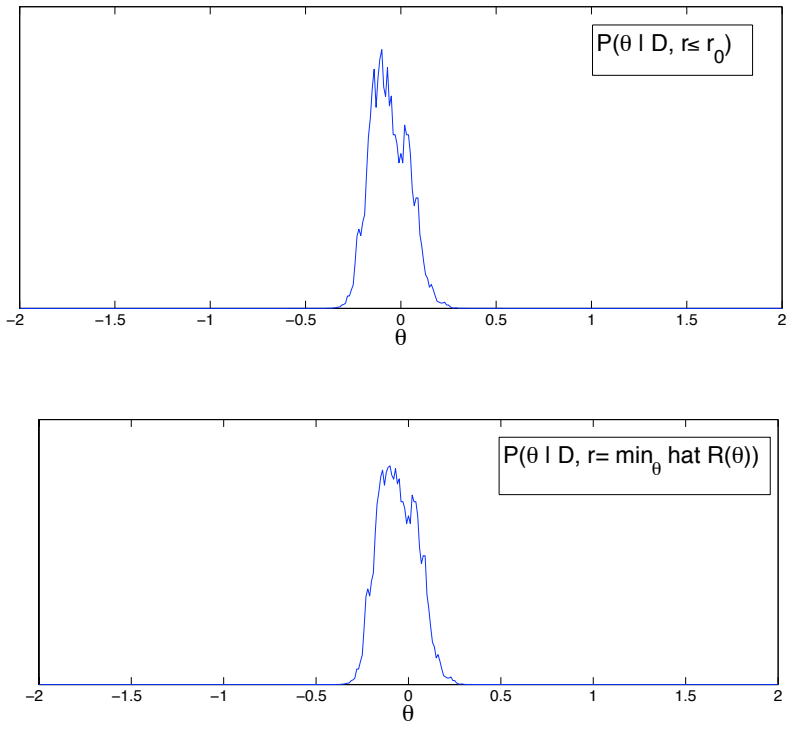


The solid line in Figure 1 represents the expected risk curve $R(\theta)$ against θ, and the dots represent the scatterplot of the MCMC draws of (θ, r). It is clearly illustrated that the MCMC draws are clustered around the true expected curve, supporting our posterior consistency result. Therefore the scatterplot of the MCMC draws can clearly demonstrate the functional relationship of the true risk $R(\theta)$ resulting from each action θ, even though the true distribution of (Y, X) is unknown in practice.

In addition, our method also gives the posterior distribution of the action θ to achieve a certain risk level. For example, if we want to control the classification risk to be no greater than the 5th percentile of the posterior distribution of r, we

could let r_0 be the 5th percentile of the MCMC draws of r, which is $r_0 = 0.163$, and focus on $P_{EL}(\theta|D, r \leq r_0)$. If we want to control the classification risk to be at the minimized empirical risk $\min_\theta \hat{R}(\theta) = 0.159$, we could focus on the posterior $P_{EL}(\theta|D, r = \min_\theta \hat{R}(\theta))$. See Figure 2 as the plotted posterior densities of θ in these two cases respectively. Both graphs demonstrate that the EL-posterior has high density levels around zero, which is the global minimizer of $R(\theta)$ in (5.1).

Figure 2: EL-posterior densities of $P_{EL}(\theta|D, r \leq r_0)$ and $P_{EL}(\theta|D, r = \min_\theta \hat{R}(\theta))$



We also plot $P_{EL}(r|D, \theta = \arg\min_\theta \hat{R}(\theta))$, which is the EL-posterior of the risk given that the empirical risk minimizer is taken. The plot can be compared with the marginal EL-posterior of r (See Figure 3), which is obtained by numerically integrating out θ. While the marginal posterior $P_{EL}(r|D)$ is supported approximately on $[0.15, 0.5]$, the density is high when r is between 0.15 and 0.2. Moreover, after conditioning on $\theta = \arg\min_\theta \hat{R}(\theta)$, the posterior of r is only supported (and has

very high posterior density) around $[0.15, 0.2]$, which is a small interval containing the minimal classification error $\min R(\theta) = 0.1667$ based on $C(X, \theta) = I_{(X-\theta>0)}$. Figure 3 also illustrates that $P_{EL}(r|D)$ has not much support when $r < \min R(\theta)$.

Figure 3: EL-posterior densities of $P_{EL}(r|D, \theta = \arg\min_\theta \hat{R}(\theta))$ and $P_{EL}(r|D)$

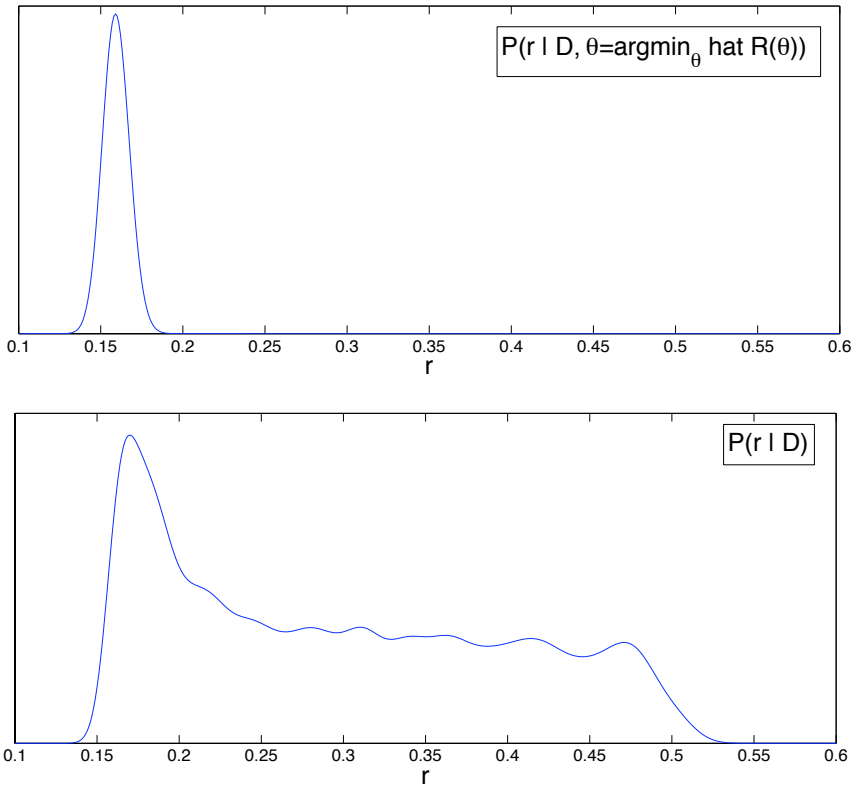
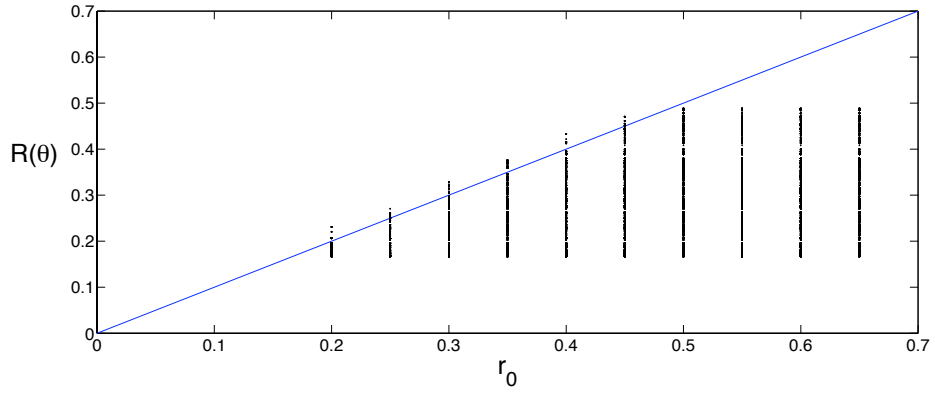


Figure 4 plots the theoretical (expected) risk $R(\theta_i) = E(\rho(W, \theta)|\theta = \theta_i)$ versus the controlled levels of risk $r_0 = 0.2, 0.25, 0.3, ..., 0.65$, where each θ_i is from the MCMC draw (θ_i, r_i), whose paired r_i is less than or equal to r_0. In other words, one can think of the dots in Figure 4 with the same horizontal coordinate r_0 as $\{R(\theta_i)\}$, where $\{\theta_i\}$ are generated from the EL-posterior density $P_{EL}(\theta|r \leq r_0, D)$. We can see that almost all the dots are below the identical line $R(\theta) = r_0$ (with very low percentage of exceptions). This indicates that, once the action θ is generated from $P_{EL}(\theta|r \leq r_0, D)$, the true risk can be effectively controlled to be no greater than

r_0 as well. Finally, none of the elements in $\{R(\theta_i)\}$ are less than 0.166. This is because the minimal expected risk that can be achieved by the linear classification rule $C(X, \theta) = I_{(X \geq \theta)}$ is $R(0) = 0.1667$.

Figure 4: Scatter plot of $(R(\theta_i), r_0)$



The dots with the same horizontal coordinate r_0 represent $\{R(\theta_i)\}$ where $\{\theta_i\}$ are generated from $P_{EL}(\theta | r \leq r_0, D)$. When $r_0 \leq 0.5$, 97.3% dots are below the identical line $R(\theta) = r_0$.

6 German Credit Data: an Empirical Application

6.1 Data set and model specification

As an empirical example, we apply the proposed Bayesian empirical likelihood method to the credit risk classification problem, using the German Credit Benchmark data set provided by Asuncion and Newman (2007). The data set consists of n =1,000 past applicants and their credit rating (GOOD or BAD), which serves as the target variable Y. In addition, there are 24 attributes served as input variables, which are used as the predictors X. The attributes are either ordered categorical, such as "Credit History", "Personal Status and Sex", "Housing", "Employment", or

numerical, such as "Credit Duration", "Credit Amount", and "Age". See Asuncion and Newman (2007) for a complete description of the data set.

The classification rule using the jth observation is $C(X_j, \theta) = I(X_{1j}\theta_0 + \theta_1 + \sum_{i=2}^{24} X_{ij}\theta_i > 0)$. Here X_{ij} denotes the realization of individual j on variable $X_i, i = 1, ..., 24$. Note that for the purpose of normalization in regular binary regression, Horowitz (1992) suggested that X_1 should be a continuous variable whose coefficient $\theta_0 \in \{-1, +1\}$. Therefore, the components of the covariates are arranged so that X_{1j} denotes the jth observation of "Credit Duration", which is a continuous variable, and it is reasonable to assume that it is related to each customer's credit behavior. We also included an intercept θ_1. In addition, the continuous attributes were normalized, i.e., subtracted mean and divided by the standard deviation.

For the credit classification problem, the cost matrix is given by the following table as in West (2000):

Table 1: Cost Matrix

		Classification	
		GOOD	BAD
Target Variable	GOOD	0	1
	BAD	5	0

Note that the cost matrix is asymmetric, this is because the penalization for misclassifying a Bad target variable Y into Good should be more severe than the opposite situation. Therefore the loss function is

$$\rho(Y, X; \theta) = I(Y = Good, C(X, \theta) = Bad) + 5I(Y = Bad, C(X, \theta) = Good).$$

The method proposed in the previous sections can be applied for variable selection. Let $\psi = (\psi_1, ..., \psi_{24})$ denote a vector of selection indicators such that for each $i =$

$1, ..., 24$, $\psi_i = 1$ if variable X_i is selected, and $\psi_i = 0$ otherwise. We then have the selected parameters $\theta^\psi = (\theta_1\psi_1, ..., \theta_{24}\psi_{24})$. A zero component of θ^ψ corresponds to an unselected covariate. We set $\psi_1 = 1$ so that the intercept θ_1 is always kept in the model. Therefore the actual classification rule based on θ_ψ is $C(X_j, \theta^\psi) = I(\theta_0 X_{1j} + \theta_1 + \sum_{i=2}^{24} X_{ij}\theta_i\psi_i > 0)$, where $\theta_0 \in \{-1, 1\}$.

The log-empirical likelihood function for (θ, r, ψ) is given by

$$\log EL(\theta, r, \psi) = -\max_{\mu\in\mathbb{R}} \sum_{i=1}^{n} \log\{1 + \mu[\rho(Y_i, X_i, \theta^\psi) - r]\}.$$

We want to obtain an action θ^ψ so as to achieve a low risk. Hence for a predetermined threshold value r_0, a conditional prior is placed: $\pi(\theta, r, \psi | r \le r_0) = \pi(\theta|\psi, r, r \le r_0)\pi(\psi, r|r \le r_0)$. It is assumed that $\pi(\theta|\psi, r, r \le r_0) = \pi(\theta^\psi|\psi)$, and $\pi(\psi, r|r \le r_0) = \pi(\psi)\pi(r|r \le r_0)$, in which

$$\pi(\psi) = \prod_{i=2}^{24} \pi(\psi_i), \qquad \begin{aligned} \pi(\theta^\psi|\psi) &\sim N(0, 10I_{|\psi|}), \\ \pi(\psi_i) &\sim \text{Binomial}(1, \lambda), \\ \pi(r|r \le r_0) &\sim \text{Uniform}[0, r_0]. \end{aligned}$$

where $|\psi|$ denotes the number of nonzero components of ψ, $I_{|\psi|}$ denotes the $|\psi| \times |\psi|$ identity matrix, and λ represents the prior expectation of the fraction of selected variables, set to 0.5 in our application. Therefore we *a priori* expect half of the variables to be selected. The posterior distribution is then given by

$$P_{EL}(\theta, \psi, r|D, r \le r_0) \propto EL(\theta, r, \psi) e^{-\frac{1}{20}\sum_{i=1}^{24}\theta_i^2\psi_i} \lambda^{|\psi|-1}(1-\lambda)^{24-|\psi|} I(0 < r < r_0).$$

6.2 Algorithm description

The Metropolis-Hastings algorithm was conducted to obtain the MCMC draws from the posterior distribution. We use the algorithm described in Chen, Jiang and

Tanner (2010), in which each iteration combines BETWEEN steps that propose changes of ψ for selecting different variables, with the WITHIN steps that propose changes of θ once ψ is fixed. These steps are described as follows: (in the algorithm below, denote $q(\theta)$ as the density function of $N(0, 0.5)$.)

BETWEEN Step Update θ to θ' with model indices changing from ψ to ψ'.

1. (Add/Delete) Randomly choose an index $j \in \{2, ..., 24\}$.

   - If $\psi_j = 1$, propose $\tilde{\psi}_j = 0$ and let $\tilde{\theta} = \theta$. Let $\tilde{\psi}_i = \psi_i$ for $i \neq j$. This proposal is accepted with probability

   $$\min\left\{1, \frac{P_{EL}(\tilde{\theta}, \tilde{\psi}, r|D)q(\theta_j)}{P_{EL}(\theta, \psi, r|D)}\right\}.$$

   - If $\psi_j = 0$, propose $\tilde{\psi}_j = 1$, and generate $\tilde{\theta}_j \sim N(0, 0.5)$ with all remaining components of θ unchanged. Let $\tilde{\psi}_i = \psi_i$ for $i \neq j$. This proposal is accepted with probability

   $$\min\left\{1, \frac{P_{EL}(\tilde{\theta}, \tilde{\psi}, r|D)}{P_{EL}(\theta, \psi, r|D)q(\tilde{\theta}_j)}\right\}.$$

2. (Swap) When $1 < |\tilde{\psi}| < 24$, randomly choose two indices $k, l \in \{2, ..., 24\}$, such that $\tilde{\psi}_k = 0$ and $\tilde{\psi}_l = 1$. Propose $\psi'_k = 1$ and $\psi'_l = 0$, and $\theta'_k \sim N(0, 0.5)$. This proposal is accepted with probability

   $$\min\left\{1, \frac{P_{EL}(\theta', \psi', r|D)q(\tilde{\theta}_l)}{P_{EL}(\tilde{\theta}, \tilde{\psi}, r|D)q(\theta'_k)}\right\}.$$

WITHIN Step Update θ' to θ^* with model indices fixed and with the nonzero values of θ' changed: For each index j such that $\psi'_j = 1$, generate $\theta^*_j \sim N(\theta'_j, 0.5)$. Generate $r^* \sim$ Uniform$[0, c]$ for some $c > 0$. Accept θ^* with probability

$$\min\left\{1, \frac{P_{EL}(\theta^*, \psi', r^*|D)}{P_{EL}(\theta', \psi', r|D)}\right\}.$$

The constant c in the WITHIN Step is chosen to be a threshold. We can simply set c to be an upper bound of the loss function. Alternatively, it can also be a level below which the classification error is expected to be controlled.

6.3 Implementation and results

The data set is randomly divided into two groups: training data (2/3) $\{(Y_i^t, X_i^t)\}_{i=1}^{n_t}$ and validation data (1/3) $\{(Y_i^v, X_i^v)\}_{i=1}^{n_v}$. To select a threshold r_0, a chain $\Theta_0 = \{\theta_i\}_{i=1}^{10,000}$ is first generated from the prior $\pi(\theta^\psi|\psi)\pi(\psi)$.[2] In order to achieve a low level of risk, r_0 is selected as the first percentile of the empirical risk based on the training data: let

$$\hat{R}^t(\theta) = \frac{1}{n_t}\sum_{i=1}^{n_t} \rho(Y_i^t, X_i^t; \theta).$$

Set r_0 to be the first percentile of $\{\hat{R}^t(\theta_i) : \theta_i \in \Theta_0\}$, which is 0.6882.

To generate samples from the posterior $P_{EL}(\theta, \psi, r|\text{training}, r \leq r_0)$, the MCMC algorithm described in the previous section is carried out for $20,000$ times based on the training data, with the first one fifth draws treated as the "burn-in period". We therefore obtain $B_1 = 16,000$ draws of the triplet: $(\Theta_1, \mathcal{R}_1, \Psi_1) = (\{\theta_i\}_{i=1}^{B_1}, \{r_i\}_{i=1}^{B_1}, \{\psi_i\}_{i=1}^{B_1})$. We then obtain a "good" action θ^* from the MCMC draws, and evaluate its performance on the validation data through the empirical risk

$$\hat{R}^v(\theta^*) = \frac{1}{n_v}\sum_{i=1}^{n_v} \rho(Y_i^v, X_i^v; \theta^*).$$

The action θ^* is obtained based on $(\Theta_1, \mathcal{R}_1, \Psi_1)$ via three different procedures as described below:

[2]Note that two chains are actually obtained, one for θ and the other for ψ. At this step we are only interested in the chain of θ.

(i) Posterior mean θ^* is the sample average of $\{\theta_i : \theta_i \in \Theta_1\}$, which also is an approximation of the posterior mean $E(\theta|\text{training}, r \leq r_0)$.

(ii) Further restricted threshold First determine r^* to be the 1st, 5th, and 10th percentiles of $\{r_i : r_i \in \mathcal{R}_1\}$ respectively. Define $\Theta_1^* = \{\theta_i \in \Theta_1 : r_i \leq r^*, i = 1, ..., B_1\}$. Let θ^* be the average of the elements in Θ_1^*, which is an approximation of $E(\theta|\text{training}, r \leq r^*)$. Therefore instead of averaging over all the draws in Θ_1, we average actions conditioning on further restricted threshold value r^*. This procedure gives us three different θ^*'s, corresponding to three chosen r^*'s.

(iii) Variable selection For each $\psi_i \in \Psi_1$, write $\psi_i = (1, \psi_{i2}, ..., \psi_{i,24})^T$. For each $j = 2, ..., 24$, let $P(j) = \frac{1}{B_1} \sum_{\psi_i \in \Psi_1} \psi_{ij}$, which calculates the sampling frequency of variable j in MCMC. Let $\psi_j^* = I(P(j) \geq k)$, for a pre-determined value $k \in [0.5, 1]$, and $\psi^* = (1, \psi_2^*, ..., \psi_{24}^*)$ as a new model index which includes all the selected variables. Hence the sampling frequency of each selected variable in ψ^* is at least k, which are identified as "important variables" by MCMC.

Now obtain a set of $B_2 = 10,000$ new MCMC draws $(\Theta_2, \mathcal{R}_2) = (\{\theta_i\}_{i=1}^{B_2}, \{r_i\}_{i=1}^{B_2})$ from $P_{EL}(\theta^{\psi^*}, r|\text{training}, \psi^*, r \leq r_0)$. Repeat either part (i) or part (ii) as described above. This procedure conducts variable selection before taking the posterior mean.

We compare our method with the classical ERM method (empirical risk minimization, e.g., Mohammadi and Van De Geer (2005)). The ERM action θ_0 is defined as $\theta_0 = \arg\min \hat{R}^t(\theta)$. The empirical risk on the validation data from ERM is $\hat{R}^v(\theta_0) = 0.6637$.

Table 2 summarizes $\hat{R}^v(\theta^*)$ on the validation data when θ^* is obtained via either

procedure (i) or (ii), as well as the actual values of r^* in (ii). The initial threshold $r_0 = 0.6882$, which is the first percentile of $\hat{R}^t(\theta)$ when θ is generated directly from the prior. The posterior mean $E(\theta|\text{training}, r \leq r_0)$ results in a smaller empirical risk on the validation data than that of ERM. In addition, the performance can be significantly improved by using $E(\theta|\text{training}, r \leq r^*)$ when r^* is further restricted to be the low percentiles of the MCMC draws $\mathcal{R}_1$.

Table 2: Empirical risk of θ^* based on validation data

	Posterior Mean	Further Restricted Threshold			
		1st	5th	10th	ERM
r^*	0.6882	0.6090	0.6285	0.6384	
$\hat{R}^v(\theta^*)$	0.6456	0.6156	0.6126	0.6186	0.6637

For the Posterior Mean, θ^* approximates $E(\theta|\text{training}, r \leq 0.6882)$; for the Further Restricted Threshold, θ^* approximates $E(\theta|\text{training}, r \leq r^*)$, where $r^* = 0.6090, 0.6285, 0.6384$ are the 1st, 5th, and 10th percentiles of MCMC draws $\mathcal{R}_1 = \{r_i\}_{i=1}^{B_1}$. Note that the empirical risk of ERM is 0.6637 on the validation data.

For the variable selection method described in part (iii) above, it is found that in addition to "Credit Duration", there are other three variables selected by all the draws in $(\Theta_1, \mathcal{R}_1, \Psi_1)$. We denote by M_1 as the model containing these four variables only (including Credit Duration). Meanwhile there are seven variables in total with $P(j) \geq 0.5$. The model containing these seven variables is denoted by M_2. We then generate a new set of MCMC draws $(\{\theta_i\}_{i=1}^{B_2}, \{r_i\}_{i=1}^{B_2})$ for $\psi^* = M_1$, and M_2 respectively, and obtain θ^* by the Posterior Mean and Further Restricted Threshold methods for both models. Table 3 summarizes the empirical risk on the validation data using θ^* obtained by variable selection.

When r^* is as low as the first percentile of the new MCMC draws $\mathcal{R}_2$, the posterior mean $E(\theta|\text{training}, M_1, r \leq r^*)$ performs the worst compared to alternative

Table 3: Empirical risk of θ^* after variable selection

		Posterior Mean	Further Restricted Threshold			
			1st	5th	7th	10th
M_1	r^*	0.6882	0.6253	0.6372	0.6407	0.6447
	$\hat{R}^v(\theta^*)$	0.6426	0.7147	0.6156	0.6216	0.6306
M_2	r^*	0.6882	0.6182	0.6378	0.6409	0.6471
	$\hat{R}^v(\theta^*)$	0.6547	0.6577	0.6306	0.6517	0.6667

For the Posterior Mean, θ^* approximates $E(\theta|\text{training}, \psi^*, r \leq 0.6882)$; for the Further Restricted Threshold, θ^* approximates $E(\theta|\text{training}, \psi^*, r \leq r^*)$, where ψ^* corresponds to M_1 and M_2 respectively, and r^* is determined as the 1st, 5th, 7th and 10th percentiles of MCMC draws $\mathcal{R}_2 = \{r_i\}_{i=1}^{B_2}$.

methods on the validation data ($\hat{R}^v = 0.7147$), because the chosen threshold value r^* (0.6253) is too low to be realistically achieved by model M_1. For M_2, when r^* is the first percentile, the empirical risk (0.6577) is also higher than those conditioned on the higher percentiles, which can be explained for the same reason: the threshold 0.6182 is probably not realistically achievable by M_2. On the other hand, when r^* is set to the 5th percentile, the posterior mean then performs well on the validation data ($\hat{R}^v = 0.6156$ for M_1 and 0.6306 for M_2.) As r^* is set to be higher percentiles, the empirical risk gradually increases. Note that all the empirical risks are either comparable to or better than the empirical risk of the ERM (0.6637), but simpler models with selected variables have better interpretability of the rationale for the credit decision in empirical applications. The selected variables in M_1 and M_2 are listed in Table 4.

We would like to emphasize that although our method is specially designed to provide a new framework to make robust inference on the risk and on the corresponding actions to take, however, this real data example demonstrates that it can

still perform comparably or favorably relative to the classical empirical risk minimization method, when it is indeed used for the purpose of risk minimization.

Table 4: List of selected variables in M_1 and M_2

Model	Variables	
M_1	Duration of Credit	Credit Amount
	Property	Age
M_2	Duration of Credit	Credit Amount
	Liable to provide maintenance	Age
	Sex and Marriage Status	Property
	Housing	

7 Conclusion

We considered an approximate joint posterior inference on the action and the associated risk in the classification problem. We introduced a prior distribution on an auxiliary parameter θ. Unlike a standard parameter, θ is not a functional of the data generating process, but only indexes a decision rule to be studied by the user. Our contribution is to provide a new framework for the posterior relationship between the risk and actions. This framework is useful in making robust joint inference of (θ, r) without knowing the distribution of the data generating process.

The posterior distribution is based on an empirical likelihood, which imposes a moment restriction relating the action to the resulting risk, but does not otherwise require a probability model for the underlying data generating process. As there is no need to assume the true likelihood function, such an EL-posterior approach based on a moment-condition likelihood is robust and computationally efficient. It

has been shown that this procedure works well when the sample size is large, since the empirical likelihood can be interpreted as the approximation to the true underlying likelihood function asymptotically. We show in the appendix that in the binary classification problem with the absolute loss function, the EL-posterior is equivalent to Schennach (2005)'s BETEL, therefore Schennach's argument also provides a distributional interpretation for the EL-posterior in our paper.

An important feature of our approach is that the parameters (θ, r) are not fully identified. The posterior density therefore does not degenerate to a point probability mass, but asymptotically concentrates around the curve $\{(\theta, r) : E[\rho(W, \theta)] = r\}$, as the functional relationship of (θ, r) is identified. Therefore we can generate the desired action θ from $P_{EL}(\theta | r < r_0, D)$, given a risk r_0 that the decision maker can tolerate. We illustrated by examples how this method is used to describe the EL-posterior of the actions to take in order to achieve a low risk, or conversely, to describe the posterior of the resulting risk for a given action. In addition, this approach can also be applied to variable selection.

A Appendix

A.1 Proof of Theorem 3.1

Proof. Define $L(\mu) = \sum_{i=1}^{n} \log[1+\mu(|Y_i - C(X_i, \theta)| - r)]$. Let n_1 be the number of i such that $|Y_i - C(X_i, \theta)| = 0$, and n_2 be the number of i such that $|Y_i - C(X_i, \theta)| = 1$. Note that $\hat{R} = n^{-1} \sum_{i=1}^{n} |Y_i - C(X_i, \theta)| = n^{-1} n_2$, and $n_1 + n_2 = n$. Since $|Y_i - C(X_i, \theta)| \in \{0, 1\}$, we have

$$\begin{aligned} L(\mu) &= \sum_{i:|Y_i - C(X_i,\theta)|=0} \log(1 - \mu r) + \sum_{i:|Y_i - C(X_i,\theta)|=1} \log(1 + \mu(1 - r)) \\ &= n_1 \log(1 - \mu r) + n_2 \log(1 + \mu(1 - r)) \end{aligned}$$

Differentiating $L(\mu)$ and setting to zero, it is straightforward to verify that the optimal $\mu^* = \frac{n_2 - nr}{nr(1-r)}$, with $\max_\mu L(\mu) = (n - n_2)\log\frac{n-n_2}{n(1-r)} + n_2 \log\frac{n_2}{nr}$. Replacing n_2 by $n\hat{R}$ yields $\max_\mu L(\mu) = nK(\hat{R}, r)$. Finally, we verify that the second derivative $L''(\mu^*) = -\frac{n^3(1-r)^2r^2}{n_2 n_1} < 0$ when $r \in (0, 1)$. Q.E.D.

A.2 Proof of Theorem 3.2

The following lemma establishes some relationships between the expression in the log-empirical likelihood and the square (or absolute value) distances.

Lemma A.1. *For $p, q \in [0, 1]$, and $K(p, q)$ as defined in (3.2),*

$$0.5(q - p)^2 \le K(p, q) \le \{\min(p, q, 1 - p, 1 - q)\}^{-2} 0.5(p - q)^2. \tag{A.1}$$

Proof. This is straightforward by a second order Taylor expansion of $-\ln(1 + \delta_{1,2})$, where $\delta_1 = q/p - 1$ and $\delta_2 = (1 - q)/(1 - p) - 1$. Q.E.D.

Proof of Theorem 3.2

Denote $R = E^*\rho(W, \theta)$ and $\Delta = \sup_\theta |\hat{R} - R|$, then

$P_{EL}[|R-r| > \epsilon|D] \le \int I(|R-r| > \epsilon) I(\Delta \le \epsilon/2) e^{-nK(\hat{R},r)} d\pi / \int e^{-nK(\hat{R},r)} d\pi +$ $I(\Delta > \epsilon/2)$. The numerator of the first term is less than $e^{-n\epsilon^2/8}$ since $|\hat{R} - r| \ge |R - r| - \Delta > \epsilon/2$ and this implies $K(\hat{R}, r) > \epsilon^2/8$ due to a previous lemma.

The denominator is bounded by

$\int e^{-nK(\hat{R},r)} d\pi \ge \int I(|R - r| \le \delta) I(\Delta \le \delta/2) I(\eta \ge \tau) e^{-nK(\hat{R},r)} d\pi$

$\ge e^{-n(\tau - \delta/2)^{-2}(9/8)\delta^2} \pi(|R - r| \le \delta, \eta \ge \tau) I(\Delta \le \delta/2)$, where $\eta = \min(R, 1 - R, r, 1 - r)$ and δ and τ are some positive constants. Here we used again a previous lemma to bound $K(\hat{R}, r) \le \{\min(\hat{R}, 1 - \hat{R}, r, 1 - r)\}^{-2} 0.5(\hat{R} - r)^2 \le (\tau - \delta/2)^{-2} 0.5(\hat{R} - r)^2 \le (\tau - \delta/2)^{-2} 0.5(\delta + \delta/2)^2$.

Combining these we obtain: the event $\Delta \leq \min\{\delta/2, \epsilon/2\}$ implies the event

$$P_{EL}(|R-r| > \epsilon | D) \leq \frac{e^{-n\epsilon^2/8+(9n/8)(\delta/(\tau-\delta/2))^2}}{\pi(|R-r| \leq \delta, \eta \geq \tau)}.$$

Note that $\pi(|R-r| \leq \delta, \eta \geq \tau) > 0$ by assumption. Choose constants τ and δ suitably, then the right hand side can be made arbitrarily close to zero (and exponentially small in n). This happens with P_W^n, the probability in D being at least $P_W^n(\Delta \leq \min\{\delta/2, \epsilon/2\})$, which converges to 1 by assumption. Q.E.D.

A.3 Equivalence between Schennach's BETEL and EL-Posterior

Suppose we observe i.i.d. data of $D = (W_1, ..., W_n)$. It is assumed that some unknown parameter (θ, r) satisfies moment condition:

$$E(\rho(W, \theta)|\theta) = r.$$

Schennach (2005) proposed a nonparametric Bayesian procedure based on the above moment condition to derive a moment-condition-based posterior, known as "Bayesian exponentially tilted empirical likelihood posterior" (BETEL):

$$P(\theta, r|D) \propto \pi(\theta, r) \prod_{i=1}^{n} w_i(\theta, r)$$

where $(w_1(\theta, r), ..., w_n(\theta, r))$ are the solutions to

$$\max_{w_1,...,w_n} \{\sum_{i=1}^{n} -w_i \log w_i | \sum_i w_i = 1, \sum_i w_i(\rho(W_i, \theta) - r) = 0\}.$$

Theorem A.1. *Consider the classification problem* $\rho = |Y - C(X, \theta)|$, $Y, C(X, \theta) \in \{0, 1\}$ *and* $r \in [0, 1]$. *Define the empirical risk* $\hat{R}(\theta) = n^{-1} \sum_{i=1}^{n} |Y_i - C(X_i, \theta)|$. *If* $\hat{R}(\theta) \in (0, 1)$*, then Schennach (2005)'s BETEL is equivalent to the EL-posterior defined in Theorem 3.1.*

Lemma A.2. *The interior of the convex hull of* $\bigcup_{i=1}^{n}\{\rho(W_i,\theta)-r\}$ *contains the origin if and only if* $r \in (0,1)$ *and* $\hat{R}(\theta) \in (0,1)$.

Proof. First of all, note that $\hat{R}(\theta) \in (0,1)$ if and only if $\min_{i\leq n}\rho(W_i,\theta) = 0$ and $\max_{i\leq n}\rho(W_i,\theta) = 1$.

As $\rho(W,\theta) = |Y - C(X,\theta)|$ is one-dimensional, the convex hull of $\bigcup_{i=1}^{n}\{\rho(W_i,\theta) - r\}$ is an interval $[\min_i \rho(W_i,\theta) - r, \max_i \rho(W_i,\theta) - r]$, whose interior is

$$(\min_i |Y_i - C(X_i,\theta)| - r, \max_i |Y_i - C(X_i,\theta)| - r)$$

When $r \in (0,1)$, as long as there exists a correctly classified (Y_i, X_i) and at least an incorrectly classified (Y_i, X_i), $\min_i |Y_i - C(X_i,\theta)| = 0$, and $\max_i |Y_i - C(X_i,\theta)| = 1$. Hence $0 \in (\min_i |Y_i - C(X_i,\theta)| - r, \max_i |Y_i - C(X_i,\theta)| - r)$. This proves the sufficiency.

On the other hand, if either $\max_i |Y_i - C(X_i,\theta)| = 0$ or $\min_i |Y_i - C(X_i,\theta)| = 1$, $0 \notin (\min_i |Y_i - C(X_i,\theta)| - r, \max_i |Y_i - C(X_i,\theta)| - r)$. If $r = \{0,1\}$, clearly $0 \notin (\min_i |Y_i - C(X_i,\theta)| - r, \max_i |Y_i - C(X_i,\theta)| - r)$ either. This proves the necessity.

Proof of Theorem A.1

Proof. The proof is based on Corollary 1 of Schennach (2005). When $r \in (0,1)$ and $\min_i |Y_i - C(X_i,\theta)| = 0, \max_i |Y_i - C(X_i,\theta)| = 1$ (which implies $\hat{R}(\theta) \in (0,1)$), by the previous lemma, the interior of the convex hull of $\bigcup_{i=1}^{n}\{\rho(W_i,\theta) - r\}$ contains the origin. By Corollary 1 and expressions (9), (10) in Schennach (2005), $P(\theta, r|D) \propto \lim_{m\to\infty}\pi(\theta,r)\int P(D|\xi^m)P(\xi^m|\theta,r)d\xi^m$ can be obtained by

$$P(\theta, r|D) \propto \pi(\theta,r)\prod_{i=1}^{n} w_i(\theta,r)$$

where

$$w_i(\theta, r) = \frac{\exp(\lambda(\theta, r)[\rho(W_i, \theta) - r])}{\sum_{i=1}^{n} \exp(\lambda(\theta, r)[\rho(W_j, \theta) - r])}$$

$$\lambda(\theta, r) = \arg\min_t \sum_{i=1}^{n} \exp(t[\rho(W_i, \theta) - r])$$

We can immediately obtain: when $\hat{R}(\theta) \in (0, 1)$, $\lambda(\theta, r) = \log((n_1 r)/(n_2(1 - r)))$ where $n_1 + n_2 = n$, $n_2 = n\hat{R}(\theta)$. Then $\prod_i w_i(\theta, r) = \exp(-nK(\hat{R}(\theta), r))n^{-n}$, where $K(,)$ is given by (3.2). If $r \in \{0, 1\}$ and $\hat{R}(\theta) \in (0, 1)$, 0 is not in the interior of convex hull of $\bigcup_{i=1}^{n}\{\rho(W_i, \theta) - r\}$. By Corollary 1 of Schennach (2005), $P(\theta, r|D) = 0$. In this case, from (3.2), $\exp(-nK(\hat{R}(\theta), r)) = 0$. Hence we conclude that when $R \in (0, 1)$, $P(\theta, r) \propto \pi(\theta, r) \exp(-nK(\hat{R}(\theta), r))$. By Theorem 3.1, $P(\theta, r|D) = P_{EL}(\theta, r|D)$. Q.E.D.

References

[1] ASUNCION, A. and NEWMAN, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[2] CHAUDHURI, S. and GHOSH, M. (2010) Empirical likelihood for small area estimation. To appear in *Biometrica*.

[3] CHEN, K., JIANG, W. and TANNER, M. (2010). A note on some algorithms for the Gibbs posterior. *Statistics and Probability Letters*. **80** 1234-1241

[4] CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics.* **115** 293-346

[5] CHERNOZHUKOV, V., HONG H. and TAMER E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*. **75** 1243-1284

[6] DEVROYE, G., GYORFI, L and LUGOSI, G (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

[7] GUSTAFSON, P. (2010). Bayesian Inference for Partially Identified Models. *The International Journal of Biostatistics*, 6, Iss. 2, Article 17.

[8] HOROWITZ, J. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica*. **60** 505-531.

[9] KIM, J. (2002). Limited information likelihood and Bayesian analysis. *Journal of Econometrics*. **107** 175-193.

[10] KITAMURA, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*. **69** 1661-1672.

[11] LAZAR, N. (2003). Bayesian empirical likelihood. *Biometrika*. **90** 319-326.

[12] LIAO, Y. and JIANG, W. (2010). Bayesian analysis of moment inequality models. *The Annals of Statistics*. **38** 275-316.

[13] MANSKI, C. (2007). *Identification for Prediction and Decision.* Harvard University Press.

[14] MONAHAN, J. and BOOS, D. (1992) . Proper likelihoods for Bayesian analysis. *Biometrika*, **79**, 271278.

[15] MOHAMMADI, L. and VAN DE GEER, S. (2005) Asymptotics in empirical risk minimization. *Journal of Machine Learning Research*. **6** 2027-2047.

[16] MOON, H. and SCHORFHEIDE, F. (2010). Bayesian and Frequentist Inference in Partially Identified Models. *Manuscript*. University of Southern California.

[17] OWEN, A. (1990). Empirical likelihood for confidence regions. *The Annals of Statistics*. **18** 90-120.

[18] POIRIER, D. (1998). Revising beliefs in nonidentified models. *Econometric Theory*. **14** 483-509.

[19] QIN, J. and LAWLESS, J. (1994). Empirical Likelihood and general estimating equations. *The Annals of Statistics*. **22** 300-325.

[20] RAGUSA, G. (2007). Bayesian Likelihoods for Moment Condition Models. *Manuscript*, University of California, Irvine.

[21] RAO, N. and WU, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of Royal Statistical Society*, Ser. B. **72** 533-544

[22] SCHENNACH, S. (2005). Bayesian exponentially tilted empirical likelihood *Biometrika*. **92** 31-46.

[23] TAMER, E. (2010). Partial identification in econometrics. *Annual Review of Economics*. **2**, 167-195.

[24] VAPNIK, V. and CHERVONENKIS, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of probability and its applications.* **16** 264-280

[25] WEST, D. (2000). Neural network credit scoring models. *Computers and Operations Research*. **27** 1131-1152.