Econometric Theory, **26**, 2010, 1437–1452.
doi:10.1017/S0266466609990636

RISK MINIMIZATION FOR TIME SERIES BINARY CHOICE WITH VARIABLE SELECTION

WENXIN JIANG AND MARTIN A. TANNER
Northwestern University

This paper considers the problem of predicting binary choices by selecting from a possibly large set of candidate explanatory variables, which can include both exogenous variables and lagged dependent variables. We consider risk minimization with the risk function being the predictive classification error. We study the convergence rates of empirical risk minimization in both the frequentist and Bayesian approaches. The Bayesian treatment uses a Gibbs posterior constructed directly from the empirical risk instead of using the usual likelihood-based posterior. Therefore these approaches do not require a correctly specified probability model. We show that the proposed methods have near optimal performance relative to a class of linear classification rules with selected variables. Such results in classification are obtained in a framework of dependent data with strong mixing.

1. INTRODUCTION

Consider a stationary econometric time series (y_t, x_t), where $y_t \in \{0, 1\}$ is the dependent variable of binary choices and $x_t \in \Re^K$ is the vector of exogenous explanatory variables. A dynamic binary choice model allows modeling of y_t by the exogenous explanatory variables as well as the lagged dependent variables, for example, by

$$y_t = I[b'z_t + u_t > 0], \tag{1}$$

where u_t is an unobserved random variable and $b'z_t$ is a linear combination of some selected components of $z_t' = (1, x_t', y_{t-1}, y_{t-2}, \ldots, y_{t-J})$.

The definition of such a model and its fundamental properties have been studied recently by de Jong and Woutersen (2007), who also discuss parameter inference with maximum likelihood and maximum score methods. The maximum score methods (Manski, 1975, 1985; and the smoothed version of Horowitz, 1992), although originally used to estimate the coefficients, b, can also be used to construct natural predictions of y_t of the form $I[b'z_t > 0]$, which is intimately related to the

We are grateful to the co-editor Jinyong Hahn and two anonymous referees for their very helpful comments. This work was supported in part by NSF Grant DMS-0706885. Address correspondence to Wenxin Jiang, Department of Statistics, Northwestern University, Evanston, IL 60208, U.S.A.; e-mail: wjiang@northwestern.edu.

© Cambridge University Press 2010 0266-4666/10 $15.00

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

method of empirical risk minimization (see, e.g., Devroye, Györfi, and Lugosi, 1996, Sect. 12.1).

The current paper studies performance of such a resulting prediction rule $I[b'z_t > 0]$ in predicting y_t, which complements the paper by de Jong and Woutersen (2007) on parameter inference. We show that even though this rule can be motivated from the dynamic binary choice model (1) on (y_t, x_t, u_t), the resulting predictive performance remains "relatively optimal" (to be specified later) even without making such model assumptions. Our results on predictive performance therefore will *not* depend on the validity of (1). We will also allow potentially high dimensionality of the predictor z_t in this paper and consider variable selection handled by nonzero components of the parameter vector b.

Recently the problem of high dimensional prediction has been of much interest in statistics and machine learning (see, e.g., Greenshtein, 2006; Jiang and Tanner, 2008a), where data are often assumed to be i.i.d. (independent and identically distributed). Advances in these fields have a great potential to be extended and applied to treat econometric problems with dependent data. The current paper is an extension of Jiang and Tanner (2008a) to a non-i.i.d. environment, which will be described in Section 3. We here first formulate the type of results that we will obtain on risk performance and convergence rates.

2. FORMULATION AND RISK PERFORMANCE

We consider classification of y_t based on $z_t' = (1, x_t', y_{t-1}, y_{t-2}, \ldots, y_{t-J})$ using linear classification rules of the form $I[b'z_t > 0]$ (where certain components of z_t are selected by the nonzero b-coefficients). The coefficient vector b is to be obtained somehow based on observed data $D = (y_t, x_t)_{t=1}^N$. We assume that the observed data D are generated by some probability mechanism p^* (which is otherwise not explicitly modeled). Typically, for the current problem of predicting binary dependent variables, the theoretical risk function of interest is

$$R(b) = \mathrm{E}^*|y_t - I[b'z_t > 0]|,$$

the expected classification error, which is to be minimized in some sense. (The symbol E^* represents expectation with respect to the underlying true probability mechanism p^*.) In practice, we will choose b to minimize an empirical risk R_N somehow, where

$$R_N(b) = (N-J)^{-1} \sum_{t=J+1}^{N} |y_t - I[b'z_t > 0]|. \tag{2}$$

Note that R_N is a sample counterpart of the theoretical risk R. We will consider convergence properties of R_N to R in a high dimensional situation, where $\dim(z_t)$ can be very high and increase with N.

This setup can be related to the principle of "empirical risk minimization" discussed in the statistics literature; see, e.g., Devroye et al. (1996, Sect. 12.1).

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

The goal of minimizing $R_N(b)$ is to achieve a small theoretical risk $R(b)$ among a "class" of models $b \in H$ (to be described later), so that $R(b) - \inf_{b \in H} R(b)$ is small. Since the good risk performance is relative to a "class" of classification rules H, this description forms an analog of the "persistence" concept described in Greenshtein (2006) in an i.i.d. context. We here allow time series data that have a weak dependence structure to be described in the next section.

We will describe both frequentist and Bayesian approaches of minimization of R_N. In a frequentist approach (in Section 4) we will consider the risk performance of a minimizer $\hat{b} = \arg\min R_N(b)$. Minimizing this empirical risk is equivalent to maximum score of Manski (1975, 1985) (although we here focus on the risk behavior $R(\hat{b})$ instead of on statistical inference). The minimization here is subject to a constraint on the maximal number of nonzero b-coefficients. Alternatively, a penalized minimization can be used, which minimizes R_N plus a term of complexity penalty. In a Bayesian approach, we will use a 'Gibbs posterior' on $b \in H$ (described in Section 5) to obtain small empirical risk $R_N(b)$ with high posterior probability.

When we consider classification of y_t based on $z_t' = (1, x_t', y_{t-1}, y_{t-2}, \ldots, y_{t-J})$, we will allow a high dimensional situation where $K = \dim(x_t)(= K_N)$ can grow polynomially as N^a for any $a > 0$ (so K can be much greater than N), and J (the number of lags allowed) can grow as $o(N^{1/2}/\ln N)$.

The class of rules H to be considered is a set of classification rules of the form $I[b' z_t > 0]$, where out of the $\tilde{K}_N = K_N + J_N + 1$ components of the candidate explanatory variables in z_t, only a smaller number, say, about $v = v_N$ (to be specified later), are selected, by setting many b-coefficients to be zero. We can also write H as a set of parameters b (which determines the classification rules). We will refer to H as a *scope*.

The results we will show are of a frequentist nature. When b is obtained from data D by empirical risk minimization in either frequentist or Bayesian approaches to be described in later sections, we will show that the theoretical classification risk $R(b)$ often will be "good" or "competitive," relative to the optimal $\inf_{b \in H} R(b)$ over a class of classification rules H, i.e.,

$$R(b) \leq \inf_{b \in H} R(b) + O_P(\delta_N), \tag{3}$$

when D is generated according to the underlying true probability mechanism $p^*(D)$. Here δ_N is a convergence rate that is related to the scope of comparison H. With a scope $H = H(v)$ where at most v variables are selected, we will show a convergence rate of order

$$\delta_N = N^{-1/2}(J+1)v \ln N. \tag{4}$$

This rate, in the special case of finite dimensionality (with J, v being constant in N), is very close (within some $\ln N$ factors) to the classical rate of empirical risk minimization $O(\sqrt{\ln N / N})$ (see, e.g., Devroye et al., 1996, Ch. 12), as well as to the optimal rate $O(1/\sqrt{N})$ (see, e.g., Devroye et al., 1996, Ch. 14). The rate

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

δ_N remains close (within some $\ln N$ factors) to $O(1/\sqrt{N})$ even when J (the number of candidate lags) and v (the maximal number of selected variables) grow as any power of $\ln N$. Note that here we can allow the number of candidate exogenous variables K to grow as any power of N. (The power turns out to affect only the coefficient of δ_N and not the rate.)

3. NOTATION AND ASSUMPTIONS

Note that the classification rule $I[b'z_t > 0]$ is unchanged with scale changes of b. We use the same normalization as Horowitz (1992). The normalization involves letting one of the predicting variables have a sign-valued coefficient in the classification rule. Let $z_t' = (1, x_t', y_{t-1}, y_{t-2}, \ldots, y_{t-J})$ and $x_t' = (x_{1t}, x_{2t}')$ with a scalar component x_{1t}. The b-coefficient can be decomposed in a similar way as $b' = (\alpha, \gamma', \rho_1, \ldots, \rho_J)$ with $\gamma' = (\gamma_1, \gamma_2')$. The normalization involves restricting $|\gamma_1| = 1$. Denote this restricted parameter space as $\Omega = \Re \times \{\pm 1\} \times \Re^{K-1+J}$. We will consider a compact subset of this parameter space $\bar{\Omega} = \{b \in \Omega, |b|_\infty \le C\}$ for some constant $C \ge 1$. In variable selection, we can also write the set of classification rules H as a set of normalized parameters b (which determines the classification rules) such that $H \subset \Omega$. (This implies that $|\gamma_1| = 1$ and x_{1t} is always a selected explanatory variable.) For example, we will consider the following scope $H = H(v)$, which is a subset of the compact parameter space $\bar{\Omega}$ and represents linear classifiers with at most v selected variables:

$$H(v) = \{b \in \bar{\Omega} : \ |b|_0 \le v\}.$$

Regarding notation used in the remainder of this paper: For a generic vector u, denote $|u|_0 = \sum_{j=1}^{\dim(u)} I(|u_j| > 0)$, $|u|_k = (\sum_{j=1}^{\dim(u)} |u_j|^k)^{1/k}$ for $k = 1, 2, \ldots$, and $|u|_\infty = \sup_{j=1}^{\dim(u)} |u_j|$. For a set S, denote $|S| = \mathrm{card}(S)$. For a real number u, define $\lceil u \rceil$ as its integer part. Constants are denoted as C, c, c_1, etc., and can take different values at different places in the text.

We now describe assumptions on the true data generating mechanism for the observed process (y_t, x_t). We here allow time series data that have a weak dependence structure of exponential strong mixing.

Assumption 1. Suppose that (y_t, x_t) is strictly stationary and has exponentially decaying strong (α-) mixing coefficients $\alpha(k) = O(s^k)$ for some $s \in (0, 1)$.

This assumption of exponential strong mixing is commonly made in the literature on predictive performance with dependent data (see, e.g., Modha and Masry, 1996; Zou and Li, 2007; and Jiang and Tanner, 2008b, for a slightly more general form). For linear processes such as the ARMA processes, exponential strong mixing occurs when there exists an invertible MA(∞) representation with exponentially decaying coefficients and with the marginal distributions of the increments satisfying a smoothness condition (see, e.g., results summarized in Davidson, 1994, Sect. 14.4). In addition, much recent research involves

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

proving models (possibly nonlinear) to be geometrically ergodic Markov processes, which then implies exponential beta mixing and consequently also exponential strong mixing. This line of research covers various GARCH (with thresholds and transforms), ACD, and stochastic volatility models. See, e.g., Carrasco and Chen (2002), Francq and Zakoïan (2006), and, more recently, Meitz and Saikkonen (2008).

The next assumption addresses the conditional distribution of the dependent variable y_t.

Assumption 2. Suppose that $P^*[y_t = 1|x_t, (y_s, x_s)_{-\infty}^{t-1}] \in (0, 1)$.

Here we assume that it is impossible to perfectly predict y_t based on $x_t, (y_s, x_s)_{-\infty}^{t-1}$. Such assumption is satisfied, e.g., if y_t is a binary choice of the form $y_t = I[g(x_t, (y_s, x_s)_{-\infty}^{t-1}) + u_t > 0]$, where u_t's are independent and have a positive density on $\Re$, such as $N(0, 1)$.

The next two assumptions address the distribution of the exogenous process x_t. These assumptions are useful in obtaining certain continuity properties of the risk function that will facilitate the proofs. Variable x_{1t} corresponds to the sign-valued coefficient in the linear classification rule. We first describe an assumption on its conditional density given x_{2t} (the remaining exogenous variables) and then describe an assumption on the remaining exogenous variables x_{2t}.

Assumption 3. The conditional density $p^*(x_{1t}|x_{2t}, (y_s, x_s)_{-\infty}^{t-1})$, with respect to the Lebesgue measure exists, is bounded above by a finite constant S and is positive, for all $x_{1t} \in \Re$ and for all possible values of the conditioned variables, for all t.

This assumption essentially asks for a "continuous" component x_{1t}, which can have, for example, a normal conditional density with variance bounded away from zero.

Assumption 4. All components of x_{2t} are standardized to be within $[-1, +1]$.

This last assumption essentially asks for components of x_{2t} to be bounded in certain ways. It can be relaxed to include some unbounded variables with first moments, e.g., by requiring $\mathrm{E}^*|x_{2t}|_\infty$ to be finite. We keep this assumption in current form for simplicity.

The following two assumptions demonstrate the potentially high dimensionality allowed in this paper by characterizing the allowed number of lags J_N and the number of candidate exogenous variables K_N.

Assumption 5. Suppose $J_N \geq 1$ and $J_N = o(N^{1/2}/\ln N)$.

Assumption 6. Suppose $1 \leq K_N \leq N^a$ for some $a > 0$, for all large enough N.

Therefore this assumption allows a high dimensional case, with K being much greater than N (with more candidate variables than the data size).

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

Comparing the two assumptions above, we note that the number of candidate lags J allowed (which is less than $\sqrt{N}/\ln N$) is much less than the number of candidate exogenous variables K allowed (which can be any power of N). Roughly speaking, this is because with too many lags, the mixing coefficients would not be decaying fast enough for the sample means to converge to the expectations properly.

With variable selection, we will obtain a convergence rate δ_N as in (4), which depends on υ, the maximal number of *selected* variables (out of the $K_N + J_N + 1$ candidate variables). This rate δ_N will be small ($\delta_N = o(1)$) if we restrict the number of selected variables as in the following assumption:

Assumption 7. We consider a scope $H = H(\upsilon) = \{b \in \bar{\Omega} : \ |b|_0 \le \upsilon\}$, where the maximal number of selected variables is υ ($1 \le \upsilon \le \tilde{K} = K + J + 1$). Assume that υ satisfies

$$\upsilon = o\left((J+1)^{-1} N^{1/2}/\ln N\right). \tag{5}$$

4. FREQUENTIST EMPIRICAL RISK MINIMIZATION

In a frequentist approach we will first consider the risk performance of a minimizer $\hat{b} = \arg\min R_N(b)$. (This is the maximum score estimator; see Manski, 1975, 1985.) The minimization is subject to the constraint $|b|_0 \le \upsilon$.

THEOREM 1 (Performance of constrained minimization). *Suppose Assumptions 1, 2, 3, 4, 5, and 6 hold. Consider a minimizer* $\hat{b} = \arg\min_{b \in H} R_N(b)$ *where* $H = H(\upsilon)$ *satisfies Assumption 7. For all* $\zeta > 0$*, there exists a constant* $c_\zeta > 0$ *such that, for all large enough* N*, we have*

$$P_D^*[R(\hat{b}) - \inf_{b \in H(\upsilon)} R(b) > 2c_\zeta N^{-1/2}(J+1)\upsilon \ln N] \le N^{-\zeta\upsilon}. \tag{6}$$

Note that this implies almost surely

$$R(\hat{b}) - \inf_{b \in H(\upsilon)} R(b) = O(N^{-1/2}(J+1)\upsilon \ln N)$$

by the Borel-Cantelli lemma. (The sum $\sum N^{-\zeta\upsilon}$ converges because $\upsilon \ge 1$ (as specified in Assumption 7) and $\zeta > 0$ is arbitrary and can be taken to be 2.) The convergence rate is therefore of the form (4), which can be close to the optimal rate $O(1/\sqrt{N})$, even if the number of candidate exogenous variables K can be much larger than N according to Assumption 6.

Although we claim a good risk performance in the above theorem, the minimization is quite difficult to carry out due to the complicated constraint of the set $H = \{b \in \bar{\Omega} : \ |b|_0 \le \upsilon\}$ and the discontinuous R_N. One may alternatively consider unconstrained minimization of R_N plus a complexity penalty, which is

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

constructed from $|b|_0$. This is similar to the structural risk minimization (see, e.g., Devroye et al., 1996, Ch. 18).

Define $\tilde{R}_N(b) = R_N(b) + \lambda|b|_0$, where λ is a regularization parameter. Define $\tilde{b} = \arg\min_{b\in\bar{\Omega}} \tilde{R}_N$. Although there is no constraint on $|b|_0$, the penalty term makes sure that any minimizer will have

$$|\tilde{b}|_0 \leq v_0 = \min\{(1 + 1/\lambda), K + J + 1\}. \tag{7}$$

(The upper bound of $K + J + 1$ is from the dimension of the parameter space $\bar{\Omega}$. The upper bound of $(1 + 1/\lambda)$ is obtained from demanding $\tilde{R}_N(\tilde{b}) \leq \tilde{R}_N(b_1)$, where b_1 is a parameter with $|b_1|_0 = 1$, which implies $\lambda|\tilde{b}|_0 \leq R_N(\tilde{b}) + \lambda|\tilde{b}|_0 \leq R_N(b_1) + \lambda|b_1|_0 \leq 1 + \lambda$.)

The following theorem describes the risk performance of penalized minimization, which will also have a convergence rate of the form (4).

THEOREM 2 (Performance of penalized minimization). *Suppose Assumptions 1, 2, 3, 4, 5, and 6 hold. Define* $\tilde{R}_N(b) = R_N(b) + \lambda|b|_0$ *and* $\tilde{b} = \arg\min_{b\in\bar{\Omega}} \tilde{R}_N(b)$. *There exists a regularization parameter*

$$\lambda = cN^{-1/2}(J + 1)\ln N \tag{8}$$

for some constant $c > 0$ *such that the corresponding* $\tilde{b}$ *has*

$$R(\tilde{b}) - \inf_{b\in H(v)} R(b) = O(N^{-1/2} v (J + 1)\ln N)$$

almost surely. This holds for any scope $H(v)$ *satisfying Assumption 7.*

Instead of penalizing on $|b|_0$, we conjecture that it is also possible to use a complexity penalty constructed from $|b|_1$ (see, e.g., Greenshtein, 2006, for the i.i.d. case). In the following section we consider a Bayesian-flavor treatment for empiricial risk minimization, which can allow the use of Markov chain Monte Carlo methods (such as the Metropolis algorithm; see, e.g., Tanner, 1996) for computation.

5. EMPIRICAL RISK MINIMIZATION WITH GIBBS POSTERIOR

In a Bayesian approach of empirical risk minimization, we will consider choosing b randomly from a Gibbs posterior $q(b|D) \propto e^{-(N-J)\psi R_N(b)} p(b)$, where $p(b)$ is a prior distribution and $1/\psi > 0$ is a constant called the temperature. (The limit of low temperature corresponds to frequentist minimization; see, e.g., Geman and Geman, 1984.) This approach is discussed in Zhang (2006) in the i.i.d. case. (Zhang (1999) considers choice of the temperature parameter using cross validation.)

In the more standard Bayesian approach, b would be chosen from a posterior distribution $p(b|D) \propto p(D|b)p(b)$, where $p(D|b)$ is a likelihood function that is

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

based on a probability model, such as the binary choice model $y_t = I(b'z_t + u_t > 0)$, where u_t's are independent $N(0,1)$. This method will have suboptimal classification performance when the probability model is misspecified, e.g., when there are higher order terms of x_t or interactions that are neglected by the binary choice model $y_t = I(b'z_t + u_t > 0)$. This is because the usual likelihood-based posterior will propose densities that are consistent for a minimizer of Kullback-Leibler difference (see, e.g., Kleijn and van der Vaart, 2006), rather than a minimizer of the classification risk, among the set of linear classifiers. Such suboptimality of the likelihood-based classification methods in the presence of model misspecification is well known. Some descriptions of this problem are seen in, e.g., Devroye et al. (1996, Sect. 15.2), and Jiang and Tanner (2008a). The Gibbs posterior used in this paper, on the other hand, will have better classification performance when the probability model is misspecified. However since the Gibbs posterior is based on the empirical risk instead of being based on the likelihood, one cannot use it for probabilistic inference. Despite its Bayesian flavor, the Gibbs posterior is mainly a random method for risk minimization, where the Bayesian flavor can be exploited to apply Markov chain Monte Carlo methods.

In Section 5.1 we will first specify assumptions on the prior distribution for parameter b, which will be denoted as P_b below. Then in Section 5.2 we will assess the classification performance of the resulting Gibbs posterior.

5.1. Prior Specification

In order that the Gibbs posterior generate classification rules that are competitive among a scope of rules, e.g., $H(\upsilon)$ with at most υ selected variables, it is natural to allow the prior to have non negligible chance to generate all such rules for comparison.

The following assumption, roughly speaking, requires the prior to have not-too-small probability in proposing any rule in a scope $H(\upsilon)$, so that all such rules can be easily proposed and compared to generate good rules in the Gibbs posterior. It will be used later when assessing the risk performance of the Gibbs posterior in Theorem 3.

Assumption 8 (Assumption on prior). The prior P_b is supported on $H(\upsilon)$ for some υ satisfying Assumption 7, and there exists a constant $c_1 > 0$ such that for all $b_0 \in H(\upsilon)$, $P_b[|b - b_0|_1 < c_1\delta_N] > e^{-c_1 N\delta_N}$ for all large enough N, where $\delta_N = N^{-1/2}(J+1)\upsilon \ln N$ (as in (4)).

For one example of such a prior, consider a uniform prior on $H = H(\upsilon)$ constructed this way. Split the components of b into parts b_2, $(b_j)_{j\in g}$, and $(b_j)_{j\notin g, j\neq 2}$ for some subset g not containing index 2. In the prior, assume that given g, independently, b_2 takes value from $\{-1, 1\}$ with equal probability, $(b_j)_{j\in g} \sim Uniform[-C, C]^{|g|}$ (this can be relaxed to a nonuniform distribution with density bounded from below in some way), and $(b_j)_{j\notin g, j\neq 2} = 0$.

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

For the prior on g, assume that the model dimension $|g|+1$ follows a distribution $f(m) = P_b[|g|+1 = m]$ on $m \in \{1, 2, 3, \ldots, \upsilon\}$ (note that $|b_2| = 1$ always), where υ is as in Assumption 7 ($1 \le \upsilon \le \tilde{K} = K + J + 1$, $\upsilon = o((J+1)^{-1} N^{1/2} / \ln N)$). Conditional on $|g| = m - 1$, assume that the elements of g are chosen randomly with probability $1 \Big/ \binom{K+J}{m-1}$ from all $K + J$ indexes $\{1, 3, \ldots, K + J + 1\}$.

Note that the above description specifies the entire prior P_b by specifying the priors on model dimension, on the model choice given the dimension, and on the parameters given the model choice. The latter two parts are fixed in a uniform scheme, and therefore Assumption 8 on the entire prior is completely specified by the assumption on $f(m)$, the prior on selected model dimension. The following assumptions and results show that a prior can be used for the selected model dimension in order to satisfy Assumption 8 on the entire prior.

Assumption 9 (Prior on selected model dimension). Regarding the prior distribution f on $\{1, 2, 3, \ldots, \upsilon\}$ for the selected model dimension, there exists a constant $c_1' > 0$ such that $f(\upsilon) \ge e^{-c_1' N \delta_N}$ for all large N. Here υ is as in Assumption 7 [$1 \le \upsilon \le \tilde{K} = K + J + 1$, $\upsilon = o((J+1)^{-1} N^{1/2} / \ln N)$].

The above assumption requires a not-too-small prior for selecting the maximal model dimension υ. Only the prior on the maximal dimension υ matters, and not $f(1), \ldots, f(\upsilon - 1)$, because a higher dimensional model (with more nonzero coefficients in the classification rule) can achieve a classification performance that is arbitrarily close to the best performance of a lower dimensional rule. We now discuss the relations between Assumptions 8 and 9.

PROPOSITION 1. *Assumption 9 is a necessary condition for Assumption 8 to hold for the entire prior.*

Assumption 9 is also a sufficient for Assumption 8 to hold for the entire prior under the previously described uniform scheme (for the priors on the model choice given the dimension and on the parameters given the model choice). This can be more generally true for nonuniform schemes that have certain prior probabilities bounded away from zero.

LEMMA 1 (Sufficient conditions for prior Assumption 8). *Suppose Assumptions 5 on J and 6 on K hold. Assume that*

(i) conditional on any selected g with maximal model dimension $|g|+1 = \upsilon$, the prior for the selected components $(b_j)_{j \in g \cup \{2\}}$ are mutually independent, and the probability density function (or probability function) of each selected b_j-component is bounded away from zero on its support by a common constant $u > 0$.

Assume also that

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

(ii) the prior of each model choice g given the choice of maximal model dimension $|g|+1=v$ differs by at most a constant factor, i.e.,

$$P[g=g_2\,||g|=v-1]/P[g=g_1\,||g|=v-1]\in[A^{-1},A], \tag{9}$$

for some constant $A>0$, for any two models g_1, g_2.

Then Assumption 9 on the model dimension implies Assumption 8 on the entire prior for b.

PROPOSITION 2. *When v is a constant of sample size N, Assumption 9 is simply equivalent to $f(v)>0$ on the maximal model dimension. When v is allowed to increase with sample size N, subject to Assumption 7, then Assumption 9 on f is satisfied if $f(m)$ is the probability function of a uniform distribution on $m\in\{1,2,3,\ldots,v\}$ or if $f(m-1)$ is the truncated geometric or truncated Poisson probability functions on $m-1\in\{0,1,2,3,\ldots,v-1\}$.*

Combining the above results, we know that Assumption 8 is very easily satisfied. For example, in the uniform scheme described above on the parameter and model choice conditional on the model dimension $|g|+1$, the prior on the model dimension can simply select the model dimension $|g|+1$ uniformly, or by truncating a geometric or Poisson random variable, in order to satisfy Assumption 8 on the entire prior.

It can be shown that Assumptions 9 and 8 will not be satisfied when the prior on $|g|+1$ has too thin a tail and when v grows with N too quickly. For example, when $|g|+1$ has a probability function $f(m)$ proportional to e^{-e^m}, $m=1,2,\ldots,v$, where v is such that $v^{-1}\ln N=o(1)$, the assumptions on the priors will fail. A more trivial counter-example for Assumption 8 can be, simply, when $f(v)=0$, so that $f(m)$ is supported on $\{1,2,\ldots,v-1\}$. In this case the resulting prior will always miss at least one variable and cannot lead to optimal classification performance relative to $H(v)$ in general.

5.2. Risk Performance of Gibbs Posterior

Now we assess the performance of classification rules that are randomly generated from the Gibbs posterior.

Suppose $b|D$ is generated according to the Gibbs posterior $q(b|D)\propto e^{-(N-J)\psi R_N(b)}P_b(b)$ as described before. We consider the risk performance when $(b,D)\sim P^*_{b,D}=q(b|D)P^*(D)$ (P^* represents the true distribution under which data D is generated).

THEOREM 3 (Performance of Gibbs posterior). *Suppose Assumptions 1, 2, 3, and 4 hold for the underlying time series. Suppose the following assumptions also hold: Assumption 5 on the number of lags J to select from, Assumption 6 on the number of exogenous variables K to select from, Assumption 7 on the class of rules $H=H(v)$, and Assumption 8 on the prior P_b. For all $\zeta>0$, there exists a*

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

constant $c_{\zeta 1} > 0$ *such that, for all large enough* N, *we have*

$$P^*_{b,D}[R(b) - \inf_{b \in H(\upsilon)} R(b) > 5c_{\zeta 1} N^{-1/2}(J+1)\upsilon \ln N] \leq N^{-\zeta \upsilon/2}. \quad \textbf{(10)}$$

Note that this implies almost surely

$$R(b) - \inf_{b \in H(\upsilon)} R(b) = O(N^{-1/2}(J+1)\upsilon \ln N)$$

by the Borel-Cantelli lemma. The resulting convergence rate is also of the form (4) when comparing to the best linear classifiers with variable selection, which is analogous to the near optimal rate $O(\sqrt{\ln N/N})$ (see, e.g., Devroye et al., 1996, Chs. 12 and 14) derived in classical situations of empirical risk minimization. In the classical situations i.i.d. data are considered for classification with finite dimensionality and with no variable selection. We, on the other hand, allow dependent data with exponential strong mixing and variable selection for classification in potentially high dimensions.

6. DISCUSSION

In this paper we have studied the convergence rates of both frequentist-style and Bayesian-style methods for empirical risk minimization, when the risk of interest is the classification error in time series of binary choice. The Bayesian-style treatment involves using a Gibbs posterior that is derived from the empirical risk and a prior. We use variable selection to handle a potentially high number of explanatory variables and lags and show that good risk performance can be achieved despite high dimensionality.

Several papers are related to the current work from different perspectives. Stock and Watson (2006) provide a comprehensive survey on forecasting with many predictors, which covers methods such as the dynamic factor models and Bayesian model selection, which all require parametric probability modeling. Our method, on the other hand, is based directly on the empirical risk of classification error and is more similar to Elliott and Lieli (2005), who study prediction of binary outcomes and actually allow more general utility functions. Elliott and Lieli (2005) focus on a finite dimensional setup. In contrast, our paper addresses high dimensional classification and variable selection and also discusses a Gibbs posterior approach for risk minimization. Recently, Lijoi, Prünster, and Walker (2007) considered Bayesian consistency in the style of Doob (1949), using the likelihood-based posterior. In contrast, our work studies convergence rates in the frequentist style and uses a nonlikelihood-based posterior that is more robust against model misspecification.

Several possible extensions may be made to the current work on dependent data:

1. It is noted that in this paper, for the Gibbs posterior approach, we have assumed a bounded support of the prior being the same as the scope of

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

comparison H. This may not be necessary, as Jiang and Tanner (2008a) show that one can also allow priors with unbounded support such as the normal priors for the selected parameter components in the i.i.d. case.

2. Here we consider a scope of comparison H that contains classification rules that only depend on a few (υ) out of possibly many ($K + J + 1$) candidate explanatory variables. More general scope H's can also be used, where H can include classifiers that depend on all candidate explanatory variables, where although only a few are important, the rest can have small and nonzero impact on classification.
3. Similar to Elliott and Lieli (2005), it is also possible to consider risk functions that are more general than the classification error, which can include risk functions of practical interest in data mining, that are related to dollar costs or profits related to marketing decisions.
4. Although we consider a Gibbs posterior constructed from the discrete sample classification error, a smoothed version can also be used, which is similar to the smoothed version (Horowitz, 1992) of the method of maximum score (Manski, 1975, 1985), which can be regarded as the frequentist counterpart of the Gibbs posterior approach in this paper.
5. Our paper considers empirical risk minimization with variable selection, which can be easily applied to obtain multistep predictions of a binary outcome in a direct procedure (as described in, e.g., Stock and Watson, 2004). Our current formalism, however, does not lead to a straightforward method to do iterative multistep prediction, since we do not make a parametric assumption on $p^*(y_t = 1|z_t)$, which is needed to derive iterative forecasts. It will be an interesting future work to make additional assumptions on $p^*(y_t = 1|z_t)$ so that it can be consistently estimated in the high dimensional case for derivation of iterative forecasts.

REFERENCES

Carrasco, M. & X. Chen (2002) Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* 18, 17–39.

Davidson, J. (1994) *Stochastic Limit Theory.* Oxford University Press.

de Jong, R.M. & T. Woutersen (2007) Dynamic Time Series Binary Choice. *Economics Working Paper Archive* 538, Johns Hopkins University.

Devroye, L., L. Györfi, & G. Lugosi (1996) *A Probabililstic Theory of Pattern Recognition.* Springer.

Doob, J.L. (1949) Application of the theory of martingales. In *Le Calcul des Probabilits et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique* 13, pp. 23–27. CNRS.

Elliott, G. & R.P. Lieli (2005) Predicting Binary Outcomes. Manuscript, University of California, San Diego.

Francq, C. & J.-M. Zakoïan (2006) Mixing properties of a general class of GARCH(1,1) models without moment assumptions on the observed process. *Econometric Theory* 22, 815–834.

Geman, S. & D. Geman (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741.

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

Greenshtein, E. (2006) Best subset selection, persistency in high dimensional statistical learning and optimization under ℓ_1 constraint. *Annals of Statistics* 34, 2367–2386.
Horowitz, J.L. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–531.
Jiang, W. & M.A. Tanner (2008a) Gibbs posterior for variable selection in high dimensional classification and data mining. *Annals of Statistics* 36, 2207–2231.
Jiang, W. & M.A. Tanner (2008b) Risk Minimization for Time Series Binary Choice with Variable Selection. Technical Report 07-02, Northwestern University.
Kleijn, B.J.K. & A.W. van der Vaart (2006) Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics* 34, 837–877.
Lijoi, A., I. Prünster, & S.G. Walker (2007) Bayesian consistency for stationary models. *Econometric Theory* 23, 749–759.
Manski, C.F. (1975) Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–228.
Manski, C.F. (1985) Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27, 313–333.
Meitz, M. & P. Saikkonen (2008) Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models. *Econometric Theory* 24, 1291–1320.
Modha, D.S. & E. Masry (1996) Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory* 42, 2133–2145.
Stock, J.H. & M.W. Watson (2006) Forecasting with many predictors. In G. Elliott, C.W.J. Granger, & A. Timmermann (eds.), *Handbook of Economic Forecasting*, pp. 515–554. Elsevier.
Tanner, M.A. (1996) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. Springer-Verlag.
Zhang, T. (1999) Theoretical analysis of a class of randomized regularization methods. In *COLT 99, Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 156–163. ACM Press.
Zhang, T. (2006) Information theoretical upper and lower bounds for statistical estimation. *IEEE Transaction on Information Theory* 52, 1307–1321.
Zou, B. & L. Li (2007) The performance bounds of learning machines based on exponentially strong mixing sequences, *Computers and Mathematics with Applications* 53, 1050–1058.

APPENDIX

The following lemma on the uniform deviation of $R_N - R$ will be used for obtaining the risk performance results in the main theorems. We omit the proof here, since a slightly more general version of the lemma is contained in Jiang and Tanner (2008b), which is a technical report version of this paper that contains the details of the proofs. (The lemma here corresponds to the special case $\beta = 1$ for Lemma 1 of Jiang and Tanner, 2008b.)

LEMMA A.1. (Uniform bound with scope H(υ)). *Suppose Assumptions 1, 2, 3, 4, 5, and 6 hold. For all $\zeta > 0$, there exist constants $c_\zeta, m_\zeta > 0$ such that, for all large enough N, for all υ satisfying*

$$1 \le \upsilon \le \min\{m_\zeta (J+1)^{-1} N^{1/2} / \ln N, K + J + 1\} \equiv \upsilon_\zeta, \tag{A.1}$$

we have

$$P_D^*[\sup_{b \in H(\upsilon)} |R_n(b) - R(b)| > c_\zeta N^{-1/2}((J+1)\upsilon \ln N)] \le N^{-\zeta \upsilon}, \tag{A.2}$$

where $H(\upsilon) = \{b \in \bar{\Omega} : |b|_0 \le \upsilon\}$.

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

Proof of Theorem 1. Note that (by applying, e.g., Lemma 8.2 of Devroye et al., 1996) $2\sup_{b\in H}|R_n(b)-R(b)| \geq R(\hat{b})-\inf_{b\in H}R(b)$ and therefore $[R(\hat{b})-\inf_{b\in H}R(b)>2\delta]$ implies $[2\sup_{b\in H}|R_n(b)-R(b)|>2\delta]$. So $P_D^*[R(\hat{b})-\inf_{b\in H}R(b)>2\delta] \leq P_D^*[2\sup_{b\in H}|R_n(b)-R(b)|>2\delta]$, which is bounded by $N^{-\zeta v}$ for all large N due to Lemma A.1, if we choose $\delta = c_\zeta N^{-1/2}(J+1)v\ln N$. ■

Proof of Theorem 2. Write the difference in two parts: $R(\tilde{b})-\inf_{b\in H(v)}R(b) = I + II \equiv \{R(\tilde{b})-\tilde{R}_N(\tilde{b})\}+\{\tilde{R}_N(\tilde{b})-\inf_{b\in H(v)}R(b)\}$. If we take $b_v = \arg\min_{b\in H(v)}R_N(b)$, and denote $\Delta(j)=\sup_{b\in H(j)}|R_N(b)-R(b)|$ for each integer j, then

$$
\begin{aligned}
II &\leq \tilde{R}_N(b_v) - \inf_{b\in H(v)} R(b) \\
&\leq \lambda v + R_N(b_v) - \inf_{b\in H(v)} R(b) \\
&= \lambda v + \inf_{b\in H(v)} R_N(b) - \inf_{b\in H(v)} R(b) \\
&\leq \lambda v + 2\Delta(v)
\end{aligned}
$$

by applying, e.g., Lemma 8.2 of Devroye et al. (1996).

Now

$$
\begin{aligned}
I &= R(\tilde{b}) - R_N(\tilde{b}) - \lambda|\tilde{b}|_0 \\
&\leq \sup_{j=1}^{v_0}\{\Delta(j)-\lambda j\},
\end{aligned}
$$

since $|\tilde{b}|_0 \leq v_0$ as argued before in (7). We will take λ to be of the form (8) for some $c>0$. Note that this and Assumption 5 on J imply that $\lambda^{-1}\geq 1$, and that $v_0 \leq \min\{(2/\lambda), K+J+1\} \leq v_\zeta$ for all large enough N, if we choose a suitably large constant c in λ. Here v_ζ is the upper limit of dimensions used in Lemma A.2. Then, for all large enough N (common to all j below), Lemma A.2 states that

$$P^*[\Delta(j) > c_\zeta(\lambda/c)j] \leq N^{-\zeta j}$$

for all $j=1,\ldots,v_0$. The above inequality also holds for $j=v$ appearing in term I, for all large N, since v satisfies Assumption 7 and thus $v=o(v_\zeta)$.

These show that for all large N, with high probability at least $1-\sum_{j=1}^{\infty}N^{-\zeta j}$,

$$
\begin{aligned}
I+II &\leq \{\lambda v + 2c_\zeta(\lambda/c)v\} + \sup_{j=1}^{v_0}\{c_\zeta(\lambda/c)j-\lambda j\} \\
&\leq 3\lambda v + 0 \\
&= 3cN^{-1/2}(J+1)v\ln N,
\end{aligned}
$$

by taking a suitably large constant c. The probability of the complement of this event is at most $\sum_{j=1}^{\infty}N^{-\zeta j} \leq 2N^{-\zeta}$ for all large N. Then taking $\zeta=2$ and applying Borel-Cantelli lemma we know that with probability one, the above inequality of $I+II$ holds for all large N. ■

Proof of Proposition 1. Take b_0 in Assumption 8 to be a vector with the first v components being 1 and other components being 0. Then any b vector in $H(v)$ must have

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

exactly υ nonzero components ($|b|_0 = \upsilon$) to satisfy $|b - b_0|_1 < 1$. So $f(\upsilon) = P[|b|_0 = \upsilon] \geq P_b[|b - b_0|_1 < 1] \geq e^{-c_1 N\delta_N}$ for all large N. ■

Proof of Lemma 1. Select any $b_0 \in H(\upsilon)$. Let g be a model that contains b_0. That is, all nonzero components b_0 have indexes in set $g \cup \{2\}$. We will choose such a g with dimension $\upsilon - 1$. This is possible since b_0 lies in $H(\upsilon)$ and cannot have more than $\upsilon - 1$ nonzero components other than $(b_0)_2$. Choose $c_1 = 4c_1'$ from Assumption 9. Then we have

$$
\begin{aligned}
P_b[|b - b_0|_1 < c_1\delta_N] &\geq P_b[g]P_b[|b - b_0|_1 < c_1\delta_N | g] \\
&\overset{(a)}{\geq} \left\{ f(\upsilon) A^{-1} \binom{K+J}{\upsilon - 1}^{-1} \right\} \times P_b \left[b_2 = (b_0)_2, \upsilon \sup_{j \in g} |b_j - b_{0j}| < c_1\delta_N \middle| g \right] \\
&\geq \left\{ f(\upsilon) A^{-1} \binom{K+J}{\upsilon - 1}^{-1} \right\} \times \inf_{b_2 \in \{\pm 1\}} P(b_2) \times \inf_b p((b_j)_{j \in g} | g) (c_1\delta_N/\upsilon)^{\upsilon - 1} \\
&\overset{(b)}{\geq} \left\{ A^{-1} \binom{K+J}{\upsilon - 1}^{-1} (c_1\delta_N/\upsilon)^{\upsilon - 1} \right\} \times f(\upsilon) \times \inf_{b_2 \in \{\pm 1\}} P(b_2) \times \inf_b p((b_j)_{j \in g} | g).
\end{aligned}
$$

For step (a), we have used assumption (ii) of Lemma 1. Note that by summing over all the possible g_2's in (9) $\left(\text{there are } \binom{K+J}{\upsilon - 1} \text{ of them} \right)$, we obtain $P[g||g| + 1 = \upsilon] \geq A^{-1} \binom{K+J}{\upsilon - 1}^{-1}$, which implies (a).

The first factor on the right-hand side of inequality (b) can be bounded by, for all large N,

$$
\left\{ A^{-1} \binom{K+J}{\upsilon - 1}^{-1} (c_1\delta_N/\upsilon)^{\upsilon - 1} \right\} \geq e^{-c\upsilon \ln N} > e^{-c_1 N\delta_N/4}
$$

for some constant $c > 0$, using

$$
\upsilon \ln N = o(N\delta_N). \tag{A.3}
$$

All of these can be easily verified due to Assumption 7 on υ, choice (4) for δ_N, and Assumptions 5 and 6 on J and K, respectively. Note that for the second to fourth factors we have

$$
f(\upsilon) \geq e^{-c_1' N\delta_N} = e^{-c_1 N\delta_N/4}
$$

due to Assumption 9; and due to assumption (i) of Lemma 1, we have

$$
\inf_{b_2 \in \{\pm 1\}} P(b_2) \geq u \geq e^{-c_1 N\delta_N/4},
$$

$$
\inf_b p((b_j)_{j \in g} | g) \geq u^{\upsilon - 1} \geq e^{-c_1 N\delta_N/4},
$$

for all large N.

Then $P_b[|b - b_0|_1 < c_1\delta_N] > e^{-c_1 N\delta_N}$ for all large N, by combining the four factors. ■

Proof of Proposition 2. The part of the statement when υ is a constant of N is trivial. In the more general case, to lower bound $f(\upsilon)$, note that the truncated probabilities will

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press

be greater than the untruncated ones. Using Assumption 7 on v and expressing $\delta_N = N^{-1/2}(J+1)v\ln N$ lead to the proof. ■

Proof of Theorem 3. By applying Proposition 6 of Jiang and Tanner (2008a), we have, for any $\delta > 0$,

$$P^*_{b,D}\left[R(b) - \inf_{b\in H} R(b) > 5\delta\right]$$
$$\leq P^*_D\left[\sup_{b\in H}|R_N(b) - R(b)| > \delta\right] + e^{-2(N-J)\psi\delta}/P_b\left[R(b) - \inf_{b\in H} R(b) < \delta\right],$$

since the prior is supported on H in this paper.

So we only need to show that $P^*_D[\sup_{b\in H}|R_N(b) - R(b)| > \delta]$ is not too large, and that $P_b[R(b) - \inf_{b\in H} R(b) < \delta]$ is not too small.

We will set

$$\delta = c_{\zeta 1}N^{-1/2}(J+1)v\ln N \equiv c_{\zeta 1}\delta_N,$$

where $c_{\zeta 1}$ is a constant to be decided later through several steps. If $c_{\zeta 1} > c_\zeta$ (†), then we can apply Lemma A.1 to ensure that for all large enough N, $P^*_D[\sup_{b\in H}|R_N(b) - R(b)| > \delta] \leq N^{-\zeta v}$.

Now consider $P_b[R(b) - \inf_{b\in H} R(b) < \delta]$. Due to (the proof of) Lemma 4 (ii) of Jiang and Tanner (2008a), R is continuous and satisfies a Lipshitz condition $R(b) - R(b') \leq W|b - b'|_1$ for all $b, b' \in \Omega$ such that $|b - b'|_1 < 1$, for some constant W, under Assumptions 3 and 4 on the data generating mechanism. Also, H is easily shown to be closed and bounded and therefore $\inf_{b\in H} R(b) = R(b_0)$ for some $b_0 \in H$. So $P_b[R(b) - \inf_{b\in H} R(b) < \delta] \geq P_b[|b - b_0|_1 < W^{-1}\delta] > e^{-N\psi\delta}$ for all large enough N, under Assumption 8 on the prior P_b and our choice $\delta = c_{\zeta 1}\delta_N$, if $c_{\zeta 1} > \max\{Wc_1, \psi^{-1}c_1\}$ (‡). Now we combine the two requirements (†)(‡) on $c_{\zeta 1}$ and choose $c_{\zeta 1} > \max\{Wc_1, \psi^{-1}c_1, c_\zeta\}$.

Then, for all large enough N,

$$P^*_{b,D}\left[R(b) - \inf_{b\in H} R(b) > 5\delta\right]$$
$$\leq N^{-\zeta v} + e^{-2(N-J)\psi\delta}e^{N\psi\delta}$$
$$\leq e^{-\zeta v\ln N} + e^{-cN\delta}$$

for some constant $c > 0$ due to Assumption 5 on J. The second term is less than the first for large N due to $v\ln N = O(N\delta)$ (which is easily verified for our choices of δ and v). Then the upper bound is $\leq 2e^{-\zeta v\ln N} \leq N^{-\zeta v/2}$ for all large N. ■

https://doi.org/10.1017/S0266466609990636 Published online by Cambridge University Press