Forthcoming in *Journal of Business & Economic Statistics*

Confidence Bands for ROC Curves with Serially Dependent Data*

Kajal Lahiri†[1] and Liu Yang‡[2]

[1]Department of Economics, University at Albany, SUNY, NY 12222, USA

[2]School of Economics, Nanjing University, Jiangsu 210093, P. R. China

Abstract

We propose serial correlation-robust asymptotic confidence bands for the receiver operating characteristic (ROC) curve and its functional, *viz.* the area under ROC curve (AUC), estimated by quasi-maximum likelihood in the binormal model. Our simulation experiments confirm that this new method performs fairly well in finite samples, and confers an additional measure of robustness to non-normality. The conventional procedure is found to be markedly undersized in terms of yielding empirical coverage probabilities lower than the nominal level, especially when the serial correlation is strong. An example from macroeconomic forecasting demonstrates the importance of accounting for serial correlation when the probability forecasts for real GDP declines are evaluated using ROC.

JEL Classifications: C01, C12, C13, C53

Key words: Receiver operating characteristic curve, AUC, Binormal model, Quasi-maximum likelihood, Serial correlation, Survey of Professional Forecasters, Robustness to skewness and excess kurtosis.

*We thank Zhongwen Liang, two referees and the co-editor for many helpful comments and suggestions. Any remaining errors are solely ours.

This research was supported by the National Institute on Minority Health and Health Disparities, National Institutes of Health (grant number 1 P20 MD003373). The content is solely the responsibility of the authors and does not represent the official views of the National Institute on Minority Health and Health Disparities or the National Institutes of Health.

†Tel.: +1 518 442 4758. E-mail address: klahiri@albany.edu.

‡Corresponding author. Tel.: +86 18710075220. E-mail address: lyang2@albany.edu.

1 Introduction

The receiver operating characteristic (ROC) curve is a popular diagnostic device, originally proposed in signal detection theory, to assess the discriminatory performance of a continuous variable representing a diagnostic test, a marker, or a classifier. The last few decades, following the pioneering work of Green and Swets (1966), have witnessed a remarkable growth of research in this field. The surge in the number of articles related to ROC analysis is well documented in Krzanowski and Hand (2009). Due to its ability to summarize all relevant information in an intuitive manner, ROC curve has received considerable attention from diverse disciplines including biomedical informatics, computer science, epidemiology, meteorology, and psychology. A general introduction to ROC methodology can be found in Fawcett (2006), Pepe (2000), Swets et al. (2000), and Zhou et al. (2002).

In the finance and banking literature, ROC curve is also commonly employed as a tool to measure the accuracy of a predictive model. Stein (2005) illustrated the use of the ROC curve generated by a credit scoring model to yield an optimal cut-off and asset pricing strategy as guidelines for a bank in its lending decisions. Blöchlinger and Leippold (2006) linked the discriminatory power of a credit scoring model, as visualized by the area under an ROC curve, with the market share, revenue, loss, and profit of a bank. Ravi and Pramodh (2008) compared the performance of alternative neural networks to predict bankruptcy with respect to the area under an ROC curve based on data of Spanish and Turkish banks.

In recent years, the ROC analysis has begun to draw increasing attention in the economics profession, especially macroeconomic forecasting. Berge and Jordà (2011), utilizing ROC curve, investigated certain issues with the business cycle indicators defined by the National Bureau of Economic Research (NBER) in terms of their skill in classifying economic activity into recessions and expansions. Lahiri and Wang (2013) noted that one important but overlooked point in forecasting relatively uncommon events is the role of a threshold or cut-off, and the usual skill measures of a binary classifier combine the true accuracy with the implicit threshold. Drehmann and Juselius (2014) used ROC analysis to assess the performance of early warning indicators for emerging financial vulnerabilities in the banking sectors. Lahiri and Yang (2013) integrated the ROC analysis into a unified framework of forecast skill evaluation for a binary outcome, and surveyed a wide range of skill scores related to the ROC curve. Other important applications include the work of Cohen et al. (2009), Gorr and Schneider (2011), and Jordà et al. (2011).

Most of the aforementioned literature concentrates on the application of the ROC curve without paying much attention to different aspects of statistical inference. There are a few exceptions that address this concern. Demidenko (2012) discussed the confidence interval and confidence band in the parametric binormal model. Pepe (2003) derived various types of confidence intervals when the ROC curve is estimated by nonsmoothing empirical methods. Hall et al. (2004) constructed confidence intervals and confidence bands for the ROC curve estimated by nonparametric kernel smoothing, and suggested a parsimonious first order asymptotic approximation. Macskassy et al. (2005) examined many of the approaches to construct confidence bands for an ROC curve, and showed their empirical performance using real-life examples.

To the best of our knowledge, no previous study has considered the impact of serial correlation on the statistical inference for an ROC curve. Virtually all papers cited above assume that the sample is independently and identically distributed (i.i.d.). This may make sense in many cross-sectional designs, which are prevalent in epidemiology and medical diagnostics. However, the legitimacy of this assumption is problematic in business and economics that are often based on time series data. A similar issue arises in meteorology and many other areas as well. The extant rich inferential procedures are not directly usable in this setting. Blaskowitz and Herwartz (2014) and Pesaran and Timmermann (2009) have developed tests designed to take care of the serial correlation while testing dependence among binary variables. Wilks (2010) has shown that the failure to accommodate serial correlation will seriously underestimate the standard error of the Brier Skill Score. Pesaran and Timmermann (2009) cited many articles that deal with the effect of serial dependence on the chi-squared tests of independence based on two-way contingency tables. Our paper parallels this literature in that we robustify the current procedures to accommodate serial dependence in a parametric binormal ROC model. Though restrictive in some cases, the binormal specification is widely used and often taken as a benchmark for comparison with more flexible semiparametric or nonparametric counterparts. See Hanley (1988) and Swets (1986) to appreciate the remarkably robust features of this model, and Lasko et al. (2005) and Devlin et al. (2013) for more recent analysis. Major econometrics packages, like Stata, also provide built-in commands to implement this model by assuming the data is i.i.d. The methodology described in this paper can be slightly modified to cope with other parametric specifications in a straightforward way. See Satchell and Xia (2008) for a number of such alternative parametric specifications for ROC analysis in the context of scoring models in banking.

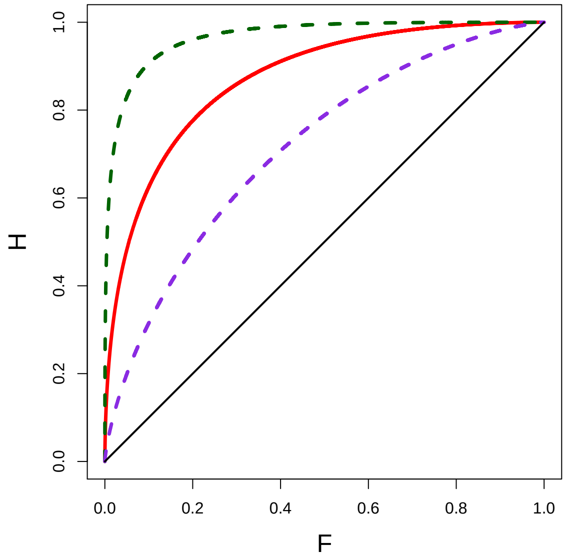
The rest of the paper is organized as follows: six types of asymptotic confidence bands in the binormal model that are robust to serial correlation are constructed in Section 2. In Section 3, results from a Monte Carlo experiment are reported to analyze the finite sample properties of these confidence bands and their robustness to departures from normality. Section 4 provides an empirical illustration using a widely used probability forecasts for real GDP declines at two quarterly horizons. The paper concludes in Section 5 with suggestions for future research. All mathematical proofs are provided in a supplementary appendix.

2 Construction of asymptotic confidence bands

2.1 The ROC curve and AUC

The ROC curve characterizes the capacity of a continuous predictor to distinguish between two possible outcomes. Specifically, let Y be the continuous variable to predict Z, the 0/1 binary target variable. Given a threshold η, we assign an observation to group 1 if Y is above η. Otherwise, it is assigned to group 0. Let the hit rate (H) or the sensitivity be the probability that the observation is correctly classified when $Z = 1$, that is, $\mathcal{P}(Y > \eta | Z = 1)$, and the false alarm rate (F) or (1-specificity) be the probability that the observation is misclassified when $Z = 0$, that is, $\mathcal{P}(Y > \eta | Z = 0)$. Ideally, we hope H could be as large as possible and F as small as possible. Both of them are functions of η. In general, given the classifier Y, it is hard to achieve a higher value of H without increasing F. The tradeoff between them is depicted by plotting the pair (F,H) in a unit square for every η. The resulting ROC curve is an increasing function from $(0,0)$ to $(1,1)$. Figure 1 presents three ROC curves. For each, the tradeoff between H and F is visually reflected by the upward sloping shape of this curve. In terms of classification performance, the upper-most curve dominates the middle curve, which in turn dominates the lower-most one. Given the tradeoff as depicted in a particular ROC curve, a decision maker can choose the optimal cut-off point for the continuous marker that would minimize the number of misclassifications and the expected loss.

Figure 1: Three ROC curves obtained from binormal models



Note that both H and F depend on the behavior of Y given $Z=1$ and $Z=0$ respectively, so neither one is affected by the prevalence of the binary event in the population. In contrast, the popular Brier Quadratic Probability Score (QPS), a probability analog of the mean squared error, depends on $\mathcal{P}(Z=1)$, which can be seen by observing $E(Z-Y)^2 = E(Z^2) - 2E(ZY) + E(Y^2) = \mathcal{P}(Z=1)(1-2E(Y|Z=1)) + E(Y^2)$. Suppose Y is a naive forecast always reporting the marginal probability of $Z=1$, that is, $Y=\mathcal{P}(Z=1)$. The QPS then reduces to $\mathcal{P}(Z=1)(1-\mathcal{P}(Z=1))$, which for rare events could be very close to zero, suggesting superior forecast skill. The lack of discriminatory power of the classifier in this case is appropriately detected by the diagonal ROC curve in Figure 1.

Sometimes, we only need a single index to summarize all information contained in an ROC curve. The area under an ROC curve (AUC) is probably the most commonly used global index of diagnostic accuracy. It is the probability that, in a randomly selected pair of values of the predictor from the two regimes (say, recessions and non-recessions), the value of the predictor is more for the recession periods, cf. Bamber (1975). Values of AUC close to one indicate a high diagnostic accuracy of the marker. Metz (1986) interpreted it as the average hit rate for all underlying values of false alarm rates, and also as the average false alarm rate for all values of hit rates, see also Lasko et al. (2005).

2.2 Asymptotic properties of the binormal estimator with serial correlation

In this section, we develop a parametric approach to construct the asymptotic confidence bands for an ROC curve based on the binormal model. However, we need to first introduce some standard assumptions and asymptotic results with respect to the fundamental parameters of the binormal model. In this subsection, we establish the strong consistency and asymptotic normality of the quasi-maximum likelihood estimator that allows for serial correlation in the score functions. In Section 2.3, we present the asymptotic confidence bands for the ROC curve with serially dependent data.

Our main results are built upon a few mild assumptions. Assumption 1 concerns the probability law governing the observed binary outcomes and classifiers.

Assumption 1 *(i)* $\{X_t = (Y_t, Z_t) : t = 1, 2, \dots\}$ *is a stochastic process on a complete probability space* $(\Omega, \mathcal{F}, \mathcal{P})$*, where* $\Omega = \times_{t=1}^{\infty} R^2$ *and* $\mathcal{F}$ *is the Borel-*σ *field generated by the measurable finite dimensional product cylinders; (ii) For some* $r' > 1$*,* $\{X_t\}$ *is a mixing sequence with either uniform mixing coefficient* ϕ_m *or strong mixing coefficient* α_m *of size* $2r'/(r'-1)$*; (iii)* $\{X_t\}$ *is strictly stationary; (iv)* $Z_t \sim$ *Bernoulli*(π^*)*, and* $\vartheta(Y_t) \sim N(\mu_Z^*, (\sigma_Z^*)^2)$ *(*$Z = 1$ *or* 0*) given* Z_t*, where* $\pi^* \in (0,1)$*,* $\theta^* \equiv (\mu_1^*, (\sigma_1^*)^2, \mu_0^*, (\sigma_0^*)^2)'$ *is a point in* $R \times R^+ \times R \times R^+$*, and*

$\vartheta(\cdot)$ *is a real-valued strictly increasing function.*

In standard ROC studies like those cited in Section 1, analysts often treat Z_t as nonstochastic. In these circumstances, Y_t (or $\vartheta(Y_t)$) is assumed to have two distributions corresponding to the two values of Z_t. Furthermore, the assumption is made that the observations are independent within the distribution as well as between distributions. This is a natural assumption under a controlled experiment. For instance, in clinical trials, the analysts are able to design an experiment to collect data recording the blood pressure (Y_t) for people from two groups of given sizes: diseased ($Z_t = 1$) and nondiseased ($Z_t = 0$). See Zhou et al. (2002) for examples of this sort. However, that Z_t is fixed is questionable in observational studies, where the recorded values of Y_t and Z_t are simultaneously determined as the realization of the underlying stochastic process. Assumption 1 specifies the joint distribution of Y_t and Z_t. 1(i) is a technical assumption. 1(ii) allows for certain degree of serial correlation in $\{X_t\}$, as long as its dependence shrinks towards zero at the stated rate. Independence is nested within 1(ii) as a special case since independent sequence must be mixing of any size. Though not necessary by itself, 1(iii) facilitates our asymptotic analysis substantially. What matters is that X_t must be identically distributed for each t in order for the population ROC curve to be well defined. 1(iv) says that the functional form of $\vartheta(\cdot)$ should be known *a priori* to transform Y_t into a mixture normal random variable. Among other things, this requires that the domain of $\vartheta(\cdot)$ must nest the range of Y_t, and the range of $\vartheta(\cdot)$ is unlimited in R. In our empirical application in Section 4, Y_t is the probability forecast. Thus any link function, like logit, probit or log-log links, for a binary dependent response in the generalized linear model would be a potential choice, see McCullagh and Nelder (1989). Alternatively, Faraggi and Reiser (2002) suggested applying a Box-Cox type power transformation before using the normal distribution. This procedure greatly robustifies the binormal model, but invokes an additional step to estimate the power parameter.

The binormal model assumes that $\{X_t\}$ is i.i.d. with $Z_t \sim \text{Bernoulli}(\pi_1)$, and $\vartheta(Y_t) \sim \text{N}(\mu_Z, \sigma_Z^2)$ ($Z = 1$ or 0) given Z_t, where $\pi_1 \in (0,1)$, $\theta \equiv (\mu_1, \sigma_1^2, \mu_0, \sigma_0^2)' \in R \times R^+ \times R \times R^+$. It derives its name from the normal specification of the two conditional distributions of $\vartheta(Y_t)$. In view of 1(iv), the model correctly specifies the distribution of X_t for each t. This, however, does not rule out the possibility of dynamic misspecification. In the presence of serial correlation as implicit in 1(ii), the distribution of $\vartheta(Y_t)$ presumably depends on previous observations of X_κ for all $\kappa < t$. Our main purpose here is to carry out the statistical inference which would be robust with respect to the possible presence of serial correlation.

We form the conditional quasi-log-likelihood function based on a sample $\{X_t : t = 1,2,...,T\}$:

$$l_T(\theta) \equiv \frac{1}{T}\sum_{t=1}^{T}(Z_t ln(f(\vartheta(Y_t);\mu_1,\sigma_1^2)) + (1-Z_t)ln(f(\vartheta(Y_t);\mu_0,\sigma_0^2))), \tag{1}$$

where $f(\cdot;\mu_Z,\sigma_Z^2)$ is the density function of a normal random variable with mean μ_Z and variance σ_Z^2, for $Z = 1$ or 0. The quasi-maximum likelihood estimator (QMLE) $\hat{\theta} \equiv (\hat{\mu}_1,\hat{\sigma}_1^2,\hat{\mu}_0,\hat{\sigma}_0^2)'$ maximizes $l_T(\theta)$. Given the normality assumption, $\hat{\theta}$ has the following explicit form

$$\hat{\mu}_1 = \frac{\sum_{t=1}^T Z_t\vartheta(Y_t)}{\sum_{t=1}^T Z_t} \tag{2a}$$

$$\hat{\sigma}_1^2 = \frac{\sum_{t=1}^T Z_t(\vartheta(Y_t)-\hat{\mu}_1)^2}{\sum_{t=1}^T Z_t} \tag{2b}$$

$$\hat{\mu}_0 = \frac{\sum_{t=1}^T (1-Z_t)\vartheta(Y_t)}{\sum_{t=1}^T (1-Z_t)} \tag{2c}$$

$$\hat{\sigma}_0^2 = \frac{\sum_{t=1}^T (1-Z_t)(\vartheta(Y_t)-\hat{\mu}_0)^2}{\sum_{t=1}^T (1-Z_t)}. \tag{2d}$$

Thus, $(\hat{\mu}_k,\hat{\sigma}_k^2)$ is the mean and variance for sub-sample of $Z_t = k$ ($k = 1$ or 0).

Theorem 1 *Under Assumption 1,* $\hat{\theta} \overset{a.s.}{\to} \theta^*$*, as* $T \to \infty$.

The strong consistency of $\hat{\theta}$ merely rests on the correct specification of the conditional distribution of $\vartheta(Y_t)$ given Z_t alone. Consequently, one can view $\{X_t : t = 1,2,...,T\}$ as a random sample and the parameters can be estimated in the usual fashion. Theorem 1 guarantees that the resulting estimator approaches the true value asymptotically.

To show asymptotic normality of $\hat{\theta}$, we have to introduce additional notations. The score function for observation t is

$$s_t(\theta) \equiv \frac{\partial(Z_t ln(f(\vartheta(Y_t);\mu_1,\sigma_1^2)) + (1-Z_t)ln(f(\vartheta(Y_t);\mu_0,\sigma_0^2)))}{\partial\theta}.$$

Define $I_T(\theta) \equiv Var(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} s_t(\theta))$, $I_T^* \equiv I_T(\theta^*)$, $J(\theta) \equiv E(\frac{\partial s_t(\theta)}{\partial\theta})$, and $J^* \equiv J(\theta^*)$.

Assumption 2 *The sequence $\{I_T^*\}$ is uniformly positive definite, i.e. I_T^* is positive definite for each $T \in N$ and there exists $\varepsilon > 0$ and a natural number $N(\varepsilon)$ such that $|I_T^*| > \varepsilon$ for all $T > N(\varepsilon)$.*

Lemma 1 *Under Assumptions 1 and 2, there exists a symmetric positive definite matrix I^* such that $I_T^* \to I^*$ as $T \to \infty$.*

Theorem 2 *Under Assumptions 1 and 2, $\sqrt{T}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, J^{*-1}I^*J^{*-1})$.*

2.3 Confidence bands for ROC curve

Based on Theorem 2, we can now get to the main objective of this paper, which is to derive the asymptotic confidence bands for an ROC curve when the data could be serially correlated. We first look at a particular point on the curve generated by fixing a threshold η. As mentioned previously, we predict Z_t to be equal to 1 whenever $\vartheta(Y_t) > \eta$. For this rule, the hit rate (H) is the conditional probability of correct classification given $Z_t = 1$, and the false alarm rate (F) is the conditional probability of incorrect classification given $Z_t = 0$. Under Assumption 1, we have

$$H^*(\eta) \equiv H(\eta;\theta^*) = 1 - \Phi(\frac{\eta - \mu_1^*}{\sigma_1^*}) = \Phi(\frac{\mu_1^* - \eta}{\sigma_1^*}) \tag{3a}$$

and

$$F^*(\eta) \equiv F(\eta;\theta^*) = 1 - \Phi(\frac{\eta - \mu_0^*}{\sigma_0^*}) = \Phi(\frac{\mu_0^* - \eta}{\sigma_0^*}), \tag{3b}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Varying η and plotting all the points $P^*(\eta) \equiv P(\eta;\theta^*) = (F(\eta;\theta^*), H(\eta;\theta^*))'$ in a unit square will produce the ROC curve, which displays the comprehensive information regarding the performance of the classifier over the entire range of η. For example, the ROC curve (the solid one) in Figure 1 corresponds to a binormal model when $\mu_1^* = -\mu_0^* = 0.8$ and $\sigma_1^* = \sigma_0^* = 1$. Note that this is the ROC curve for $\vartheta(Y_t)$, which is not our interest. However, the ROC curve for the original classifier Y_t is not altered by the strictly increasing transformation $\vartheta(\cdot)$, according to the invariance property. See Krzanowski and Hand (2009) for a rigorous proof.

The binormal model simplifies the analysis by characterizing (3a) and (3b) in terms of four parameters only. A natural estimator for each of them is obtained by replacing θ^* by its QMLE $\hat{\theta}$ in Section 2.2. For instance, $H(\eta;\hat{\theta})$ produces an estimate of $H^*(\eta)$ in (3a). However, objects like $H(\eta;\hat{\theta})$ are subject to sampling uncertainties, which must be properly accounted for. For this purpose, the confidence intervals for a given η should be used. Define

$$k_1(\eta;\theta) \equiv \frac{\mu_1 - \eta}{\sigma_1},\ k_2(\eta;\theta) \equiv \frac{\mu_0 - \eta}{\sigma_0},$$

and $k'(\eta;\theta) \equiv (k_1(\eta;\theta), k_2(\eta;\theta))'$.

Moreover, we let $\Phi(c \pm d)$ be the closed interval $[\Phi(c-d), \Phi(c+d)]$ for any positive c and d in R. For a nonempty set $O \subset R^2$, $\Phi(O) \equiv \{(\Phi(o_1), \Phi(o_2))' \in R^2 : (o_1, o_2)' \in O\}$. The $1-\alpha$ asymptotic confidence intervals for $H^*(\eta)$ and $F^*(\eta)$ are given by (4a) and (4b), respectively, in Theorem 3. To measure the joint uncertainty in estimating $H^*(\eta)$ and $F^*(\eta)$, (4c) provides the $100(1-\alpha)\%$ simultaneous confidence region for $(H^*(\eta), F^*(\eta))'$.

Theorem 3 *Suppose* $\sqrt{T}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, J^{*-1} I^* J^{*-1})$ *and* $\alpha \in (0,1)$. *For each* $\eta \in R$, *we have*

$$\lim_{T\to\infty} \mathcal{P}\left(H^*(\eta) \in \Phi\left(k_1(\eta;\hat{\theta}_T) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\partial k_1(\eta;\theta^*)}{\partial \theta} H^{*-1} I^* H^{*-1} \frac{\partial k_1(\eta;\theta^*)'}{\partial \theta} / T}\right)\right) = 1 - \alpha, \tag{4a}$$

$$\lim_{T\to\infty} \mathcal{P}\left(F^*(\eta) \in \Phi\left(k_2(\eta;\hat{\theta}_T) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\partial k_2(\eta;\theta^*)}{\partial \theta} H^{*-1} I^* H^{*-1} \frac{\partial k_2(\eta;\theta^*)'}{\partial \theta} / T}\right)\right) = 1 - \alpha, \tag{4b}$$

and

$$\lim_{T\to\infty} \mathcal{P}((H^*(\eta), F^*(\eta))' \in \Phi(\Gamma(\eta,\alpha))) = 1 - \alpha, \tag{4c}$$

where $\Gamma(\eta,\alpha)$ *is the set defined as*

$$\{O \in R^2 : T(k'(\eta;\hat{\theta}_T) - O)'(\frac{\partial k'(\eta;\theta^*)}{\partial\theta} H^{*-1} I^* H^{*-1} \frac{\partial k'(\eta;\theta^*)}{\partial\theta}')^{-1}(k'(\eta;\hat{\theta}_T) - O) \leq \chi^2_\alpha(2)\},$$

$z_{\frac{\alpha}{2}} \equiv \Phi^{-1}(1-\frac{\alpha}{2})$*, and* $\chi^2_\alpha(2)$ *is the* $1-\alpha$ *quantile of the chi-squared distribution with* 2 *degrees of freedom.*

An alternative way to define the ROC curve is to rewrite (3a)-(3b) in such a way that η does not enter the expression explicitly. It follows from (3b) that $\eta = \mu_0^* - \sigma_0^*\Phi^{-1}(F^*(\eta))$, which can be plugged in (3a) to get

$$y^*(x) \equiv y(x;\theta^*) = \Phi(\frac{\mu_1^* - \mu_0^* + \sigma_0^*\Phi^{-1}(x)}{\sigma_1^*}), \tag{5}$$

for $x \equiv F^*(\eta) \in [0,1]$. (5) is the functional form for the ROC curve in a unit square with F as horizontal axis and H as vertical axis, as is shown in Figure 1. Again, $y^*(x)$ can be estimated by $y(x;\hat{\theta})$, where the true parameter θ^* is replaced by its estimate $\hat{\theta}$. Now, define

$$k_3(x;\theta) \equiv \frac{\mu_1 - \mu_0 + \sigma_0\Phi^{-1}(x)}{\sigma_1} \text{ and } k(\theta) \equiv (\frac{\mu_1-\mu_0}{\sigma_1}, \frac{\sigma_0}{\sigma_1})'.$$

For any $0 < a < b < 1$, A(a,b) is a square matrix defined as

$$\begin{pmatrix} 1 & \Phi^{-1}(a) \\ 1 & \Phi^{-1}(b) \end{pmatrix}.$$

Furthermore, let $f_\Sigma(\cdot,\cdot;a,b)$ be the density function of a bivariate normal random vector with zero mean and covariance matrix

$$\Sigma(a,b) \equiv A(a,b)\frac{\partial k(\theta^*)}{\partial\theta} J^{*-1} I^* J^{*-1} \frac{\partial k(\theta^*)}{\partial\theta}' A(a,b)'.$$

Finally,

$$F_{sup}(u;a,b) \equiv \begin{cases} \int_{-u}^{u}\int_{-u}^{u} f_\Sigma(x_1,x_2;a,b)dx_1dx_2, & \text{if } u > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 4 offers the confidence interval of $y^*(x)$ in (5) for a given x and the uniform confidence band when x is allowed to be any value in a closed interval. For the latter, we first construct the uniform band for $k_3(x;\theta^*)$ since it is linear in $\Phi^{-1}(x)$, and scale the band by the nonlinear transformation $\Phi(\cdot)$.

Theorem 4 *Suppose $\sqrt{T}(\hat{\theta}-\theta^*) \xrightarrow{d} N(0, J^{*-1}I^*J^{*-1})$ and $\alpha \in (0,1)$. For each $x \in (0,1)$ and $0 < a < b < 1$, we have*

$$\lim_{T\to\infty} \mathcal{P}(y^*(x) \in \Phi(k_3(x;\hat{\theta}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\partial k_3(x;\theta^*)}{\partial\theta} J^{*-1}I^*J^{*-1} \frac{\partial k_3(x;\theta^*)}{\partial\theta}' /T})) = 1-\alpha \tag{6a}$$

and

$$\lim_{T\to\infty} \mathcal{P}(\forall x \in [a,b], y^*(x) \in \Phi(k_3(x;\hat{\theta}) \pm \frac{f_\alpha}{\sqrt{T}})) = 1-\alpha, \tag{6b}$$

where $z_{\frac{\alpha}{2}} \equiv \Phi^{-1}(1-\frac{\alpha}{2})$ and $f_\alpha \equiv F_{sup}^{-1}(1-\alpha;a,b)$.

We now define the confidence interval for AUC. As its name suggests, AUC is defined as the integral of $y^*(x)$ over its domain $[0,1]$, that is,

$$AUC^* \equiv AUC(\theta^*) \equiv \int_0^1 y^*(x)dx = \Phi(\frac{\mu_1^* - \mu_0^*}{\sqrt{(\sigma_1^*)^2 + (\sigma_0^*)^2}}). \tag{7}$$

The last equality of (7) is due to Krzanowski and Hand (2009). Let

$$k_4(\theta) \equiv \frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}.$$

The confidence interval for AUC^* is given in Theorem 5.

Theorem 5 *Suppose $\sqrt{T}(\hat{\theta}-\theta^*) \xrightarrow{d} N(0, J^{*-1}I^*J^{*-1})$ and $\alpha \in (0,1)$. We have*

$$\lim_{T\to\infty} \mathcal{P}(AUC^* \in \Phi(k_4(\hat{\theta}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\partial k_4(\theta^*)}{\partial\theta} J^{*-1}I^*J^{*-1} \frac{\partial k_4(\theta^*)}{\partial\theta}' /T})) = 1-\alpha, \tag{8}$$

where $z_{\frac{\alpha}{2}} \equiv \Phi^{-1}(1-\frac{\alpha}{2})$.

Compared to a poor competitor, a good classifier is rewarded by a high $H^*(\eta)$ and a low $F^*(\eta)$ for a given η, as well as by a high $y^*(x)$ for a given x, and by a high AUC^* overall. The upper-most curve in Figure 1 is generated for the case where $\mu_1^* = -\mu_0^* = 1.3$ and $\sigma_1^* = \sigma_0^* = 1$. The mean difference $\mu_1^* - \mu_0^*$ is larger than that for the solid curve, whereas both curves share the same standard deviation. This does make intuitive sense in the context of classification. If the difference of two means is small or two distributions of $\vartheta(Y_t)$ overlap to a large extent, it is hard to distinguish one from the other. In other words, a large proportion of observations could be misclassified, which is reflected by the poorer (solid) curve. The lower-most (dotted) curve has the same mean values as the solid one except for the fact that here $\sigma_1^* = \sigma_0^* = 2$. The higher standard deviation effectively dilutes the mean difference. Even if the means of two distributions are very different, the two regimes cannot be distinguished sharply unless the two distributions have small dispersions with little overlap.

Note that the confidence intervals in (4), (6) and (8) are of different forms compared to those commonly used. Nevertheless, one may construct the interval of the form $[c-d, c+d]$, where both c and d are positive. However, it is possible that either $c+d > 1$ or $c-d < 0$. This should be avoided given that the ROC curve must lie in the unit square. The merit for constructing Φ-scaled confidence bands is that the resulting estimators cannot fall beyond the feasible range in finite samples. For example, the function $\Phi(\cdot)$ ensures that both the upper and the lower bounds must be numbers between zero and one in (8). The equations (4) and (6a) are valid in the pointwise sense for a particular η or x. By contrast, (6b) is the uniform confidence band for all values of $x \in [a,b]$. In order for (6b) to be useful, f_α needs to be evaluated for any $\alpha \in (0,1)$. This is the two-tailed equicoordinate quantile of a bivariate normal distribution, which can be easily calculated by most statistical packages for a given mean vector and a covariance matrix.

A point worth emphasizing here is that not all types of confidence bands presented above are suitable for a specific decision problem under consideration. For example, with a well specified loss function, a decision maker needs to concentrate only on one η, which minimizes the expected loss over the entire range of η, as documented by Blöchlinger and Leippold (2006), Jordà (2014) and Stein (2005). Thus, the probability forecasts will be economically valuable if the point on the ROC curve corresponding to this optimal η is significantly above the diagonal. Otherwise, the decision maker would rather depend on the coin-toss naive forecast. In this situation, more attention should be paid to the pointwise confidence bands given η, instead of the confidence band of H when F is fixed or the uniform band. In other situations, the decision maker may not know the loss function. However, it may be required by law to attain a minimum permissible level of H given that a value of F is achieved. This is often encountered in medical diagnosis, where a new diagnostic device is required to satisfy a minimum value of H given an allowable F to meet the criteria set by an administrative agency. The confidence band of H given F is a legitimate solution to this problem. Finally, if we only care about the overall performance of the forecasts without regard to any η or F, we have two choices: the uniform band and the confidence interval of the AUC. The former is much more conservative in that the resulting band is wider than its pointwise colleagues.

Agresti (2007) reported the exact confidence intervals for $H^*(\eta)$ and $F^*(\eta)$ in the absence of serial correlation. Agresti and Coull (1998) argued that the coverage probabilities for these exact confidence intervals tend to be unduly large because of their inherent conservativeness. They proposed the so-called "score confidence interval" and demonstrated that it performs much better than the exact and asymptotic intervals in finite samples. Stephenson (2000) applied the score confidence intervals to judge whether the observed hit and false alarm rates in a contingency table could be obtained purely by chance. Ma and Hall (1993) took a regression view towards ROC curve and constructed the uniform band. Demidenko (2012) studied pointwise and uniform confidence bands with the shortest width. If the serial correlation is detected, none of these approaches is appropriate, and those in Theorems 3-5 should be used.

In order to be useful, all unknown terms in Theorems 3-5 have to be estimated. It is straightforward to find consistent estimators for some of them. For example, J^* can be estimated by the Hessian matrix of (1) evaluated at $\hat{\theta}$. Any partial derivative appearing in the asymptotic variances should be evaluated at $\hat{\theta}$ also. Estimating I^* is somewhat more complicated in the presence of serial correlation. Fortunately, a number of positive semi-definite estimators have been proposed in the literature dealing with heteroskedasticity and autocorrelation robust estimation. The basic idea is to use a finite sum of sample autocovariances to approximate the population infinite sum, allowing for the truncation lag to increase to infinity at an appropriate rate as the sample size grows. To ensure the positive semi-definiteness of the resulting estimators, appropriate weights, like the Bartlett sequence, are necessary. Conventionally, the ratio b of the truncation lag to the sample size is assumed to approach zero asymptotically (Andrews (1991) and Newey and West (1987, 1994)). Under this so-called “small-b asymptotics”, the resulting estimator of I^* is consistent and the t statistic is asymptotically normally distributed. As pointed out by Kiefer and Vogelsang (2005), this procedure is blind as to which weighting scheme (kernel) and the value of b should be used in any finite sample because the asymptotic distribution of the t statistic does not depend on the choice of kernel and b. They proposed a new asymptotic theory by assuming b to be a constant. However, their asymptotic distribution is nonstandard, and the critical values are obtained by simulation. Thus, in order to keep our procedure simple, we use the asymptotic F approximation developed by Sun (2013, 2014) to construct various confidence bands due to its computational ease.

3 Simulation experiments

3.1 Under correct model specification

This section serves as an illustration to shed light on the finite sample properties of the methods proposed in Section 2 when the binormal model coincides with the true underlying process. In Section 3.2 we will examine the robustness of these results when normality assumption is violated. The data is generated from two mutually independent autoregressive processes of order 1, that is,

$$Z_t^* = \tau + \rho Z_{t-1}^* + \varepsilon_t^Z \text{ and } Y_t^* = \rho Y_{t-1}^* + \varepsilon_t^Y,$$

where ε_t^Z and ε_t^Y are normal white noises and mutually independent. The variance of ε_t^Z is 1, and the variance of ε_t^Y is determined in such a way that $Var(Y_t^*) = 1$. Throughout this section, $\vartheta(\cdot)$ is taken to be the identity transformation, i.e. $\vartheta(Y_t) = Y_t$. $X_t = (Y_t, Z_t)$ is obtained by letting $Z_t = I(Z_t^* > 0)$, and $Y_t = \mu_{Z_t}^* + Y_t^*$, where $\mu_1^* = -\mu_0^* = 1$ and $I(\cdot)$ is the indicator function that is 1 only when the condition in $(\cdot)$ is met (otherwise it is 0). Two values of τ are computed so that the corresponding π^* equals 0.5 and 0.15, indicating that the event $Z_t = 1$ is balanced in the first case and relatively uncommon in the other. For independence ($\rho = 0$) and each dependence strength ($\rho = 0.3$, 0.5, 0.7, 0.9), we simulate 1,000 Monte Carlo replications of the processes. We consider samples of size $T = 200$ and 500.

To construct confidence bands, the asymptotic covariance matrix must be estimated first. Suppose we treat the sample as i.i.d., as is often done in practice. The asymptotic covariance matrix of $\hat{\theta}$ is $-J^{*-1}$, which can be estimated by its sample analog evaluated at $\hat{\theta}$. When serial correlation is accommodated, we use Andrews' (1991) quadratic spectral kernel ("small-b asymptotics") and Sun's (2014) Bartlett kernel ("fixed-b asymptotics") HAC estimators to approximate the long run variance I^*. All computations are performed in the R system with the aid of functions in the package **sandwich**. See Zeileis (2004, 2006) for additional functions in this package to compute the long run variance. The two-tailed equicoordinate quantile f_α for the uniform bands is obtained using the package **mvtnorm**. The significance level α is always set to be 5% so that the usual 95% two-tailed bands are produced. For ease of exposition, we set $\eta = 0$ and $x = 0.5$ in (4) and (6a), that is, only one point on the ROC curve is considered to avoid unnecessary clutter. For (6b), $a = 0.01$ and $b = 0.99$.

The simulation results when $\pi^* = 0.5$ are summarized in Table 1. Notably, the empirical coverage probabilities for all types of confidence bands are quite close to 95% when $\rho = 0$. When $\rho > 0$, the independent bands (under "Ind." column) cover the truth with lower frequencies than the other two bands, and the gap between them gets remarkably larger when ρ increases. Under low serial correlation ($\rho = 0.3$), the independent bands are still able to cover the truth at frequencies higher than 90% for all sample sizes. As the dependence becomes much stronger ($\rho = 0.9$), most of these coverage probabilities fall below 50%, and some of them are around 30% only even when $T = 500$. If the data display strong serial correlation, the independent bands are far too narrow to cover the true ROC curve at the nominal frequency 95%. In contrast, both Andrews' (under "Adw." column) and Sun's (under “Sun.” column) bands are more robust, and the differences between their empirical and nominal coverage probabilities are much smaller than those for independent bands. For instance, when $T = 500$ and $\rho = 0.9$, the coverage probability of Sun’s (Andrews’) interval for AUC is 95.1% (88.8%) - a sizeable improvement over the independent interval with coverage probability of only 51.2%. Although most of the coverage probabilities of Andrews’ bands are higher than 90% when $\rho > 0$, the serial correlation does weaken its performance, and appear to get narrower as ρ goes up. This may arise from finite sample bias in estimating the covariance matrix. In contrast, Sun’s bands are much more robust to serial correlation in that nearly all empirical rates are closer to 95%. This is obvious by looking at $\rho = 0.9$ and $T = 200$, where Sun’s band includes the true hit rate and false alarm rate in 97 out of 100 cases while only 77 cases are correctly covered by Andrews’ counterpart. The finding here lends further evidence for the better finite sample properties of “fixed-b asymptotics” over “small-b asymptotics”, especially when the data exhibits strong autocorrelation.

Table 2, which is qualitatively similar to Table 1, displays the coverage probabilities when $\pi^* = 0.15$. As the event $Z_t = 1$ gets relatively uncommon or rare, the finite sample distortion for some of the bands becomes more severe as ρ gets larger. In particular, when $\rho = 0.9$, it is hard for the two robust bands to achieve a rate higher than 90%, but still are significantly better that the independent band. Note that with $\rho = 0$, the rareness of the outcome variable does not create any special problem in the coverage rates. In the current context, Sun’s approach is again slightly better than Andrews’ when $T = 500$. For $T = 200$, the performance of both of them deteriorates similarly. However, it is expected. Given fewer observations for $Z_t = 1$ in a sample, any parameter relevant to the occurrence of this event is estimated less accurately. This is the case for all types of bands except $F(0)$. In a similar vein, King and Zeng (2001) found that in finite samples the slope parameter of a logit model would be less precisely estimated when the event is rare. It follows from (3b) that $F^*(0)$ is a function of μ_0^* and σ_0^* solely, both of which are related to the more dominant event $Z_t = 0$. Thus, the finite sample distortion for $F(0)$ is substantially alleviated by exploiting information contained in more observations for $Z_t = 0$. To confirm this intuition, we conducted an additional simulation study with $T = 3,000$ and $\rho = 0.9$, and indeed found that the finite sample biases of

Table 1: Empirical coverage probabilities when $\pi^* = 0.5$

	T=200			T=500		
$H(0)$	Ind.	Adw.	Sun.	Ind.	Adw.	Sun.
$\rho = 0.0$	0.946	0.936	0.936	0.949	0.946	0.941
$\rho = 0.3$	0.948	0.944	0.946	0.920	0.937	0.932
$\rho = 0.5$	0.869	0.936	0.930	0.868	0.936	0.929
$\rho = 0.7$	0.763	0.919	0.916	0.760	0.931	0.931
$\rho = 0.9$	0.479	0.839	0.905	0.487	0.878	0.942
$F(0)$						
$\rho = 0.0$	0.949	0.938	0.943	0.955	0.947	0.951
$\rho = 0.3$	0.927	0.918	0.923	0.927	0.941	0.936
$\rho = 0.5$	0.867	0.934	0.923	0.864	0.939	0.933
$\rho = 0.7$	0.749	0.918	0.924	0.748	0.926	0.929
$\rho = 0.9$	0.482	0.844	0.923	0.481	0.908	0.943
$(H(0), F(0))$						
$\rho = 0.0$	0.949	0.935	0.934	0.950	0.948	0.945
$\rho = 0.3$	0.936	0.922	0.934	0.909	0.934	0.936
$\rho = 0.5$	0.829	0.929	0.912	0.836	0.937	0.935
$\rho = 0.7$	0.670	0.902	0.933	0.659	0.906	0.931
$\rho = 0.9$	0.276	0.774	0.973	0.293	0.858	0.978
$y(0.5)$						
$\rho = 0.0$	0.955	0.950	0.952	0.949	0.946	0.945
$\rho = 0.3$	0.950	0.937	0.944	0.939	0.942	0.943
$\rho = 0.5$	0.891	0.932	0.943	0.898	0.938	0.936
$\rho = 0.7$	0.805	0.924	0.914	0.793	0.918	0.918
$\rho = 0.9$	0.520	0.846	0.894	0.526	0.877	0.926
$\{y(x) : x \in [0.01, 0.99]\}$						
$\rho = 0.0$	0.961	0.953	0.955	0.951	0.946	0.947
$\rho = 0.3$	0.945	0.937	0.934	0.945	0.946	0.944
$\rho = 0.5$	0.899	0.922	0.933	0.919	0.941	0.934
$\rho = 0.7$	0.850	0.924	0.922	0.833	0.928	0.931
$\rho = 0.9$	0.508	0.852	0.936	0.504	0.881	0.948
AUC						
$\rho = 0.0$	0.950	0.941	0.945	0.963	0.954	0.950
$\rho = 0.3$	0.936	0.925	0.928	0.937	0.948	0.948
$\rho = 0.5$	0.879	0.936	0.936	0.871	0.945	0.935
$\rho = 0.7$	0.764	0.914	0.918	0.753	0.917	0.927
$\rho = 0.9$	0.472	0.842	0.929	0.512	0.888	0.951

Notes: The columns "Ind.", "Adw." and "Sun." contain empirical coverage probabilities for the independent, Andrews' and Sun's HAC-based autocorrelation-robust 95% confidence bands respectively. The left panel presents results when T=200, while the results when T=500 are shown in the right panel. $H(0), F(0)$ and $(H(0), F(0))$ correspond to (4a), (4b) and (4c) respectively, for $\eta = 0$. $y(0.5)$ is (6a) for $x = 0.5$. $\{y(x) : x \in [0.01, 0.99]\}$ is (6b) for $a = 0.01$ and $b = 0.99$. AUC is (8).

Table 2: Empirical coverage probabilities when $\pi^* = 0.15$

	T=200			T=500		
$H(0)$	Ind.	Adw.	Sun.	Ind.	Adw.	Sun.
$\rho = 0.0$	0.941	0.921	0.927	0.955	0.940	0.942
$\rho = 0.3$	0.936	0.913	0.913	0.940	0.940	0.942
$\rho = 0.5$	0.918	0.900	0.891	0.917	0.922	0.918
$\rho = 0.7$	0.840	0.869	0.856	0.826	0.912	0.903
$\rho = 0.9$	0.574	0.790	0.742	0.576	0.827	0.850
$F(0)$						
$\rho = 0.0$	0.942	0.937	0.937	0.954	0.944	0.945
$\rho = 0.3$	0.906	0.941	0.930	0.903	0.953	0.947
$\rho = 0.5$	0.838	0.931	0.946	0.826	0.944	0.943
$\rho = 0.7$	0.718	0.928	0.944	0.702	0.944	0.941
$\rho = 0.9$	0.474	0.889	0.976	0.431	0.902	0.975
$(H(0), F(0))$						
$\rho = 0.0$	0.944	0.909	0.921	0.959	0.941	0.944
$\rho = 0.3$	0.902	0.913	0.914	0.924	0.941	0.941
$\rho = 0.5$	0.840	0.895	0.915	0.837	0.928	0.934
$\rho = 0.7$	0.690	0.874	0.918	0.673	0.900	0.920
$\rho = 0.9$	0.320	0.749	0.961	0.312	0.823	0.978
$y(0.5)$						
$\rho = 0.0$	0.950	0.917	0.926	0.952	0.933	0.937
$\rho = 0.3$	0.950	0.921	0.911	0.962	0.944	0.945
$\rho = 0.5$	0.928	0.903	0.898	0.921	0.933	0.910
$\rho = 0.7$	0.864	0.867	0.873	0.863	0.917	0.907
$\rho = 0.9$	0.631	0.782	0.733	0.587	0.815	0.835
$\{y(x) : x \in [0.01, 0.99]\}$						
$\rho = 0.0$	0.947	0.926	0.924	0.955	0.943	0.940
$\rho = 0.3$	0.947	0.925	0.918	0.965	0.956	0.950
$\rho = 0.5$	0.938	0.903	0.888	0.936	0.933	0.917
$\rho = 0.7$	0.878	0.882	0.877	0.875	0.916	0.904
$\rho = 0.9$	0.654	0.790	0.770	0.615	0.827	0.860
AUC						
$\rho = 0.0$	0.938	0.903	0.912	0.950	0.940	0.942
$\rho = 0.3$	0.922	0.918	0.914	0.938	0.945	0.938
$\rho = 0.5$	0.894	0.900	0.898	0.910	0.934	0.911
$\rho = 0.7$	0.794	0.891	0.888	0.821	0.916	0.914
$\rho = 0.9$	0.504	0.808	0.822	0.558	0.845	0.884

Notes: The columns "Ind.", "Adw." and "Sun." contain empirical coverage probabilities for the independent, Andrews' and Sun's HAC-based autocorrelation-robust 95% confidence bands respectively. The left panel presents results when T=200, while the results when T=500 are shown in the right panel. $H(0), F(0)$ and $(H(0), F(0))$ are (4a), (4b) and (4c), respectively, for $\eta = 0$. $y(0.5)$ is (6a) for $x = 0.5$. $\{y(x) : x \in [0.01, 0.99]\}$ is (6b) for $a = 0.01$ and $b = 0.99$. AUC is (8).

the robust bands largely vanish, and the empirical rate of $H(0)$ was 93.1% (Andrews') and 94.5% (Sun's).

In summary, the correlated confidence band proposed in this paper offers a significantly more robust procedure than the conventional independent band. In terms of the finite sample performance and computational cost, Sun's band based on F approximation is suggested. Whether the serial correlation is present or not, this procedure performs very well in finite samples. However, it is still subject to a moderate amount of small sample bias whose size depends on the strength of the serial dependence ρ and the rareness of $Z_t = 1$ in an expected way. As noted before, the bias could be partially attributed to the imprecision in estimating the covariance matrix using finite samples. Even when the true covariance matrix is known *a priori*, the bias might still be present since the properties stated in Theorems 3-5 are justified as $T \to \infty$. When T is small, the true coverage probabilities in unbalanced samples can be quite distinct from the prescribed level $1-\alpha$ for any type of band.

3.2 Under model misspecifications

Swets (1986) and Hanley (1988) demonstrated the remarkable robustness property of the binormal model. Walsh (1997) pointed out that even though the ROC curve under binormal assumption "fits well" under model misspecification, the inferences based on the misspecified model could be misleading. He demonstrated this point in a simulation study of the binormal estimator when the data came from a bilogistic distribution. Faraggi and Reiser (2002) and Devlin et al. (2013) have also studied the effect of different types of model misspecification in biomedical contexts. To examine how robust our confidence bands are to the possible model misspecification and data configurations that are more typical in economics and business, we consider two additional data generating processes (DGP) which are extensions of the models in Section 3.1. As before, $Z_t^* = \tau + \rho Z_{t-1}^* + \varepsilon_t^Z$, ε_t^Z follows the standard normal distribution, and $Z_t = I(Z_t^* > 0)$. However, Y_t is no longer normally distributed given Z_t. Specifically, we let $Y_{st}^* = \rho Y_{s(t-1)}^* + \varepsilon_{st}^Y$, where ε_{st}^Y is a normal white noise whose variance is determined such that $Var(Y_{st}^*) = 1$ for each $s = 1, 2, ..., S+1$. Furthermore, $(\varepsilon_{1t}^Y, \varepsilon_{2t}^Y, ..., \varepsilon_{(S+1)t}^Y, \varepsilon_t^Z)$ are mutually independent. In the first scenario, $Y_t = \mu^*_{Z_t} + \frac{\sum_{s=1}^{S} Y^{*2}_{st} - S}{\sqrt{2S}}$, where $\mu^*_1 = -\mu^*_0 = 1$. Since Y^*_{st} is a standard normal variate, $\sum_{s=1}^{S} Y^{*2}_{st}$ follows χ^2 distribution with S degrees of freedom. This process is motivated by observing that given Z_t, Y_t has mean $\mu^*_{Z_t}$ and unit variance, and thus shares the same first two moments as the process in Section 3.1. However, the distribution of Y_t is asymmetric and the degree of skewness is completely controlled by S in that a higher S is associated with a lower skewness. In the second scenario, $Y_t = \mu^*_{Z_t} + \frac{\sqrt{S-2}}{\sqrt{S}} \frac{Y^*_{(S+1)t}}{\sqrt{\sum_{s=1}^{S} Y^{*2}_{st}/S}}$. Note that given Z_t, $\frac{Y^*_{(S+1)t}}{\sqrt{\sum_{s=1}^{S} Y^{*2}_{st}/S}}$ follows the t distribution with S degrees of freedom. Consequently, Y_t is symmetrically distributed, and it has mean $\mu^*_{Z_t}$ and unit variance given Z_t. Thus, as in scenario 1, in scenario 2, Y_t shares the same first two moments as process in Section 3.1. However, the tail of its distribution is fat relative to that of a normal distribution, and the excess kurtosis is non-zero. The larger the degree of freedom S, the more the distribution resembles a normal distribution. Although these two scenarios by no means exhaust the possibilities where the maintained binormal model could be misspecified, they are of particular interest. In both cases, the first two moments, namely, the mean and the variance, match those values in Section 3.1. The robustness of our model is examined when the third and fourth moments deviate from those values corresponding to the normal benchmark. Faraggi and Reiser (2002) cautioned that a moderate bias in AUC is likely to arise when the underlying distributions are complex like bimodal mixtures and the two conditional forecast distributions are poorly separated (e,g., AUC= 0.7 or less). In these circumstances, Devlin et al. (2013) recommended estimating the ROC curve directly within a parametric family rather than modeling the marker distributions that induce the ROC curve. However, with a high discrimination power of Y_t (e.g., AUC= 0.9), they found the binormal model to be reasonably robust to bimodal distributions too. Since most of the classifiers in economics and related disciplines are unimodally distributed, as those in Section 4, we will not consider this case in the simulation. For ease of exposition, we focus on $\pi^* = 0.5$. The results with $\pi^* = 0.15$ are similar, and are available upon request.

Table 3 shows the empirical coverage probabilities of three types of confidence intervals for AUC when the DGP is skewed, where a specific skewness is obtained as $\sqrt{8/df}$. In this table, ΔAUC is the simulated difference between the AUC of the true process and that of the estimated binormal model. In order to minimize sampling variability, our computation of ΔAUC is based on a sample of size $1,000,000$ with $\rho = 0$ (ΔAUC is roughly the same for other values of ρ). The relatively minor effect of skewness on AUC is reflected by a small bias in the estimation of AUC ($\Delta AUC = 0.016$) - a finding that is consistent with the "forgiving" property of the binormal assumption as demonstrated by Hanley (1988). Furthermore, ΔAUC shrinks towards zero as the degree of freedom of χ^2 distribution (*df*) rises. This is not surprising since χ^2 distribution with a large degree of freedom is roughly symmetric, and thus can be approximated fairly well by a normal distribution.

Except for *df*=30, the independent interval performs poorly, and a larger sample ($T = 500$) does not help in improving its coverage properties. As *df* gets larger, the true AUC is approximated by the estimated AUC with a higher precision. Accordingly, the coverage rates get closer to the nominal level 95%. By ignoring the strong serial correlation in the data, the performance of the independent interval deteriorates as ρ rises in all cases. In contrast, our robust intervals provide a substantial improvement in the presence of serial correlation, especially when $\rho = 0.9$. For instance, all of the coverage probabilities of Sun's interval exceed 0.91 when $\rho = 0.9$ and $T = 500$ even though the skewness is strong (*df*=1), while the conventional independent interval covers the truth at frequency lower than 0.62. Again, Andrews' interval is slightly worse than Sun's if the data is strongly correlated. By comparing Tables 1 and 3, it is clear that all intervals are adversely impacted by the skewness, but our robust intervals seem to be affected less. In other words, our serial correlation-robust band provides an extra measure of robustness to skewness.

As noted above, $\mu_1^* = -\mu_0^* = 1$ in Table 3. When *df*=1, the population value of AUC is 0.938 and ΔAUC= 0.016 - a bias of moderate size. We also carried out another experiment (not reported here) where forecasts are much less discriminatory with $\mu_1^* = -\mu_0^* = 0.26$. In the case of *df*=1, the population value of AUC reduces to 0.75 and ΔAUC rises to 0.105. Thus, as was first pointed out by Faraggi and Reiser (2002), the bias from the misspecified binormal model due to skewness is more severe when the forecasts are less discriminatory. When $T = 200$ and $\rho = 0.7$, the empirical coverage rates for three intervals are 0.314 (independent), 0.489 (Andrews'), and 0.530 (Sun's) respectively in this scenario of low discrimi-

Table 3: Coverage probabilities for AUC in the presence of skewness

	T=200			T=500		
df=1(ΔAUC=0.016)	Ind.	Adw.	Sun.	Ind.	Adw.	Sun.
$\rho = 0.0$	0.691	0.878	0.883	0.622	0.851	0.854
$\rho = 0.3$	0.706	0.896	0.900	0.612	0.862	0.862
$\rho = 0.5$	0.649	0.895	0.896	0.567	0.856	0.854
$\rho = 0.7$	0.550	0.853	0.868	0.508	0.877	0.887
$\rho = 0.9$	0.386	0.812	0.874	0.361	0.883	0.915
df=2(ΔAUC=0.011)						
$\rho = 0.0$	0.805	0.902	0.911	0.745	0.897	0.899
$\rho = 0.3$	0.807	0.917	0.922	0.748	0.890	0.889
$\rho = 0.5$	0.769	0.898	0.902	0.715	0.897	0.897
$\rho = 0.7$	0.691	0.899	0.888	0.639	0.910	0.911
$\rho = 0.9$	0.455	0.839	0.882	0.456	0.896	0.923
df=3(ΔAUC=0.009)						
$\rho = 0.0$	0.866	0.915	0.923	0.818	0.897	0.898
$\rho = 0.3$	0.850	0.926	0.924	0.811	0.912	0.903
$\rho = 0.5$	0.819	0.923	0.926	0.765	0.918	0.911
$\rho = 0.7$	0.739	0.909	0.916	0.710	0.929	0.916
$\rho = 0.9$	0.496	0.826	0.889	0.494	0.878	0.920
df=4(ΔAUC=0.007)						
$\rho = 0.0$	0.886	0.936	0.944	0.864	0.925	0.922
$\rho = 0.3$	0.878	0.925	0.930	0.844	0.917	0.921
$\rho = 0.5$	0.844	0.930	0.935	0.823	0.936	0.936
$\rho = 0.7$	0.760	0.913	0.915	0.764	0.942	0.941
$\rho = 0.9$	0.517	0.838	0.916	0.504	0.902	0.930
df=5(ΔAUC=0.006)						
$\rho = 0.0$	0.909	0.928	0.935	0.864	0.918	0.920
$\rho = 0.3$	0.876	0.918	0.921	0.857	0.920	0.914
$\rho = 0.5$	0.858	0.931	0.925	0.809	0.919	0.916
$\rho = 0.7$	0.770	0.903	0.912	0.774	0.945	0.936
$\rho = 0.9$	0.551	0.847	0.906	0.524	0.908	0.921
df=6(ΔAUC=0.005)						
$\rho = 0.0$	0.899	0.944	0.948	0.868	0.927	0.927
$\rho = 0.3$	0.901	0.938	0.939	0.896	0.933	0.933
$\rho = 0.5$	0.856	0.927	0.927	0.842	0.926	0.920
$\rho = 0.7$	0.786	0.934	0.929	0.792	0.942	0.942
$\rho = 0.9$	0.568	0.868	0.922	0.554	0.904	0.923
df=30(ΔAUC=0.002)						
$\rho = 0.0$	0.941	0.951	0.955	0.945	0.953	0.952
$\rho = 0.3$	0.935	0.935	0.934	0.940	0.947	0.947
$\rho = 0.5$	0.910	0.932	0.932	0.911	0.942	0.938
$\rho = 0.7$	0.870	0.947	0.946	0.852	0.937	0.936
$\rho = 0.9$	0.611	0.885	0.926	0.617	0.909	0.943

Notes: The columns “Ind.”, “Adw.” and “Sun.” contain empirical coverage probabilities for the independent, Andrews’ and Sun’s HAC-based autocorrelation-robust 95% confidence intervals for AUC respectively. The left panel presents results when $T = 200$, while the results when $T = 500$ are shown in the right panel. “df” is the degree of freedom of the χ^2 distribution in DGP. “ΔAUC” is the simulated difference between the AUC of the true process and that of the estimated binormal model.

Table 4: Coverage probabilities for AUC in the presence of excess kurtosis

	T=200			T=500		
df=4(ΔAUC=0.017)	Ind.	Adw.	Sun.	Ind.	Adw.	Sun.
$\rho = 0.0$	0.786	0.902	0.905	0.667	0.854	0.852
$\rho = 0.3$	0.777	0.894	0.894	0.684	0.862	0.863
$\rho = 0.5$	0.764	0.888	0.887	0.646	0.864	0.857
$\rho = 0.7$	0.693	0.887	0.898	0.592	0.873	0.896
$\rho = 0.9$	0.448	0.810	0.896	0.462	0.844	0.916
df=5(ΔAUC=0.010)						
$\rho = 0.0$	0.862	0.919	0.923	0.799	0.896	0.897
$\rho = 0.3$	0.837	0.919	0.920	0.787	0.904	0.903
$\rho = 0.5$	0.805	0.913	0.916	0.735	0.889	0.898
$\rho = 0.7$	0.722	0.905	0.922	0.667	0.901	0.911
$\rho = 0.9$	0.504	0.843	0.919	0.453	0.843	0.925
df=6(ΔAUC=0.007)						
$\rho = 0.0$	0.885	0.926	0.932	0.863	0.922	0.922
$\rho = 0.3$	0.875	0.931	0.929	0.838	0.924	0.920
$\rho = 0.5$	0.826	0.932	0.934	0.786	0.910	0.911
$\rho = 0.7$	0.765	0.928	0.939	0.721	0.913	0.933
$\rho = 0.9$	0.493	0.833	0.934	0.495	0.867	0.948
df=7(ΔAUC=0.006)						
$\rho = 0.0$	0.901	0.935	0.939	0.881	0.929	0.930
$\rho = 0.3$	0.897	0.940	0.937	0.858	0.937	0.932
$\rho = 0.5$	0.869	0.937	0.933	0.819	0.922	0.923
$\rho = 0.7$	0.751	0.919	0.936	0.752	0.937	0.943
$\rho = 0.9$	0.479	0.824	0.917	0.472	0.848	0.928
df=8(ΔAUC=0.005)						
$\rho = 0.0$	0.920	0.938	0.941	0.889	0.919	0.919
$\rho = 0.3$	0.906	0.947	0.946	0.885	0.930	0.932
$\rho = 0.5$	0.864	0.938	0.933	0.857	0.943	0.941
$\rho = 0.7$	0.743	0.924	0.928	0.766	0.941	0.952
$\rho = 0.9$	0.478	0.839	0.922	0.494	0.885	0.934
df=9(ΔAUC=0.004)						
$\rho = 0.0$	0.910	0.926	0.933	0.920	0.945	0.948
$\rho = 0.3$	0.932	0.965	0.961	0.889	0.924	0.922
$\rho = 0.5$	0.862	0.940	0.936	0.845	0.930	0.932
$\rho = 0.7$	0.743	0.921	0.917	0.752	0.932	0.945
$\rho = 0.9$	0.458	0.836	0.925	0.476	0.865	0.936
df=100(ΔAUC=0.001)						
$\rho = 0.0$	0.941	0.935	0.937	0.958	0.954	0.956
$\rho = 0.3$	0.921	0.940	0.943	0.949	0.960	0.964
$\rho = 0.5$	0.903	0.942	0.938	0.906	0.958	0.953
$\rho = 0.7$	0.772	0.918	0.921	0.801	0.929	0.935
$\rho = 0.9$	0.508	0.843	0.937	0.525	0.890	0.960

Notes: The columns “Ind.”, “Adw.” and “Sun.” contain empirical coverage probabilities for the independent, Andrews’ and Sun’s HAC-based autocorrelation-robust 95% confidence intervals for AUC respectively. The left panel presents results when $T = 200$, while the results when $T = 500$ are shown in the right panel. “*df*” is the degree of freedom of the t distribution in DGP. The excess kurtosis is well defined only if *df*$>$ 4. “ΔAUC” is the simulated difference between the AUC of the true process and that of the estimated binormal model.

nation. Again, our robust bands suffer from the misspecification bias, but they still perform better than the independent band in the presence of serial correlation. In these circumstances, Faraggi and Reiser (2002) found that a Box-Cox transformation of the forecasts before applying the binormal model performed admirably well.

When the DGP exhibits excess kurtosis, the coverage properties can be found in Table 4. Here the excess kurtosis is obtained as 6/(*df*-4). By comparing this table with the corresponding figures in Table 1, we find that excess kurtosis has an additional adverse effect on the coverage rates that gets worse with higher autocorrelation. As in Table 3, the performance of all types of intervals varies with ρ and *df* in an expected manner, but our robust bands significantly improve the situation across the board. For instance, with *df*=6 (i.e., excess kurtosis = 3), $T = 500$, and $\rho = 0.9$, the coverage rate for the independent interval is only 0.495, compared to 0.948 for Sun's interval. It is interesting to note that in both Tables 3 and 4, the independent interval is strictly worse than the two robust intervals even without serial correlation ($\rho = 0$) for smaller *df*'s. The reason is that the independent interval is based on the inverse of negative Hessian matrix, i.e. $-J^{*-1}$. In contrast, both robust intervals are based on the sandwich covariance matrix $J^{*-1}I^{*}J^{*-1}$, which is equal to $-J^{*-1}$ by the well-known information equality when the model is correctly specified. However, $J^{*-1}I^{*}J^{*-1}$ usually yields larger variances than $-J^{*-1}$ when the model is misspecified. This explains why all types of intervals are equally good in Tables 1 - 4 (*df*=30) when $\rho = 0$, as in all of these circumstances, the binormal model is exactly or roughly correct and the information equality holds. When *df* in Tables 3 or 4 is small, the information equality fails and the independent interval tends to be unduly narrower than the robust counterparts.

The implication of these simulation results is that i) a suitably transformed binormal model is pretty robust to all but poorly separated complex distributions, and ii) the autocorrelation-robust confidence intervals are always preferred if the time dependence is present, and iii) our robust bands provide an extra measure of robustness when the binormal model is misspecified due to non-zero skewness and excess kurtosis.

4 An empirical application to SPF forecasts of GDP downturns

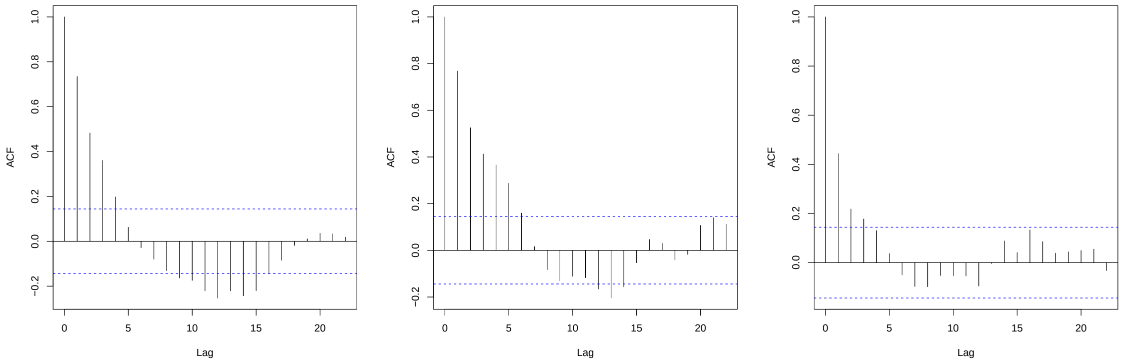
We apply the methodology outlined above to evaluate the accuracy of the subjective probability forecasts of real GDP downturns in the *Survey of Professional Forecasters* (SPF). The SPF, conducted by the Federal Reserve Bank of Philadelphia, is a leading U.S. survey collecting subjective probability predictions in economics. The respondents of this survey are asked to indicate the probability they would attach to a decline in the level of real GDP in the current and the next four quarters. For the sake of illustration, we focus on the current-quarter and three-quarter-ahead probability forecasts averaged over individuals. Our sample covers the period from 1968:Q4 to 2014:Q4. In order to evaluate the forecasts in real time, we calculated the actual GDP growth rates based on values known one month after the quarter. The total number of observations is 185, and the fraction of real GDP declines is about 13%, lower than the mean recorded forecasts 19.3% (for current-quarter forecasts) and 17.2% (for three-quarter-ahead forecasts). The reader is referred to Croushore (1993) for a general introduction to SPF. Lahiri and Wang (2013) used a battery of diagnostic tools including the ROC curve to examine the value of these forecasts over five available horizons. They found that the current-quarter forecasts have impressive discriminatory power, whereas the quality of three-quarter-ahead forecasts is marginal. Thus, studying the SPF forecasts at these two horizons will allow us to examine if the quality of the forecasts affects the relative distortion of the confidence bands in a misspecified model with autocorrelation.

Our serial correlation-robust ROC analysis is motivated by Figure 2, which depicts the sample autocorrelation functions of SPF forecasts and the actual, respectively. The plots display the presence of moderate serial correlation, especially for the forecasts, where all the autocorrelation coefficients up to four-quarter lag are significantly different from zero. Like the binary target variable, real GDP growth in real time also exhibited significant autocorrelation up to lag 3. Ignoring this temporal dependence, which is typical in most economic time series data, would make the resulting inference misleading.

To implement QMLE, the probability forecasts with zero and one as two natural bounds need to be transformed using $\vartheta(\cdot)$. As mentioned in Section 2.2, we adopt the bellwether probit link function, i.e. we set $\vartheta(Y_t) = \Phi^{-1}(Y_t)$. Table 5 shows several descriptive statistics of the Φ^{-1}-transformed SPF forecasts given the two economic states, that is, $Z = 1$ and $Z = 0$. Except for the current-quarter forecasts conditional on $Z = 0$, all other statistics are close to zero, validating the normality assumption. For the three-quarter-ahead forecasts, the Kolmogorov-Smirnov (K-S) test of normality in Table 5 lends further support to the binormal model. For the current-quarter forecasts, the normality assumption is reasonable when $Z = 1$ although it is rejected when $Z = 0$. We could have experimented with alternative link functions to attain normality for the current-quarter forecasts when $Z = 0$. However, since the conditional distributions of the transformed forecasts are unimodal and the current-quarter forecasts are highly discriminatory, we kept the probit link, noting that the mild skewness and excess kurtosis in the current-quarter forecasts should not generate any bias in our inference according to the results summarized in Section 3.2.

Figures 3 and 4 present five types of confidence bands, as in Theorems 3-4, for the current-quarter and three-quarter-ahead SPF forecasts, respectively. We only report Sun's band ("Corr. 95% bands") because of its superior finite sample properties, as found in Sec-

Figure 2: Autocorrelation functions of SPF and the Actual



(a) Current-quarter forecasts (b) Three-quarter-ahead forecasts (c) Real GDP declines

Notes: The autocorrelation function of real GDP declines is based on $0/1$ binary series $\{Z_t : t = 1, ..., T\}$, where Z_t equals 1 (0) if real GDP declined (rose) in quarter t. The dotted lines are 95% confidence band about the zero line.

tion 3. Observe that these bands are not symmetric around the estimated ROC curve. This is simply due to the scaling of $\Phi(\cdot)$, which, as argued before, guarantees that the bands cannot go outside the unit square. As is evident from these graphs, all confidence bands become wider when the serial correlation is taken into consideration. This is because of the autocorrelation functions in Figure 2. Given that the estimated long run variance is a weighted sum of the sample variance and autocovariances of the score vector, the independent confidence bands, by ignoring the autocovariances, suffer from a downward bias.

It is interesting to note that inference based on the independence assumption in Figures 4(a)-4(d) leads us to believe that the three-quarter-ahead SPF outperforms a coin-toss naive forecast whose ROC curve is represented by the diagonal ROC line. However, when looking at the confidence bands under autocorrelation, a different picture emerges because the lower bounds of the robust bands do not stay over the diagonal line uniformly. As a comparison, the confidence bands for the current-quarter probability forecasts in Figure 3 indicate that the professional forecasters perform much better at this shorter horizon, as reflected by a higher ROC curve. As in the case of three-quarter-ahead forecasts, here too the confidence band accounting for the serial correlation is always wider than its independent benchmark. However, the wider confidence bands for the current-quarter forecasts do not include the diagonal, meaning that both methods lead to essentially the same inference in rejecting zero forecast skill. In fact, all bands in Figure 3 lie in the upper left triangle, producing an overwhelming evidence on the high accuracy of the current-quarter forecasts. Only in the case of three-quarter-ahead forecasts do the two methods of constructing confidence band make

Table 5: Descriptive statistics of the transformed SPF forecasts

	Current-quarter forecasts		Three-quarter-ahead forecasts	
Statistics	$Z = 1$	$Z = 0$	$Z = 1$	$Z = 0$
skewness	-0.579	1.124	-0.022	0.235
excess kurtosis	-0.080	1.686	-0.920	0.226
K-S test statistic	0.126	0.111^{*}	0.078	0.060

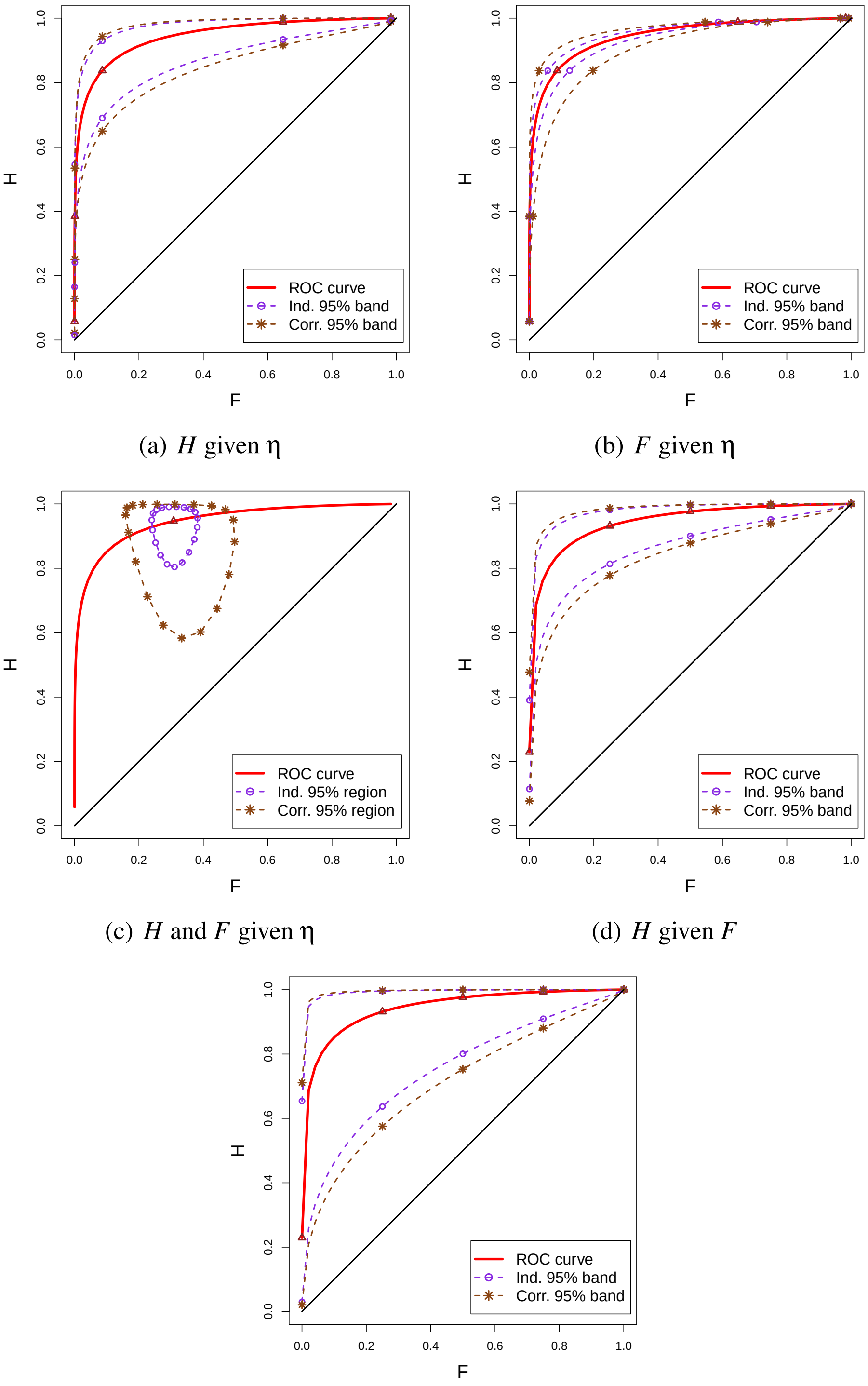
Notes: Column "Z = 1" ("Z = 0") contains statistics for the Φ^{-1}-transformed SPF forecasts when real GDP declines (rises).* means significance at 5% level.

a difference in conclusion. Interestingly, unlike in Wilks (2010), the current-quarter forecast bands are not affected relatively more than the three-quarter bands because the autocorrelations were largely the same for the two forecasts.

The uniform band and the confidence interval for AUC can be informative if one is interested in evaluating the value of the forecasts over the whole range. For the two forecast horizons, the bounds for the uniform band are reported in Figures 3(e) and 4(e) such that we are 95% certain that the true ROC curve when $F \in [0.01, 0.99]$ would fall strictly between the two bounds. To put it differently, if the uniform band excludes the diagonal and the estimated ROC curve lies in the upper triangular area, we are more certain about the performance of the forecasts. This is true for the current-quarter forecasts, but not so for the three-quarter-ahead forecasts. From Figure 4(e), it is clear that we cannot reject that the population ROC curve lies below the diagonal using either of the bands, although the curve becomes more likely to lie below the diagonal when serial correlation is adjusted for. For the current-quarter forecasts, the estimated AUC value is 0.945, with the two 95% confidence intervals: $[0.875, 0.980]$ (assuming independence) and $[0.825, 0.988]$ (not assuming independence). Either interval indicates an AUC significantly higher than 0.5, which is the AUC of the coin-toss naive forecast. For the three-quarter-ahead forecasts, the AUC estimate is only 0.649, and the two intervals are $[0.541, 0.746]$ (assuming independence) and $[0.409, 0.840]$ (not assuming independence). Clearly, taking serial correlation into account ends up with a drastically different conclusion with respect to three-quarter-ahead forecasts. To reiterate, a comparison of the five types of confidence bands in Figures 3 and 4 highlights the importance of allowing for autocorrelation in making inference about ROC and its various functionals. Also, ignoring autocorrelation tends to underestimate the true sampling uncertainty, irrespective of the quality of forecasts.

Our illustrative examples also suggest that inference on AUC should be combined with other types of confidence bands (as in Figures 3 and 4) to draw meaningful conclusions regarding the quality of forecasts. Consider the case where the ROC curve crosses the no-skill diagonal line. The AUC could still be larger than 0.5 as long as the part above the diagonal dominates. A significant AUC may tell nothing useful to a decision maker if only

Figure 3: Estimated 95% confidence bands of the ROC curve for current-quarter forecasts



(a) H given η

(b) F given η

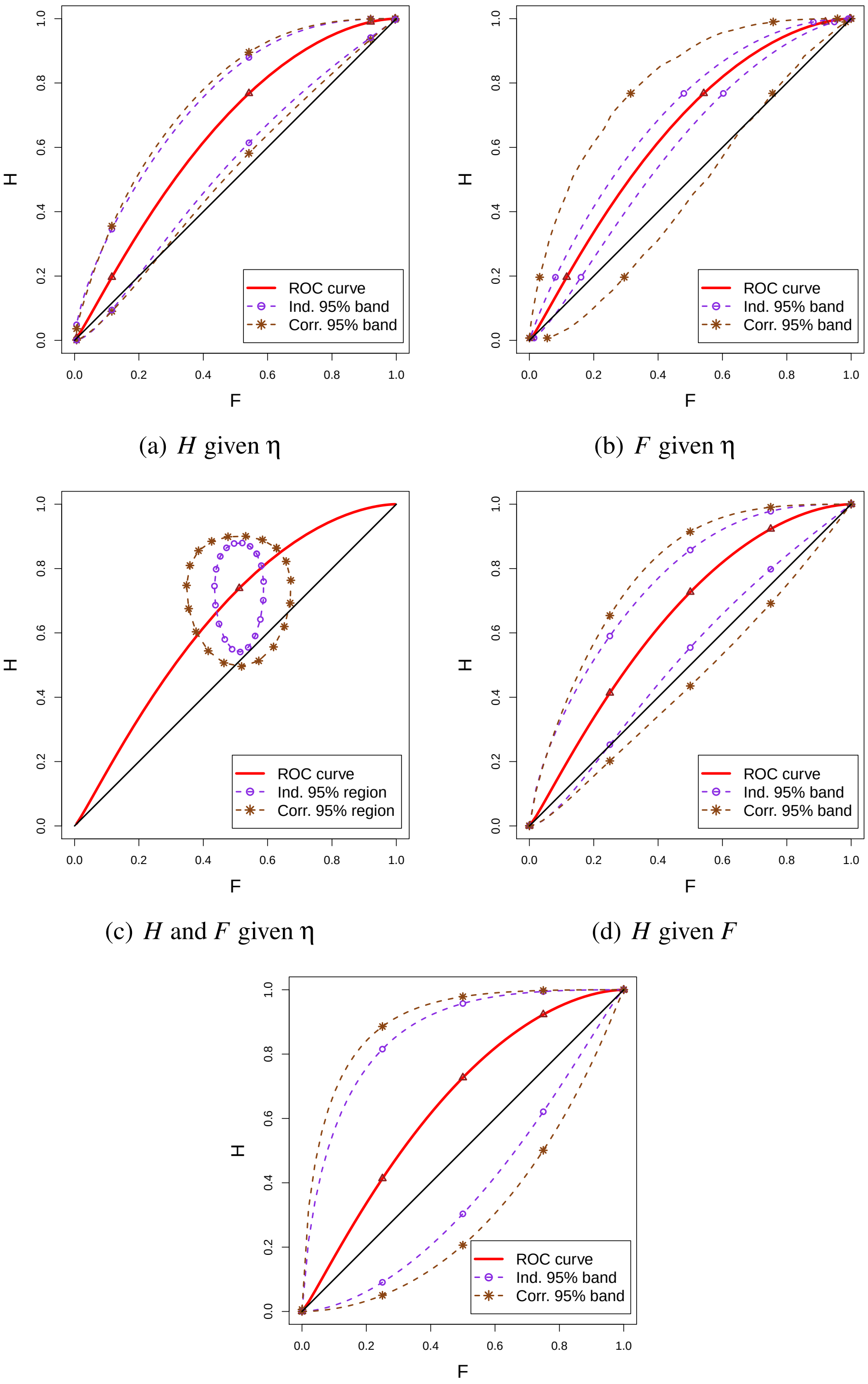
(c) H and F given η

(d) H given F

(e) Uniform

Notes: (a)-(e) present the confidence bands of (4a), (4b), (4c), (6a), and (6b), respectively. We consider only one value of η to make (c) readable. For (e), the band for $F \in [0.01, 0.99]$ is plotted.

Figure 4: Estimated 95% confidence bands of the ROC curve for three-quarter-ahead forecasts



(a) *H* given η

(b) *F* given η

(c) *H* and *F* given η

(d) *H* given *F*

(e) Uniform

Notes: (a)-(e) present the confidence bands of (4a), (4b), (4c), (6a), and (6b), respectively. We consider only one value of η to make (c) readable. For (e), the band for $F \in [0.01, 0.99]$ is plotted.

the cut-off values corresponding to the part under the diagonal are relevant in the context of a particular loss structure. Only when we are sure that the ROC curve cannot lie in the lower triangular area, reporting the global AUC makes an unequivocal sense.

5 Conclusion and further remarks

In this paper, we developed six types of serial correlation-robust confidence bands for ROC curves in the binormal model. Our asymptotic theory is based on the law of large numbers and central limit theorem for a mixing sequence. The confidence bands are obtained from a direct application of the (functional) delta method. The simulation experiment we conduct shows a better finite sample performance of these robust confidence bands than conventional independent bands regardless of the underlying processes. Depending on the sample size, the event's base rate, and the type of band considered, we find about $32\% - 102\%$ improvement in our robust bands in terms of the coverage probabilities in the presence of strong serial correlation. Simulation experiments show that accounting for positive serial correlation would widen the confidence bands for small as well as for large sample sizes. In our illustrative example, we show how the conclusion regarding forecast skill for the three-quarter-ahead SPF probability forecasts is reversed when serial correlation is accounted for while constructing the various types of confidence intervals. The underestimation of the coverage rates of the bands when autocorrelation is ignored does not seem to depend on the quality of forecasts.

We emphasized that our robust procedure still suffers from finite sample distortion in small samples with strong serial correlation, especially when the event being predicted is relatively rare. However, in large samples, where the number of realizations of the rare event is adequate, the problem goes away and the use of robust bands improves the situation quite remarkably. As an added bonus, our autocorrelation-robust bands provide an extra measure of robustness in the presence of skewness and excess kurtosis.

We only considered confidence bands in a fully parametric binormal model, which might be misspecified due to skewness and excess kurtosis in the underlying process. Unless the

DGP is close to the binormal model, the misspecification bias cannot be overlooked particularly when the forecasts are not very powerful. However, the binormal models only needs to assume that there is a monotone function that will simultaneously transform the forecasts for the two regimes into normal distributions, because the ROC curve remains invariant to such transformations. The Box-Cox transformation to normality works very well in a wide variety of cases, but needs an additional step to estimate the power parameter. If such a transformation is not feasible because the distributions are too complex, a semiparametric version of the model is a potential choice (Cai and Moskowitz (2004), Cai and Pepe (2002), Hsieh and Turnbull (1998), and Metz et al. (1998)). If one wants to discard the binormal assumption altogether, nonparametric method is the most robust approach to follow, provided we have a fairly large sample and adequate number of points on the curve (Lloyd (1998)). Unfortunately, there is no study on semiparametric and nonparametric ROC confidence bands robust to serial correlation. We leave these for future research.

References

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, John Wiley & Sons.

Agresti, A. and Coull, B. A. (1998), 'Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions', *The American Statistician* **52**, 119–126.

Andrews, D. W. K. (1991), 'Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation', *Econometrica* **59**, 817–858.

Bamber, D. (1975), 'The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph', *Journal of Mathematical Psychology* **12**, 387–415.

Berge, T. J. and Jordà, Ò. (2011), 'Evaluating the Classification of Economic Activity into Recessions and Expansions', *American Economic Journal: Macroeconomics* **3**, 246–277.

Blaskowitz, O. and Herwartz, H. (2014), 'Testing Directional Forecast Value in the Presence of Serial Correlation'. *International Journal of Forecasting* **30**, 30–42.

Blöchlinger, A. and Leippold, M. (2006), 'Economic Benefit of Powerful Credit Scoring', *Journal of Banking & Finance* **30**, 851–873.

Cai, T. and Moskowitz, C. S. (2004), 'Semi-parametric Estimation of the Binormal ROC Curve for a Continuous Diagnostic Test', *Biostatistics* **5**, 573–586.

Cai, T. and Pepe, M. S. (2002), 'Semiparametric Receiver Operating Characteristic Analysis to Evaluate Biomarkers for Disease', *Journal of the American Statistical Association* **97**, 1099–1107.

Cohen, J., Garman, S. and Gorr, W. (2009), 'Empirical Calibration of Time Series Monitoring Methods using Receiver Operating Characteristic Curves', *International Journal of Forecasting* **25**, 484–497.

Croushore, D. (1993), Introducing: The Survey of Professional Forecasters. Federal Reserve Bank of Philadelphia Business Review, November/December, 3-13.

Demidenko, E. (2012), ‘Confidence Intervals and Bands for the Binormal ROC Curve Revisited’, *Journal of Applied Statistics* **39**, 67–79.

Devlin, S. A., Thomas, E. G. and Emerson, S. S. (2013), ‘Robustness of Approaches to ROC Curve Modeling under Misspecification of the Underlying Probability Model’, *Communications in Statistics-Theory and Methods* **42**, 3655–3664.

Drehmann, M. and Juselius, M. (2014), ‘Evaluating Early Warning Indicators of Banking Crises: Satisfying Policy Requirements’, *International Journal of Forecasting* **30**, 759–780.

Faraggi, D. and Reiser, B. (2002), ‘Estimation of the Area under the ROC Curve’, *Statistics in Medicine* **21**, 3093–3106.

Fawcett, T. (2006), ‘An Introduction to ROC Analysis’, *Pattern Recognition Letters* **27**, 861–874.

Gorr, W. and Schneider, M. J. (2011), ‘Large-Change Forecast Accuracy: Reanalysis of M3-Competition Data using Receiver Operating Characteristic Analysis’, *International Journal of Forecasting* **29**, 274–281.

Green, D. M. and Swets, J. A. (1966), *Signal Detection Theory and Psychophysics*, John Wiley & Sons.

Hanley, J. A. (1988), ‘The Robustness of the “Binormal” Assumptions Used in Fitting ROC Curves’, *Medical Decision Making* **8**, 197–203.

Hsieh, F. and Turnbull, B. W. (1996), ‘Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve’, *The Annals of Statistics* **24**, 25–40.

Jordà, Ò. (2014), ‘Assessing the Historical Role of Credit: Business Cycles, Financial Crises and the Legacy of Charles S. Peirce’, *International Journal of Forecasting* **30**, 729–740.

Jordà, Ò, Schularick, M. and Taylor, A. M. (2011), ‘Financial Crises, Credit Booms, and External Imbalances: 140 Years of Lessons’, *IMF Economic Review* **59**, 340–378.

Kiefer, N. M. and Vogelsang, T. J. (2005), ‘A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests’, *Econometric Theory* **21**, 1130–1164.

King, G. and Zeng, L. (2001), 'Logistic Regression in Rare Events Data', *Political Analysis* **9**, 137–163.

Krzanowski, W. J. and Hand, D. J. (2009), *ROC Curves for Continuous Data*, Chapman & Hall.

Lahiri, K. and Wang, J. G. (2013), 'Evaluating Probability Forecasts for GDP Declines using Alternative Methodologies', *International Journal of Forecasting* **29**, 175–190.

Lahiri, K. and Yang, L. (2013), Forecasting Binary Outcomes, *in* A. Timmermann and G. Elliott, eds, 'Handbook of Economic Forecasting Volume 2B', North-Holland Amsterdam, pp. 1025–1106.

Lasko, T. A., Bhagwat, J. G., Zou, K. H. and Ohno-Machado, L. (2005), 'The Use of Receiver Operating Characteristic Curves in Biomedical Informatics', *Journal of Biomedical Informatics* **38**, 404–415.

Lloyd, C. J. (1998), 'Using Smoothed Receiver Operating Characteristic Curves to Summarize and Compare Diagnostic Systems', *Journal of the American Statistical Association* **93**, 1356–1364.

Ma, G. and Hall, W. J. (1993), 'Confidence Bands for Receiver Operating Characteristic Curves', *Medical Decision Making* **13**, 191–197.

Macskassy, S., Provost, F. and Rosset, S. (2005), ROC Confidence Bands: An Empirical Evaluation. In Proceedings of the 22nd International Conference on Machine Learning (ICML-2005).

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall.

Metz, C. E. (1986), 'ROC Methodology in Radiologic Imaging', *Investigative Radiology* **21**, 720–733.

Metz, C. E., Herman, B. A. and Shen, J. (1998), 'Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves from Continuously Distributed Data', *Statistics in Medicine* **17**, 1033–1053.

Newey, W. K. and West, K. D. (1987), ‘A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix’, *Econometrica* **55**, 703–708.

Newey, W. K. and West, K. D. (1994), ‘Automatic Lag Selection in Covariance Matrix Estimation’, *Review of Economic Studies* **61**, 631–653.

Pepe, M. S. (2000), ‘Receiver Operating Characteristic Methodology’, *Journal of the American Statistical Association* **95**, 308–311.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press.

Pesaran, M. H. and Timmermann, A. (2009), ‘Testing Dependence among Serially Correlated Multi-Category Variables’, *Journal of the American Statistical Association* **104**, 325–337.

Ranjan, R. and Gneiting, T. (2010), ‘Combining Probability Forecasts’, *Journal of the Royal Statistical Society, Series B* **72**, 71–91.

Ravi, V. and Pramodh, C. (2008), ‘Threshold Accepting Trained Principal Component Neural Network and Feature Subset Selection: Application to Bankruptcy Prediction in Banks’, *Applied Soft Computing* **8**, 1539–1548.

Satchell, S. E. and Xia, W. (2008), Analytic Models of the ROC Curve: Applications to Credit Rating Model Validation, *in* G. Christodoulakis and S. Satchell, eds, ‘The Analytics of Risk Model Validation’, Academic Press.

Stein, R. M. (2005), ‘The Relationship between Default Prediction and Lending Profits: Integrating ROC Analysis and Loan Pricing’, *Journal of Banking & Finance* **29**, 1213–1236.

Stephenson, D. B. (2000), ‘Use of the ‘Odds Ratio’ for Diagnosing Forecast Skill’, *Weather Forecasting* **15**, 221–232.

Sun, Y. (2013), ‘A Heteroskedasticity and Autocorrelation Robust F Test using an Orthonormal Series Variance Estimator’, *Econometrics Journal* **16**, 1–26.

Sun, Y. (2014), ‘Let’s Fix it: Fixed-b Asymptotics versus Small-b Asymptotics in Heteroscedasticity and Autocorrelation Robust Inference’, *Journal of Econometrics* **178**, 659–677.

Swets, J. A. (1986), ‘Form of Empirical ROCs in Discrimination and Diagnostic Tasks: Implications for Theory and Measurement of Performance’, *Psychological Bulletin* **99**, 181–198.

Swets, J. A., Dawes, R. M. and Monahan, J. (2000), ‘Better Decisions through Science’, *Scientific American* **283**, 82–87.

Walsh, S. J. (1997), ‘Limitations to the Robustness of Binormal ROC Curves: Effects of Model Misspecification and Location of Decision Thresholds on Bias, Precision, Size and Power’, *Statistics in Medicine* **16**, 669–679.

Wilks, D. S. (2010), ‘Sampling Distributions of the Brier Score and Brier Skill Score under Serial Dependence’, *Quarterly Journal of the Royal Meteorological Society* **136**, 2109–2118.

Zeileis, A. (2004), ‘Econometric Computing with HC and HAC Covariance Matrix Estimators’, *Journal of Statistical Software* **11**, 1–17.

Zeileis, A. (2006), ‘Object-oriented Computation of Sandwich Estimators’, *Journal of Statistical Software* **16**, 1–16.

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002), *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons.

Mathematical Appendix

This appendix accompanies the paper "Confidence Bands for ROC Curves with Serially Dependent Data". It contains proofs of all theorems in the text.

Proof (Theorem 1): By (2a),

$$\hat{\mu}_1 = \frac{\sum_{t=1}^T Z_t \vartheta(Y_t)/T}{\sum_{t=1}^T Z_t/T}.$$

It follows from the strong law of large number for mixing sequences (cf. McLeish (1975)) that $\sum_{t=1}^T Z_t\vartheta(Y_t)/T \overset{a.s.}{\to} E(Z_t\vartheta(Y_t)) = \pi^*\mu_1^*$ and $\sum_{t=1}^T Z_t/T \overset{a.s.}{\to} \pi^*$. Hence, $\hat{\mu}_1 \overset{a.s.}{\to} \mu_1^*$. Consistency of $\hat{\mu}_0$ is established in a similar fashion.

By (2b),

$$\hat{\sigma}_1^2 = \frac{\sum_{t=1}^T Z_t(\vartheta(Y_t)-\mu_1^*)^2/T}{\sum_{t=1}^T Z_t/T} + \frac{2\sum_{t=1}^T Z_t(\vartheta(Y_t)-\mu_1^*)(\mu_1^*-\hat{\mu}_1)/T}{\sum_{t=1}^T Z_t/T} + \frac{\sum_{t=1}^T Z_t(\mu_1^*-\hat{\mu}_1)^2/T}{\sum_{t=1}^T Z_t/T},$$

where the first term converges to $(\sigma_1^*)^2$ by the strong law, and the last two terms converge to 0 due to the consistency of $\hat{\mu}_1$. Therefore, $\hat{\sigma}_1^2 \overset{a.s.}{\to} (\sigma_1^*)^2$. The same reasoning applies to $\hat{\sigma}_0^2$. □

Proof (Lemma 1): Let $s_{t,q}(\theta)$ be the qth element of $s_t(\theta)$ for $q = 1,2,3,4$. It follows that

$$\begin{aligned}
s_{t,1}(\theta^*) &= Z_t(\frac{\vartheta(Y_t)-\mu_1^*}{(\sigma_1^*)^2}) \\
s_{t,2}(\theta^*) &= Z_t(\frac{(\vartheta(Y_t)-\mu_1^*)^2}{2(\sigma_1^*)^4} - \frac{1}{2(\sigma_1^*)^2}) \\
s_{t,3}(\theta^*) &= (1-Z_t)(\frac{\vartheta(Y_t)-\mu_0^*}{(\sigma_0^*)^2}) \\
s_{t,4}(\theta^*) &= (1-Z_t)(\frac{(\vartheta(Y_t)-\mu_0^*)^2}{2(\sigma_0^*)^4} - \frac{1}{2(\sigma_0^*)^2}).
\end{aligned}$$

Further, let $a \equiv [r'/(r'-1)]$ be the largest integer less than or equal to $r'/(r'-1)$ and $b \equiv 2a/(a-1)$. By 1(ii), $r' > 1$, so $b > 2$. For any q and t, $\|s_{t,q}(\theta^*)\|_b \leq \|s_{t,q}(\theta^*)\|_{[b]+1} < \infty$, where $\|\cdot\|_b$ is the L_b norm on $(\Omega, \mathcal{F}, \mathcal{P})$. Here we use Jensen's inequality to derive the first inequality. Then,

$$|Cov(s_{1,i}(\theta^*),s_{1+m,j}(\theta^*))| \leq 2a(2\alpha_m)^{\frac{1}{a}}\|s_{1,i}(\theta^*)\|_b\|s_{1+m,j}(\theta^*)\|_b$$

$$= 2a(2\alpha_m)^{\frac{1}{a}}\|s_{1,i}(\theta^*)\|_b\|s_{1,j}(\theta^*)\|_b, \tag{9}$$

for any i,j and $m \in N$. The first line in (9) is Davydovs inequality, and the second is due to stationarity. Take summation on both sides of (9) over m to yield

$$\begin{aligned}\sum_{m=1}^{\infty}|Cov(s_{1,i}(\theta^*),s_{1+m,j}(\theta^*))| &\leq 2a(2)^{\frac{1}{a}}\|s_{1,i}(\theta^*)\|_b\|s_{1,j}(\theta^*)\|_b\sum_{m=1}^{\infty}\alpha_m^{\frac{1}{a}} \\ &= C\sum_{m=1}^{\infty}\alpha_m^{\frac{1}{a}} < \infty,\end{aligned}$$

where $C \equiv 2a(2)^{\frac{1}{a}}\|s_{1,i}(\theta^*)\|_b\|s_{1,j}(\theta^*)\|_b$. The finiteness of $\sum_{m=1}^{\infty}|Cov(s_{1,i}(\theta^*),s_{1+m,j}(\theta^*))|$ comes from two facts, that is, $C < \infty$ and $\sum_{m=1}^{\infty}\alpha_m^{\frac{1}{a}} < \infty$. Since i,j are arbitrary, we have verified that the autocovariance matrix $\Gamma_m \equiv Cov(s_t(\theta^*),s_{t+m}(\theta^*))$ is absolutely summable. Note that

$$\begin{aligned}I_T^* &= Var(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}s_t(\theta^*)) \\ &= \Gamma_0+\sum_{m=1}^{T-1}\frac{T-m}{T}(\Gamma_m+\Gamma_m').\end{aligned}$$

For any nonzero $\lambda \in R^4$,

$$\begin{aligned}\lambda' I_T^*\lambda &= \lambda'\Gamma_0\lambda+\sum_{m=1}^{T-1}\frac{T-m}{T}(\lambda'\Gamma_m\lambda+\lambda'\Gamma_m'\lambda) \\ &= \lambda'\Gamma_0\lambda+\sum_{m=1}^{T-1}(\lambda'\Gamma_m\lambda+\lambda'\Gamma_m'\lambda)-\frac{1}{T}\sum_{m=1}^{T-1}m(\lambda'\Gamma_m\lambda+\lambda'\Gamma_m'\lambda).\end{aligned}$$

Since $\sum_{m=1}^{\infty} |\lambda'\Gamma_m\lambda| < \infty$, dominated convergence implies that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{m=1}^{T-1} m\lambda'\Gamma_m\lambda = 0,$$
$$\lim_{T\to\infty} \frac{1}{T}\sum_{m=1}^{T-1} m\lambda'\Gamma_m'\lambda = 0.$$

Hence,

$$\lim_{T\to\infty} \lambda' I_T^* \lambda = \lambda'\Gamma_0\lambda + \sum_{m=1}^{\infty} (\lambda'\Gamma_m\lambda + \lambda'\Gamma_m'\lambda) = \lambda' I^* \lambda < \infty,$$

where $I^* \equiv \Gamma_0 + \sum_{m=1}^{\infty}(\Gamma_m + \Gamma_m')$. Since λ is arbitrarily chosen, we have $I_T^* \to I^*$. Obviously, I^* is symmetric. By Assumption 2, I_T^* is positive definite for each $T \in N$, implying $\lambda' I_T^* \lambda > 0$. Therefore, $\lambda' I^* \lambda \geq 0$ for any nonzero $\lambda \in R^4$. I^* is positive semidefinite as a result. Moreover, $|I_T^*| > \varepsilon$ for all $T > N(\varepsilon)$, so $|I^*| \geq \varepsilon > 0$ and I^* is positive definite, which completes the proof. □

Proof (Theorem 2): Define $\tilde{\theta}$ as

$$\left(\frac{\sum_{t=1}^T Z_t \vartheta(Y_t)}{\sum_{t=1}^T Z_t}, \frac{\sum_{t=1}^T Z_t(\vartheta(Y_t)-\mu_1^*)^2}{\sum_{t=1}^T Z_t}, \frac{\sum_{t=1}^T (1-Z_t)\vartheta(Y_t)}{\sum_{t=1}^T 1-Z_t}, \frac{\sum_{t=1}^T (1-Z_t)(\vartheta(Y_t)-\mu_0^*)^2}{\sum_{t=1}^T 1-Z_t}\right)'.$$

Suppose $\sqrt{T}(\tilde{\theta}-\theta^*) \xrightarrow{d} N(0, J^{*-1} I^* J^{*-1})$. If we can show $\sqrt{T}(\hat{\theta}-\tilde{\theta}) = o_p(1)$, then the conclusion of Theorem 2 follows as a result of asymptotic equivalence lemma.

Note that

$$\begin{aligned} & \sqrt{T}\frac{\sum_{t=1}^T Z_t((\vartheta(Y_t)-\hat{\mu}_1)^2 - (\vartheta(Y_t)-\mu_1^*)^2)}{\sum_{t=1}^T Z_t} \\ = \quad & \sqrt{T}(\mu_1^*-\hat{\mu}_1)^2 + \frac{2\sum_{t=1}^T Z_t(\vartheta(Y_t)-\mu_1^*)/\sqrt{T}}{\sum_{t=1}^T Z_t/T}(\mu_1^*-\hat{\mu}_1). \end{aligned}$$

Since $\sum_{t=1}^T Z_t(\vartheta(Y_t)-\mu_1^*)/\sqrt{T} = O_p(1)$ by central limit theorem, the second term is $o_p(1)$. $(\mu_1^*-\hat{\mu}_1)^2$ is $O_p(1/T)$ by central limit theorem and continuous mapping theorem, which implies the first term is also $o_p(1)$. Similarly, we can show $\sqrt{T}\sum_{t=1}^T(1-Z_t)((\vartheta(Y_t)-\hat{\mu}_0)^2 - (\vartheta(Y_t)-\mu_0^*)^2)/\sum_{t=1}^T(1-Z_t) = o_p(1)$. It then follows that $\sqrt{T}(\hat{\theta}-\tilde{\theta}) = o_p(1)$.

Define Π_T to be

$$\begin{pmatrix} \frac{\pi^* T}{\sum_{t=1}^T Z_t} & 0 & 0 & 0 \\ 0 & \frac{\pi^* T}{\sum_{t=1}^T Z_t} & 0 & 0 \\ 0 & 0 & \frac{(1-\pi^*)T}{\sum_{t=1}^T (1-Z_t)} & 0 \\ 0 & 0 & 0 & \frac{(1-\pi^*)T}{\sum_{t=1}^T (1-Z_t)} \end{pmatrix}.$$

It remains to show $\sqrt{T}(\tilde{\theta}-\theta^*) \xrightarrow{d} N(0, J^{*-1}I^*J^{*-1})$. This is true because $\sqrt{T}(\tilde{\theta}-\theta^*) = \Pi_T J^{*-1}\sum_{t=1}^T s_t(\theta^*)/\sqrt{T}$, which converges to $N(0, J^{*-1}I^*J^{*-1})$ in distribution by Slutsky's theorem and central limit theorem for mixing sequences (cf. Wooldridge (1986)). □

Proof (Theorem 3): Denote $[c-d, c+d]$ as $c \pm d$ for two real scalars c and d.

$$\begin{aligned} & \mathcal{P}(H^*(\eta) \in \Phi(k_1(\eta;\hat{\theta}) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}'/T})) \\ = \;& \mathcal{P}(k_1(\eta;\theta^*) \in k_1(\eta;\hat{\theta}) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}'/T}) \\ = \;& \mathcal{P}(k_1(\eta;\hat{\theta}) - k_1(\eta;\theta^*) \in \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}'/T}) \\ = \;& \mathcal{P}(\sqrt{T}(k_1(\eta;\hat{\theta}) - k_1(\eta;\theta^*)) \in \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}'}), \end{aligned}$$

which converges to $1-\alpha$ since

$$\sqrt{T}(k_1(\eta;\hat{\theta}) - k_1(\eta;\theta^*)) \xrightarrow{d} N(0, \frac{\partial k_1(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k_1(\eta;\theta^*)}{\partial\theta}')$$

by the delta method. (4b) can be proved in the same way. Analogously,

$$\sqrt{T}(k'(\eta;\hat{\theta}) - k'(\eta;\theta^*)) \xrightarrow{d} N(0, \frac{\partial k'(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k'(\eta;\theta^*)}{\partial\theta}').$$

We have

$$T(k'(\eta;\hat{\theta}) - k'(\eta;\theta^*))'(\frac{\partial k'(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k'(\eta;\theta^*)}{\partial\theta}')^{-1}(k'(\eta;\hat{\theta}) - k'(\eta;\theta^*)) \xrightarrow{d} \chi^2(2),$$

where $\chi^2(2)$ is a random variable with chi-squared distribution of 2 degrees of freedom. Recall that $\Gamma(\eta,\alpha)$ is the set defined as

$$\{O \in R^2 : T(k'(\eta;\hat{\theta}) - O)'(\frac{\partial k'(\eta;\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k'(\eta;\theta^*)}{\partial\theta}')^{-1}(k'(\eta;\hat{\theta}) - O) \le \chi^2_\alpha(2)\}.$$

It follows that

$$\begin{aligned} & \mathcal{P}((H^*(\eta),F^*(\eta))' \in \Phi(\Gamma(\eta,\alpha))) \\ = \ & \mathcal{P}(\Phi(k'(\eta;\theta^*)) \in \Phi(\Gamma(\eta,\alpha))) \\ = \ & \mathcal{P}(k'(\eta;\theta^*) \in \Gamma(\eta,\alpha)), \end{aligned}$$

which converges to $1-\alpha$ by the preceding argument and (4c) holds. □

Proof (Theorem 4): (6a) holds by the same argument as in the proof of Theorem 3. To show (6b), applying the delta method again, we have

$$\sqrt{T}(k(\hat{\theta}) - k(\theta^*)) \xrightarrow{d} W \equiv (W_1,W_2)', \tag{10}$$

where

$$W \sim N(0, \frac{\partial k(\theta^*)}{\partial\theta}J^{*-1}I^*J^{*-1}\frac{\partial k(\theta^*)}{\partial\theta}').$$

Consider the map $F_1 : (o_1,o_2)' \in R^2 \mapsto o_1 + o_2\Phi^{-1}(x) \in C([a,b])$, where $C([a,b])$ is the set of all continuous real-valued functions on $[a,b]$. We want to check that $F_1(\cdot)$ as a map between two normed spaces is continuous and linear. For any two vectors $\tilde{o} \equiv (\tilde{o}_1,\tilde{o}_2)'$, $\bar{o} \equiv (\bar{o}_1,\bar{o}_2)'$ in R^2 and any real scalar α, we have

$$\begin{aligned} F_1(\tilde{o}+\bar{o}) &= \tilde{o}_1 + \bar{o}_1 + (\tilde{o}_2+\bar{o}_2)\Phi^{-1}(x) \\ &= \tilde{o}_1 + \tilde{o}_2\Phi^{-1}(x) + \bar{o}_1 + \bar{o}_2\Phi^{-1}(x) \\ &= F_1(\tilde{o}) + F_1(\bar{o}), \end{aligned}$$

and

$$\begin{aligned} F_1(\alpha\tilde{o}) &= \alpha\tilde{o}_1 + \alpha\tilde{o}_2\Phi^{-1}(x) \\ &= \alpha(\tilde{o}_1 + \tilde{o}_2\Phi^{-1}(x)) \\ &= \alpha F_1(\tilde{o}), \end{aligned}$$

establishing linearity. For continuity, we have to equip $C([a,b])$ with some appropriate norm. As usual, we adopt the supremum norm $\|f\|_\infty \equiv \sup_{x\in[a,b]}|f(x)|$. For the previously defined $\tilde{o}$ and $\bar{o}$,

$$\begin{aligned} 0 \leq \|F_1(\tilde{o}) - F_1(\bar{o})\|_\infty &= \sup_{x\in[a,b]} |\tilde{o}_1 - \bar{o}_1 + (\tilde{o}_2 - \bar{o}_2)\Phi^{-1}(x)| \\ &\leq |\tilde{o}_1 - \bar{o}_1| + |\tilde{o}_2 - \bar{o}_2| \sup_{x\in[a,b]} |\Phi^{-1}(x)| \\ &= |\tilde{o}_1 - \bar{o}_1| + |\tilde{o}_2 - \bar{o}_2| K(a,b) \to 0, \end{aligned}$$

as $\tilde{o} \to \bar{o}$, since $K(a,b) \equiv \sup_{x\in[a,b]} |\Phi^{-1}(x)|$ is a finite constant. By arbitrariness of $\bar{o}$, $F_1(\cdot)$ is continuous on R^2. In order to apply the functional delta method to (10), $F_1(\cdot)$ must be Hadamard-differentiable at $k(\theta^*) \in R^2$. For this purpose, let $\{t_n\}$ and $\{h_n\}$ be any two converging sequences such that $t_n \to 0 \in R$ and $h_n \to h \in R^2$ as $n \to \infty$. We have

$$\begin{aligned} \frac{F_1(k(\theta^*) + t_n h_n) - F_1(k(\theta^*))}{t_n} &= \frac{F_1(k(\theta^*)) + t_n F_1(h_n) - F_1(k(\theta^*))}{t_n} \\ &= F_1(h_n), \end{aligned}$$

which converges to $F_1(h)$ by the continuity of $F_1(\cdot)$. Thus, $F_1(\cdot)$ is Hadamard-differentiable at $k(\theta^*)$ tangentially to R^2. According to the functional delta method (cf. Kosorok (2008)),

$$\sqrt{T}(F_1(k(\hat{\theta})) - F_1(k(\theta^*))) \Longrightarrow F_1(W),$$

where $\Longrightarrow$ stands for weak convergence. Moreover, the continuous mapping theorem implies that

$$\sqrt{T}\|F_1(k(\hat{\theta})) - F_1(k(\theta^*))\|_\infty \Longrightarrow \|F_1(W)\|_\infty,$$

or

$$\sqrt{T}\sup_{x\in[a,b]}|k_3(x;\hat{\theta}) - k_3(x;\theta^*)| \xrightarrow{d} \sup_{x\in[a,b]}|W_1 + W_2\Phi^{-1}(x)|. \tag{11}$$

Without loss of generality, let $W_2 \neq 0$. The maximizer x^* for $|W_1 + W_2\Phi^{-1}(x)|$ over $[a,b]$ satisfies

$$x^* = \begin{cases} a, & \text{if } -\frac{W_1}{W_2} \geq \frac{\Phi^{-1}(b)-\Phi^{-1}(a)}{2} + \Phi^{-1}(a); \\ b, & \text{otherwise.} \end{cases}$$

Therefore $\sup_{x\in[a,b]}|W_1 + W_2\Phi^{-1}(x)| = \max\{|W_1 + W_2\Phi^{-1}(a)|, |W_1 + W_2\Phi^{-1}(b)|\}$. Observe that

$$A(a,b)W \sim N(0,\Sigma(a,b)).$$

We have

$$\begin{aligned} & \mathcal{P}(\sup_{x\in[a,b]}|W_1 + W_2\Phi^{-1}(x)| \leq u) \\ = \ & \mathcal{P}(\max\{|W_1 + W_2\Phi^{-1}(a)|, |W_1 + W_2\Phi^{-1}(b)|\} \leq u) \\ = \ & \mathcal{P}(|W_1 + W_2\Phi^{-1}(a)| \leq u, |W_1 + W_2\Phi^{-1}(b)| \leq u) \\ = \ & \mathcal{P}(-u \leq W_1 + W_2\Phi^{-1}(a) \leq u, -u \leq W_1 + W_2\Phi^{-1}(b) \leq u) \\ = \ & F_{sup}(u;a,b). \end{aligned}$$

Therefore, $F_{sup}(\cdot; a, b)$ is the distribution function of $\sup_{x\in[a,b]} |W_1 + W_2\Phi^{-1}(x)|$. Finally,

$$\begin{aligned}
& \mathcal{P}(\forall x \in [a,b], y^*(x) \in \Phi(k_3(x;\hat{\theta}) \pm \frac{f_\alpha}{\sqrt{T}})) \\
= \; & \mathcal{P}(\forall x \in [a,b], k_3(x;\theta^*) \in k_3(x;\hat{\theta}) \pm \frac{f_\alpha}{\sqrt{T}}) \\
= \; & \mathcal{P}(\forall x \in [a,b], |k_3(x;\hat{\theta}) - k_3(x;\theta^*)| \leq \frac{f_\alpha}{\sqrt{T}}) \\
= \; & \mathcal{P}(\sqrt{T} \sup_{x\in[a,b]} |k_3(x;\hat{\theta}) - k_3(x;\theta^*)| \leq f_\alpha),
\end{aligned}$$

which converges to $1-\alpha$ by (11). (The measurability of $\sup_{x\in[a,b]} |k_3(x;\hat{\theta}) - k_3(x;\theta^*)|$ can be argued from Appendix C of Pollard (1984)). □

Proof (Theorem 5): (8) is a trivial application of the delta method. □