*Econometric Theory*, **32**, 2016, 122–153.
doi:10.1017/S0266466614000802

# MODEL-FREE INFERENCE FOR TAIL RISK MEASURES

KE-LI XU
*Texas A&M University*

Understanding uncertainty in estimating risk measures is important in modern financial risk management. In this paper we consider a nonparametric framework that incorporates auxiliary information available in covariates and propose a family of inferential methods for the value at risk, expected shortfall, and related risk measures. A two-step generalized empirical likelihood test statistic is constructed and is shown to be asymptotically pivotal without requiring variance estimation. We also show its validity when applied to a semiparametric index model. Asymptotic theories are established allowing for serially dependent data. Simulations and an empirical application to Canadian stock return index illustrate the finite sample behavior of the methodologies proposed.

## 1. INTRODUCTION

Since its introduction at J.P. Morgan about two decades ago, the value at risk (VaR) has become the benchmark for managing financial risk in most financial institutions. It measures the severity of the risk of holding a portfolio of risky assets over a time period, and it extends the concept of the maximum possible loss by giving an upper quantile of the loss distribution which is not exceeded with a given high probability. However, as pointed out by Artzner, Delbaen, Eber, and Heath (1999) and Acerbi and Tasche (2002), the VaR disregards information beyond the quantile reported; e.g., it is not sensitive to the catastrophic loss incurred by a damage event. Moreover, the VaR does not satisfy the subadditivity axiom and is therefore not a coherent risk measure. This undesirable property implies that a portfolio can be riskier, when measured by VaR, than its individual components and thus discourages diversification. The expected shortfall, which is defined as the expected loss given that the loss exceeds the VaR, provides a robust alternative as a coherent measure of financial risk and has gained increasing popularity among practitioners. Basak and Shapiro (2001) and Yamai and Yoshiba (2005) showed that utility-maximizing investors with expected shortfall

The author thanks Tim Bollerslev, Oliver Linton, Taisuke Otsu, Peter Phillips, Kenneth Wolpin, anonymous referees, and participants of the third CIREQ time series conference at Montreal for suggestion and comments. The author is grateful to University of Alberta School of Business for partial financial support through Canadian Utilities faculty award and Killam research grant. Address correspondence to Ke-Li Xu, Department of Economics, Texas A&M University, 3063 Allen Building, 4228 TAMU, College Station, Texas 77843-4228, USA; e-mail: keli.xu@tamu.edu.

 © Cambridge University Press 2014

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

constraint tend to optimally construct more invulnerable positions than those with VaR constraint.

This paper studies statistical inference for the expected shortfall (hereinafter, ES) in a nonparametric framework and focuses on the cases when there is auxiliary information available in covariates. Most extant methods of estimating the ES strongly depend on various parametric specification. For example, a widely used approach amounts to postulating the return process as a linear or nonlinear ARCH/GARCH volatility process with a multiplicative error term. Depending on the treatment of the error distribution as known or unknown, the ES is then estimated by the direct approach or extreme value theory (McNeil, Frey, and Embrechts, 2005; Tsay, 2005, Chap. 7). Komunjer (2007) constructed a parametric ES estimator based on a flexible family of asymmetric power distributions. Such volatility-based methods assume the central role of the volatility played in the return dynamics and are challenged for potential model misspecification which may lead to substantial estimation inaccuracy for risk measures (Mittnik and Rachev, 2000; Chan, Deng, Peng, and Xia, 2007, among others). In contrast, nonparametric methods (Scaillet, 2004, 2005; Cai and Wang, 2008; Chen, 2008; Linton and Xiao, 2013) do not rely on any particular formulation of the return process. They directly explore the unconditional/conditional expectation representation of the ES and can accommodate return dynamics in a flexible way given the covariates that are appropriately chosen.

The current work contributes to the growing literature on nonparametric modeling of ES by providing a convenient and reliable procedure of conducting inference, i.e., constructing a confidence interval or testing a hypothesis on the expected shortfall. This problem is important since it offers information on a safety margin of the risk measure with a specified confidence level and is crucial for decision making of financial managers, as argued by Christoffersen and Gonçalves (2005) and Chan et al. (2007).[1] While the papers cited above are mainly restricted to the parametric volatility-based framework, the estimation risk is much less addressed in a nonparametric setting. Komunjer (2007) noticed that "Despite a strong appeal of the semi- and nonparametric methods . . . their difficulty lies in the ability to estimate the expected shortfall variance" and pointed out that the difficulty is inherited from a similar situation arising in quantile regression.

We aim to close the gap in the literature and propose a general procedure based on empirical likelihood (Owen, 1988, 1990; Chen and Qin, 2000) for a family of nonparametric kernel-weighed local polynomial estimators. We also demonstrate its validity when the new method is applied to a semiparametric index model, which is useful with multiple covariates and still achieves the one-dimensional convergence rate. Empirical likelihood (EL) has attracted substantial attention in statistics and econometrics over the past decades[2] for its ability to construct test statistics in which the studentization is achieved automatically; no explicit variance estimation is needed. Under mild conditions the EL test statistic obeys a nonparametric version of Wilks theorem with a chi-square limit distribution and

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

the associated confidence intervals are flexibly shaped by data. These properties motivate us to extend the EL to conduct inference for the expected shortfall.

As is customary in the EL based method, our test statistic is constructed via estimating equations. However, the estimating equation for ES that is obtained from the local polynomial estimator cannot be directly used. It involves an unknown quantity, the VaR, which also has to be estimated. We bring in the estimating equation for the VaR and consider first a joint empirical likelihood test statistic for VaR and ES. This statistic is of independent interest and it can be used to perform simultaneous inference for VaR and ES. In the second step the test statistic for ES is formed by profiling out the nuisance parameter. We show that the methodology is also applicable to other risk measures that involve both VaR and ES, e.g., the mean excess loss, which is defined as the difference between ES and VaR.

Following Baggerly (1998), Corcoran (1998), and Kitamura (2006), we regard the EL as defining a divergence measure from the underlying estimator and consider a general framework of generalized empirical likelihood (GEL) (Smith, 1997; Newey and Smith, 2004) which nests as special cases the EL and other popular divergence measures such as maximum entropy and euclidean likelihood as well. The asymptotic theories are developed allowing for general dynamic dependence in data.

The empirical likelihood method can be also applied to unconditional nonparametric model that does not use covariates, but certain modification is needed to allow for serially dependent data. It is relevant in applications to financial data, e.g., to capture the effects of autoregressive conditional heteroskedasticity. In contrast, EL works directly in conditional models under weak dependence, thanks to the nature of spatial smoothing that generally weakens serial dependence (Hart, 1996; Fan and Yao, 2003). This observation motivates the recent work by Xu (2009) and Chan, Peng, and Zhang (2011) who also considered EL in conditional time series models within different contexts.

The paper is organized as follows. In Section 2 we introduce nonparametric estimators of the VaR, the expected shortfall, and other related quantities. A general inference procedure is proposed in Section 3 and is applied to several risk measures. Asymptotic theories are also presented. A semiparametric index model is considered in Section 4. Simulations and an empirical application to Canadian stock return index are reported in Sections 5 and 6. Section 7 concludes. All proofs are collected in the Appendix.

## 2. THE MODEL AND THE ESTIMATOR

### 2.1. Risk Measures

Let $Y_i$ be the return process on a financial asset or portfolio over a prescribed holding period. Suppose we are interested in estimating conditional value at risk (VaR) and conditional expected shortfall (ES) given $X_i = x$ and a probability value $\tau$

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

(usually a small number, e.g., 1% or 5%).[3] The explanatory variable $X_i$ contains information that can be utilized to estimate or forecast some aspects of interest of the distribution of $Y_i$. It typically includes (but not necessarily) the lags of $Y_i$.[4] Let $F(\cdot|x)$ and $f(\cdot|x)$ be the conditional cumulative distribution function (CDF) and conditional density function, respectively, of $Y_i$ given $X_i = x$. We focus on the case when $X_i$ is univariate in what follows and consider a semiparametric index model that is useful when $X_i$ is multivariate.

Define $\beta_0(x,\tau) = F^{-1}(\tau|x) = \inf\{y : F(y|x) \geq \tau\}$ and $\theta_0(x,\tau) = \mathrm{E}(Y_i|Y_i \leq \beta_0, X_i = x)$. In practice, the VaR and the ES of holding the asset or portfolio under investigation are usually cited as the negatives of $\beta_0(x,\tau)$ and $\theta_0(x,\tau)$ respectively. By defining $Y_i$ as returns (instead of losses), we simply refer to $\beta_0(x,\tau)$ and $\theta_0(x,\tau)$ as VaR and ES respectively in what follows and take the negative signs as meaning of losses. For ease of exposition, we omit the explicit references to $x$ and $\tau$ and write $\beta_0(x,\tau)$ and $\theta_0(x,\tau)$ simply as $\beta_0$ and $\theta_0$ respectively. ES assigns equal weights to the losses larger than a prespecified level and is thus sensitive to catastrophic events, unlike VaR which depends only on the relative frequencies (not the levels) of extreme realizations.[5] Ruszczyski and Vanderbei (2003) showed that replacing the quadratic risk measure in the classical Markowitz criteria with an ES-type risk will produce efficient frontiers of stochastically nondominated portfolios, a property not shared by a classical Markowitz model.

To estimate the ES, we rewrite it as

$$\theta_0 = \tau^{-1}\int_{-\infty}^{\beta_0} yf(y|x)dy = \tau^{-1}\mathrm{E}(Y_i\mathbb{I}(Y_i \leq \beta_0)|X_i = x),$$

where $\mathbb{I}(\cdot)$ is the indicator function. If $f$ is Gaussian with mean $m(x)$ and variance $\sigma^2(x)$, then $\theta_0 = m(x) - \sigma(x)\phi(\Phi^{-1}(\tau))/\tau$ where $\phi$ and $\Phi$ respectively are the density and distribution functions of a standard normal variable.

If $\beta_0$ is known, the problem of estimating $\theta_0$ is just estimation of a conditional expectation, and the nonparametric method like kernel smoothing can be used. Lack of knowledge of $\beta_0$ poses little problem for point estimation since plugging in a consistent estimator of $\beta_0$ would be a handy solution. However, it makes inference of $\theta_0$ more of an issue since estimation risk (namely the standard error of the estimator) has to be determined, which is usually nontrivial.

## 2.2. Local Polynomial Estimation

Given an estimator of $\beta_0$, say $\widehat{\beta}$, we consider the following kernel-weighted local $p$-th polynomial estimator (Fan and Yao, 2003) of $\theta_0$, defined as $\widehat{\theta} = \widehat{a}_0$, where

$$\{\widehat{a}_j, 0 \leq j \leq p\} = \min_{\{a_j:0\leq j\leq p\}} \sum_{i=1}^{n}\left(\frac{1}{\tau}Y_i G\left(\frac{\widehat{\beta}-Y_i}{h_0}\right) - \sum_{j=0}^{p} a_j(X_i - x)^j\right)^2 K_h(X_i - x). \quad (1)$$

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

It is a nonparametric estimator in the sense that it does not require global functional specification and it fits the data locally by a polynomial. Here $G$ is a user-specified distribution function and $K_h(X_i - x) = h^{-1}K((X_i - x)/h)$, for a kernel function $K$ and smoothing bandwidths $h$ and $h_0$. When $h_0$ is close to zero, $Y_i G((\widehat{\beta} - Y_i)/h_0)/\tau$ approaches $Y_i \mathbb{I}(Y_i \leq \widehat{\beta})/\tau$ which seems more natural to be used in the conditional expectation estimation (1). We introduce a bandwidth parameter $h_0$ for technical reasons to obtain smooth estimating equations that facilitate the profiling-out procedure in constructing the empirical likelihood (cf. Section 3.3).[6]

The local polynomial estimator $\widehat{\theta}$ can be expressed in closed form. Define

$$\widehat{S}_k = n^{-1}\sum_{i=1}^{n}[(X_i - x)/h]^k K_h(X_i - x),$$

$k = 0, 1, \ldots, 2p$, and $\widehat{S}$ is the $(p+1)$ by $(p+1)$ matrix with the $(i, j)$-th element $\widehat{S}_{i+j-2}$. Let $e_1$ be the $(p+1)$-dimension vector $(1, 0, \ldots, 0)^\top$. Then $\widehat{\theta}$ can be written as

$$\widehat{\theta} = \frac{1}{nh}\sum_{i=1}^{n} W_p((X_i - x)/h) Y_i G((\widehat{\beta} - Y_i)/h_0)/\tau, \tag{2}$$

where $W_p(u) = e_1^\top H^{-1}\widehat{S}^{-1}H^{-1}(1, uh, \ldots, (uh)^p)^\top K(u)$ with $H = diag(1, h, \ldots, h^p)$ satisfying

$$\frac{1}{nh}\sum_{i=1}^{n} W_p((X_i - x)/h) = 1.$$

For example, when $p = 0$, $\widehat{\theta}$ is the Nadaraya–Watson or local constant estimator considered by Scaillet (2005) with $W_p^{NW}((X_i - x)/h) = [(nh)^{-1}\sum_{i=1}^{n} K((X_i - x)/h)]^{-1} \cdot K((X_i - x)/h)$. When $p = 1$, $\widehat{\theta}$ is the popular local linear estimator with $W_p^{LL}((X_i - x)/h) = [(nh)^{-1}\sum_{i=1}^{n} K((X_i - x)/h)(M_{n,2} - ((X_i - x)/h)M_{n,1})]^{-1} \cdot K((X_i - x)/h)[M_{n,2} - ((X_i - x)/h)M_{n,1}]$, where $M_{n,j} = (nh)^{-1}\sum_{i=1}^{n}[(X_i - x)/h]^j K((X_i - x)/h)$ for $j = 1, 2$.

Implementation of $\widehat{\theta}$ depends on an estimator of the conditional VaR. We consider the local polynomial estimator $\widehat{\beta}$ which solves[7]

$$\frac{1}{nh}\sum_{i=1}^{n} W_p((X_i - x)/h) G((\widehat{\beta} - Y_i)/h_0) = \tau. \tag{3}$$

The local polynomial estimator (especially the local linear estimator) enjoys superior minimax efficiency among the class of linear smoothers that also includes orthogonal series estimators and spline smoothers (Fan and Gijbels, 1996, Sect. 3.4). Both estimators $\widehat{\beta}$ and $\widehat{\theta}$ are consistent, which follows from the standard limit

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

theory for the local polynomial estimator and the uniform law of large numbers. The plug-in effect of $\widehat{\beta}$ in the distribution of $\widehat{\theta}$ is nontrivial and generally leads to a larger standard error than that of the oracle estimator of $\theta_0$ which assumes the true value $\beta_0$. The inference procedure considered in the next section can automatically capture such an effect.

An alternative estimator of $\beta_0$ is based on minimizing a kernel-weighted piecewise linear loss function and is first-order asymptotically equivalent to $\widehat{\beta}$ in (3). Bassett, Koenker, and Kordas (2004) showed that ES can then be estimated by the difference between the above minimum (scaled by $\tau^{-1}$) and the average return.

We will also consider another closely related risk measure, the mean excess loss $\kappa_0 = \beta_0 - \theta_0$, which is defined as the difference between the expected shortfall and the VaR. It is the expectation of the excess loss function which gauges the probability of suffering the loss between $\beta_0$ and $\beta_0 - \varepsilon$ for a positive value of $\varepsilon$. The mean excess loss can be estimated as $\widehat{\kappa} = \widehat{\beta} - \widehat{\theta}$.

## 3. INFERENCE

### 3.1. The Joint Generalized Empirical Likelihood (GEL)

In this section we introduce a novel procedure to draw statistical inference for risk measures in a nonparametric framework. Suppose that the base measures VaR $\beta_0$ and ES $\theta_0$ are estimated by local polynomial estimators $\widehat{\beta}$ and $\widehat{\theta}$ defined by (3) and (2) respectively. We treat $\widehat{\beta}$ and $\widehat{\theta}$ as maximum empirical likelihood (EL) estimators of Owen (1988) and Qin and Lawless (1994), i.e., $(\widehat{\beta}, \widehat{\theta}) = \arg\max_{(\beta,\theta)\in\mathcal{B}\times\Theta\subset\mathbb{R}^2} \overline{L}(\beta, \theta)$, where $\mathcal{B}$ and $\Theta$ are compact sets in $\mathbb{R}^1$. The empirical likelihood $\overline{L}(\beta, \theta)$ solves the constrained optimization problem

$$\overline{L}(\beta, \theta) = \max_{\{\pi_i : 1\le i\le n\}} \prod_{i=1}^{n} \pi_i \tag{4}$$

subject to

$$\sum_{i=1}^{n} \pi_i W_p((X_i - x)/h)[G((\beta - Y_i)/h_0) - \tau] = 0 \text{ (VaR)}, \tag{5}$$

$$\sum_{i=1}^{n} \pi_i W_p((X_i - x)/h)[Y_i G((\beta - Y_i)/h_0) - \tau\theta] = 0 \text{ (ES)}, \tag{6}$$

and

$$\sum_{i=1}^{n} \pi_i = 1. \tag{7}$$

Optimization restrictions (5) and (6) are for VaR and ES respectively. It is easy to see that $\widehat{\beta}$ and $\widehat{\theta}$ maximize $\overline{L}(\beta, \theta)$ since the maximization problem (4), subject

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

to (7) only, is solved at $\pi_i = 1/n$ for all $i$, and by definition $\widehat{\beta}$ and $\widehat{\theta}$ satisfy the restrictions (5) and (6) when $\pi_i = 1/n$.

Let $m_i(\beta,\theta) = (W_p((X_i - x)/h)[G((\beta - Y_i)/h_0) - \tau], W_p((X_i - x)/h)[Y_i G((\beta - Y_i)/h_0) - \tau\theta])^\top$, then the restrictions (5) and (6) reduce to $\sum_{i=1}^{n} \pi_i m_i(\beta,\theta) = 0$. For given values of $\beta$ and $\theta$, by solving (4) subject to restrictions (5)−(7), $\overline{L}(\beta,\theta)$ is maximized at $\widehat{\pi}_i = [n(1 + \lambda^\top m_i(\beta,\theta))]^{-1}$, where $\lambda$ is the Lagrange multiplier of (5) and (6) satisfying $\sum_{i=1}^{n}[1 + \lambda^\top m_i(\beta,\theta)]^{-1} m_i(\beta,\theta) = 0$. Our test statistic is then defined as

$$\mathcal{L}(\beta,\theta) = -2\log\left(\overline{L}(\beta,\theta)/\overline{L}(\widehat{\beta},\widehat{\theta})\right) = -2\log\left(\overline{L}(\beta,\theta)/\prod_{i=1}^{n}(1/n)\right) \tag{8}$$

$$= \min_{\{\pi_i : 1 \le i \le n\}} -2\sum_{i=1}^{n}\log(n\pi_i), \text{ subject to (5)−(7)},$$

which can be interpreted as $-2$ times the logarithm of the nonparametric likelihood ratio (empirical likelihood ratio). Just like in the parametric framework, the likelihood ratio statistic $\mathcal{L}(\beta,\theta)$ rejects parameter values at which likelihood ratios are large. It is based on a nonparametric likelihood in the sense that we assume no functional distribution on the underlying unobservable data generating process and instead assign multinomial weights to observations.

Written in terms of the Lagrange multiplier, the likelihood ratio becomes $\mathcal{L}(\beta,\theta) = -2\sum_{i=1}^{n}\log(n\widehat{\pi}_i) = 2\sum_{i=1}^{n}\log[1 + \lambda^\top m_i(\beta,\theta)]$. The $n$-variable optimization problem in (8) can then be reduced to a two-dimensional problem by using its dual form. Since $\partial\mathcal{L}(\beta,\theta)/\partial\lambda = 0$ and $\partial^2\mathcal{L}(\beta,\theta)/\partial\lambda\partial\lambda^\top = -\sum_{i=1}^{n} 2[1 + \lambda^\top m_i(\beta,\theta)]^{-2} m_i(\beta,\theta) m_i(\beta,\theta)^\top$, which is negative semidefinite, $\mathcal{L}(\beta,\theta)$ can be rewritten as

$$\mathcal{L}(\beta,\theta) = \sup_{\lambda \in \Lambda_n(\beta,\theta)} 2\left(\sum_{i=1}^{n}\log[1 + \lambda^\top m_i(\beta,\theta)]\right), \tag{9}$$

where $\Lambda_n(\beta,\theta) = \{\lambda \in \mathbb{R}^2 : \lambda^\top m_i(\beta,\theta) \in J \text{ for } 1 \le i \le n\}$ and $J$ is an open interval containing 0. In what follows we focus on the dual representation (9), which is computationally more attractive than the primal problem (8).

By introducing a nonparametric likelihood, empirical likelihood delivers tests and confidence regions (based on the log EL ratio $\mathcal{L}(\beta,\theta)$) in an analogous way to a parametric likelihood. More generally, we shall consider the following test statistic

$$\mathcal{L}(\beta,\theta) = \sup_{\lambda \in \Lambda_n(\beta,\theta)} 2\left[\sum_{i=1}^{n}\rho(\lambda^\top m_i(\beta,\theta)) - n\rho(0)\right], \tag{10}$$

where $\rho(v)$ is a user-chosen function. We assume that $\rho$ is twice continuously differentiable in a neighborhood of zero such that $\dot{\rho}(0) = -1$ and $\ddot{\rho}(0) = -1$, where $\dot{\rho}$ and $\ddot{\rho}$ are the first and second derivatives of $\rho$. The statistic in (10) reduces to

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

the one in (9) when $\rho(v) = \log(1 - v)$. Without stipulating $\rho$ as of logarithmic functional form as in case of EL, the parameter values $\beta$ and $\theta$ that minimize $\mathcal{L}(\beta, \theta)$ in (10) are the generalized empirical likelihood (GEL) estimators introduced by Smith (1997) and Newey and Smith (2004). To motivate the statistic $\mathcal{L}(\beta, \theta)$ in (10), note that besides the log EL ratio interpretation, $\mathcal{L}(\beta, \theta)$ in (8) can be alternatively viewed as Kullback–Leibler divergence from the probability distribution $\pi_i$ of the uniform distribution with equal weights $1/n$, and $\widehat{\beta}$ and $\widehat{\theta}$ (which maximize $\overline{L}(\beta, \theta)$ therefore minimize $\mathcal{L}(\beta, \theta)$) are the corresponding minimum discrepancy estimators (e.g., Baggerly, 1998; Corcoran, 1998; Kitamura, 2006). By considering a general function $\rho$, we allow for the Kullback–Leibler divergence as well as other divergence measures, e.g., maximum entropy, Neyman-modified chi-square statistic (or euclidean likelihood), Hellinger divergence, or others in the power divergence family of Cressie and Read (1984) (see Newey and Smith, 2004, for the corresponding choices of $\rho$).

### 3.2. Asymptotic Theory

We make the following assumptions throughout the paper. Let $C$ be a generic finite constant.

**Assumption.**

A1. The kernel $K$ is a density function on a bounded support $[-1,1]$ satisfying $\int uK(u)du = 0$ and $\int u^2 K(u)du > 0$.

A2. Let $f_{X_1,X_l}(x_1, x_l)$ be the joint density of $X_1$ and $X_l$ for any $l \geq 2$. Assume that $f_{X_1,X_l}(x_1, x_l) \leq C$ and $\mathrm{E}(Y_1^2 + Y_l^2 | X_1 = x_1, X_l = x_l) \leq C$ for $x_1$ and $x_l$ in a neighborhood of $x$.

A3. The process $\{X_i, Y_i\}$ is stationary $\alpha$-mixing, with mixing coefficients $\alpha(l)$, $\sum_{l=1}^{\infty} l^a [\alpha(l)]^{1-2/\delta} < \infty$, $\mathrm{E}(|Y_1|^\delta | X_1 = x_1) \leq C$ for $x_1$ in a neighborhood of $x$, $\delta > 2$ and $a > 1 - 2/\delta$.

A4. For the $\alpha$-mixing process, we assume that there is a sequence of positive integers $s_n$ such that $s_n \to \infty$, $s_n = o(\sqrt{nh})$, and $\sqrt{n/h}\alpha(s_n) \to 0$, as $n \to \infty$.

A5. Let $\mu(\cdot)$ be the marginal density function of $X_i$. Assume that $F(y|\cdot)$ and $\mu(\cdot)$ are continuous at $x$, for any $y$, and $\mu(x) > 0$.

A6. $nh \to \infty$ and $nh^{2p+3} \to 0$, as $n \to \infty$.

A7. $nhh_0^4 \to 0$, as $n \to \infty$.

The assumptions above are standard in nonparametric modeling of time series (Fan and Yao, 2003). They exclude the cases when at least one of $X_i$ and $Y_i$ is nonstationary, e.g., exponentially smoothing model of Taylor (2008).[8] The assumptions require the finite variance in the spatial neighborhood of $X_i = x$. Failure of this assumption generally leads to asymptotic nonnormality of the estimator $\widehat{\theta}$ (Linton and Xiao, 2013) thus invalidates the standard chi-square limit of

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

the empirical likelihood ratio.[9] Note that it does not rule out the data generating process with infinite global variance. In Section 6, we demonstrate that in practice the distribution for observations in a spatial neighborhood could have much thinner tails than the global distribution. We also assume undersmoothing to eliminate the smoothing bias which is commonly recommended when making inference for a nonparametric function (e.g., Owen, 2001). The following result gives the null distribution of $\mathcal{L}(\beta, \theta)$.

THEOREM 1. *Under the assumptions and the null hypothesis* $H_0 : (\beta, \theta) = (\beta_0, \theta_0)$, $\mathcal{L}(\beta_0, \theta_0) \xrightarrow{d} \chi^2(2)$ *as* $n \to \infty$.

It is a bit surprising that the (generalized) empirical likelihood, which was originally designed for independent data, gives an asymptotically pivotal test statistic under serial dependence. In fact, neglecting smaller order terms $\mathcal{L}(\beta_0, \theta_0)$ can be written as the square of a self-normalized sum (cf. the proof of Theorem 1 in Appendix) $\mathcal{L}(\beta_0, \theta_0) = A^\top B^{-1} A + o_p(1)$, where $A = (nh)^{-1/2} \sum_{i=1}^n m_i(\beta_0, \theta_0)$ and $B = (nh)^{-1} \sum_{i=1}^n m_i(\beta_0, \theta_0) m_i^\top(\beta_0, \theta_0)$. The long run variance of $A$ has asymptotically vanishing covariance terms due to the weakened serial dependence of $Y_i$ by spatial smoothing along the direction of the covariate $X_i$, so that it can be effectively approximated by the sum of squares $B$. The corresponding results do not hold when EL is applied to unconditional VaR or ES models since the covariance terms in the long run variance play a role even asymptotically (Chen, 2008).

### 3.3. Profiled Likelihood

Theorem 1 shows that a simultaneous confidence region or a joint test for VaR and ES can be constructed based on the statistic $\mathcal{L}(\beta, \theta)$.[10] In many cases we are interested in inference about ES separately. This can be achieved by profiling out the nuisance parameter $\beta$ in the statistic $\mathcal{L}(\beta, \theta)$ :

$$\mathcal{L}(\theta) = \min_{\beta \in \mathcal{B} \subset \mathbb{R}^1} \mathcal{L}(\beta, \theta). \tag{11}$$

The limit behavior of $\mathcal{L}(\theta)$ is described in the following result.

THEOREM 2. *Under the assumptions stated and the null hypothesis* $H_0' : \theta = \theta_0$, $\mathcal{L}(\theta_0) \xrightarrow{d} \chi^2(1)$ *as* $n \to \infty$.

The confidence interval (CI) for $\theta$ can be constructed by inverting the test based on $\mathcal{L}(\theta)$, that is, by collecting all the values of $\theta$ that are not rejected by the test. It is also noteworthy that the CI such formed using the concentrated statistic $\mathcal{L}(\theta)$ is different from the one obtained from projection of a simultaneous CI. The projection-based method is very commonly used in applied research and is known to deliver conservative and thus less informative confidence intervals.[11]

Scaillet (2005) considered kernel smoothing (local constant) estimation of the ES and provided a standard error. The standard errors for general local polynomial

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

ES estimators are unknown. Such a Wald approach is not available for the mean excess loss estimator or for conducting any joint inference. We circumvent the difficulties by relying on an automatic method to select standard errors implicitly. The associated confidence interval is flexible to allow for an asymmetric shape unlike the Wald interval constructed with the explicit standard error. Scaillet's variance estimator is constructed in a way that does not guarantee positivity in finite samples. We find that negative variance estimate, thus unavailability of a standard error, is not a rare event in our simulations.

More generally, we can construct the test statistic for a linear function of $\beta$ and $\theta$, denoted as $\kappa = r_1\beta + r_2\theta$, where $r_1$ and $r_2$ are known real numbers. One example is the mean excess loss when $(r_1, r_2) = (1, -1)$, which by definition is nonnegative. In this case we modify the restriction (6) by

$$\sum_{i=1}^{n} \pi_i W_p((X_i - x)/h)[Y_i G((\beta - Y_i)/h_0) - \tau(\kappa - r_1\beta)/r_2] = 0. \tag{12}$$

Define the test statistic for $\kappa$ as

$$\mathcal{L}(\kappa) = \min_{\beta \subset \mathcal{B} \in \mathbb{R}^1} \mathcal{L}(\beta, \kappa), \tag{13}$$

where $\mathcal{L}(\beta, \kappa)$ is defined as in (8) except that the restriction (6) is replaced by (12).

THEOREM 3. *Under the assumptions stated and the null hypothesis* $H_0''$ : $\kappa = \kappa_0$, *for real numbers* $r_1, r_2$ *such that* $r_2 \neq 0$, $\mathcal{L}(\kappa_0) \xrightarrow{d} \chi^2(1)$ *as* $n \to \infty$.

## 4. A SEMIPARAMETRIC INDEX MODEL

Suppose we have multiple covariates that are contained in the $d$-dim ($d > 1$) vector $Z_i$. The VaR and ES of $Y_i$ given $Z_i = z$ are then defined as $\beta_0(z, \tau) = F^{-1}(\tau|z) = \inf\{y : F(y|z) \geq \tau\}$ and $\theta_0(z, \tau) = \mathrm{E}(Y_i | Y_i \leq \beta_0, Z_i = z)$, respectively, where $F^{-1}(\cdot|z)$ denotes the conditional distribution of $Y_i$ given $Z_i = z$.

In this section, we consider a semiparametric index model, which assumes that VaR and ES depend on $Z_i$ only through a linear combination of $Z_i$ (the so-called index variable) $Z_i^\top \gamma_0$, that is, $\beta_0(z, \tau) = \beta_0(z^\top \gamma_0, \tau)$ and $\theta_0(z, \tau) = \theta_0(z^\top \gamma_0, \tau)$. This assumption is restrictive, but it achieves dimension reduction so that $\beta_0$ and $\theta_0$ can still be estimated at the one-dimensional nonparametric rate. Note that for simplicity we also restrict the index parameter $\gamma$ to be the same in the conditional VaR and ES functions, which holds, e.g., if the conditional distribution function of $Y_i$ assumes an index model, as considered by Hall and Yao (2005). For identification purpose, assume that the functions $\beta_0(\cdot, \tau)$ and $\theta_0(\cdot, \tau)$ are differentiable and nonconstant, and the first element of $\gamma$ is one.

Suppose we have a consistent estimator $\widehat{\gamma}$ of $\gamma_0$. The semiparametric VaR and ES estimators $\widehat{\beta}^{SP}$ and $\widehat{\theta}^{SP}$ are obtained by solving $m_i^{SP}(\widehat{\beta}^{SP}, \widehat{\theta}^{SP}) = 0$, where

$$m_i^{SP}(\beta, \theta) = \begin{pmatrix} W_p((Z_i - z)^\top \widehat{\gamma}/h)[G((\beta - Y_i)/h_0) - \tau] \\ W_p((Z_i - z)^\top \widehat{\gamma}/h)[Y_i G((\beta - Y_i)/h_0) - \tau\theta] \end{pmatrix}.$$

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

The GEL criterion function based on the estimating functions $m_i^{SP}(\beta,\theta)$ is then denoted as $\mathcal{L}^{SP}(\beta,\theta)$. Suppose that the conditions in the assumptions hold for $Y_i$ and the covariates $Z_i$.

THEOREM 4. *Suppose that (a) $\gamma \in \Gamma$, where $\Gamma \subset \mathbb{R}^d$ is compact; (b) the density of $Z_i^\top \gamma$ is continuous in $\gamma$, and uniformly bounded on $\Gamma$; and (c) $n^{1/2}(\widehat{\gamma}-\gamma_0) = O_p(1)$. Then the chi-square limits of $\mathcal{L}^{SP}(\beta,\theta)$ under the hypotheses $H_0$ and $H_0'$, as in Theorems 1 and 2, still hold.*

A natural estimate $\widehat{\gamma}$ may be based on semiparametric quantile regression or the average-derivative estimator (Chaudhuri, Doksum, and Samarov, 1997; Wu, Yu, and Yu, 2010). Root-n convergence of $\widehat{\gamma}$ is imposed for simplicity as it is satisfied by most standard semiparametric estimators. It is not a necessary assumption though (see the proof of Theorem 4 in the Appendix).

## 5. FINITE SAMPLE PERFORMANCE

In this section we provide evidence on the finite sample performance of the methods proposed and focus on those based on empirical likelihood. In the experiments we generate the data from the following two designs (data generating processes or DGPs):

DGP 1: $r_i = 0.017 + \sigma_i \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0,1)$, $\sigma_i^2 = 0.013 + 0.353(r_{i-1} - 0.0174)^2$.

DGP 2: The same as DGP 1 except that $\varepsilon_i$ has Student's $t$ distribution with degrees of freedom five.

The initial observation is set to zero and for each sample we discard the first 200 observations to eliminate the initiation effects. The chosen coefficient values are obtained by fitting linear ARCH models to a empirical stock return dataset (Tsay, 2005). Let $Y_i = r_i$, $X_i = r_{i-1}$, $\tau = 0.05$ and thus we are interested in forecasting the next day VaR or expect shortfall (at the 5% level) given the current information. In what follows we only report simulation results regarding forecast of the risk measures when the lagged return is zero (i.e., $x = 0$) and those for other values of $x$ are found to be similar. The sample size is $n = 1,000$ and the number of replications is 2,000.

For point estimation of VaR and ES we consider the popular local linear estimator ($p = 1$). The kernels used are the standard normal distribution function for $G$ and the Epanechnikov kernel $K(u) = 0.75(1-u^2)\mathbb{I}_{[-1,1]}$. For illustration, we consider undersmoothing bandwidths $h_0 = \widehat{\sigma}_Y n^{-1/2}$ and $h = 3\widehat{\sigma}_X n^{-1/4}$, where $\widehat{\sigma}$ means sample standard deviation. The point estimates of VaR, ES, and mean excess loss (MXL) are reported in Table 1 together with true values. All three risk measures are estimated with reasonable biases, with larger standard errors for DGP 2 than DGP 1.

We consider tests of hypotheses $H_0 : (\beta,\theta) = (\beta_0,\theta_0)$, $H_0' : \theta = \theta_0$, and $H_0'' : \kappa = \kappa_0$ on VaR and ES jointly, ES, and MXL respectively. To calculate the statistic $\mathcal{L}(\beta,\theta)$ for any given $(\beta,\theta)$, we use the standard empirical likelihood

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

**TABLE 1.** Point estimation of VaR, expected shortfall (ES), and mean excess loss (MXL)

| | Design 1 | | | Design 2 | | |
|---|---|---|---|---|---|---|
| | VaR | ES | MXL | VaR | ES | MXL |
| True | .1680 | .2151 | .0471 | .2097 | .3084 | .0987 |
| Mean | .1703 | .2166 | .0463 | .2187 | .3166 | .0979 |
| Median | .1701 | .2168 | .0461 | .2179 | .3151 | .0972 |
| STD | .0124 | .0160 | .0124 | .0195 | .0347 | .0311 |

(EL) algorithm.[12] A Nelder–Mead simplex-based search is then used to calculate the concentrated statistic $\mathcal{L}(\theta)$. The similar algorithm is used to compute $\mathcal{L}(\kappa)$. To test $H_0'$, besides the empirical likelihood, we also consider the method based on the Wald statistic $W_*^2$, where $W_* = \sqrt{nh}(\widehat{\theta} - \theta_0)/\sqrt{\widehat{\Gamma}}$ and $\widehat{\Gamma}$ is the variance estimate proposed by Scaillet (2005, Sect. 3.2, Prop. 2). Under the assumptions and the null $H_0'$, the Wald statistic has a $\chi^2(1)$ limit. The estimated variance $\widehat{\Gamma}$ is constructed by plug-in estimation of unknown quantities in the asymptotic variance and is not guaranteed to be positive in finite samples, which may lead to nonsensical inference based on the associated Wald statistic. This is verified in the simulations. Figure 1 displays all 2,000 realizations of variance estimates for the two designs, and it is clear that a nonnegligible portion of them falls below the zero line (75 and 40 out of 2,000 samples respectively).[13] Note that the Wald statistic for a joint test on VaR and ES or a test on MXL ($H_0$ or $H_0''$) has not been available

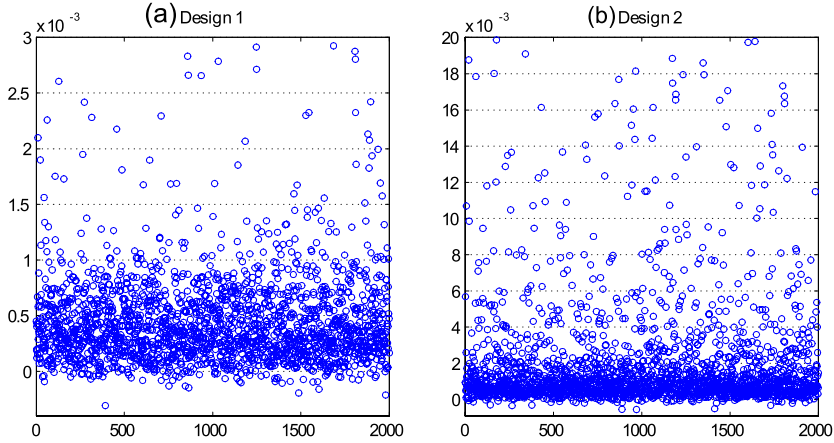



**FIGURE 1.** 2,000 realizations of estimated asymptotic variances (the 2.5% large outliers are not displayed). Among 2,000 samples, 75 and 40 estimated variances are negative for Design 1 and Design 2 respectively.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

in the literature yet and the empirical likelihood method proposed in this paper seems to be the only practical choice.

Tables 2 and 3 report the finite sample size and power of the tests considered at the significance levels 5% and 10%. For $H_0$ and $H_0'$, the power is calculated when the values of ES in the null are $\pm 10\%$, $\pm 20\%$, and $\pm 30\%$ away from true values used in DGPs. The similar is true for the power of $H_0''$. We focus on $H_0'$ below. The Wald test overrejects seriously (especially under DGP 2) and the EL test also overrejects but in a milder way. This is consistent with the fact reported in the tables (see the columns labeled "Quant.") that finite-sample quantiles of the EL statistic are closer to the $\chi^2(1)$ quantiles than those of the Wald statistic. Regarding the power, the EL test has the uniformly higher rejection rates than the Wald test when the value of ES in the null is less than the true value, but it is not so otherwise (e.g., the Wald test is more powerful for 10%, 20%, and 30% deviations from the true value under DGP 2). By using finite-sample quantiles as critical values to adjust for size distortion, the EL test uniformly dominates the Wald test in terms of such adjusted power.

For the joint test of VaR and ES and the test of MXL, Tables 2 and 3 show that the proposed EL test has reasonable size, and the power approaches one when the alternative hypothesis drifts away from the null.

**TABLE 2.** (Design 1) The size and power of Wald and EL tests for the expected shortfall (ES), the joint EL test for the VaR and ES, and the EL test for the mean excess loss (MXL). The power and adjusted power are calculated when the value in the null is $\pm 10\%$, $\pm 20\%$, and $\pm 30\%$ away from its true value. Note that 5% and 10% quantiles are 3.84 and 2.71 for $\chi^2(1)$, and 5.99 and 4.61 for $\chi^2(2)$ respectively

| Test | Size | Quant. | −30% Dev. | −20% | −10% | 10% | 20% | 30% |
|---|---|---|---|---|---|---|---|---|
| | | | Power | | | | | |
| Wald 5% ES | .108 | 16.6 | .88 | .61 | .21 | .26 | .59 | .86 |
| Wald 10% ES | .144 | 4.40 | .93 | .73 | .23 | .35 | .69 | .90 |
| EL 5% ES | .062 | 4.18 | 1.00 | .94 | .38 | .23 | .62 | .88 |
| EL 10% ES | .121 | 3.02 | 1.00 | .96 | .49 | .34 | .75 | .95 |
| EL 5% Joint | .059 | 6.31 | 1.00 | 1.00 | .48 | .36 | .77 | .95 |
| EL 10% Joint | .118 | 4.96 | 1.00 | 1.00 | .65 | .47 | .86 | .99 |
| EL 5% MXL | .039 | 3.29 | 1.00 | 1.00 | .31 | .41 | .85 | .98 |
| EL 10% MXL | .077 | 2.32 | 1.00 | 1.00 | .54 | .56 | .91 | 1.00 |
| | | | Adjusted Power | | | | | |
| Wald 5% ES | | | .38 | .16 | .09 | .07 | .17 | .39 |
| Wald 10% ES | | | .83 | .60 | .22 | .22 | .55 | .83 |
| EL 5% ES | | | 1.00 | .94 | .38 | .19 | .57 | .87 |
| EL 10% ES | | | 1.00 | .96 | .50 | .29 | .70 | .93 |

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

**TABLE 3.** (Design 2) The size and power of Wald and EL tests for the expected shortfall (ES), the joint EL test for the VaR and ES, and the EL test for the mean excess loss (MXL). The power and adjusted power are calculated when the value in the null is ±10%, ±20%, and ±30% away from its true value. Note that 5% and 10% quantiles are 3.84 and 2.71 for $\chi^2(1)$, and 5.99 and 4.61 for $\chi^2(2)$ respectively

| Test | Size | Quant. | −30% Dev. | −20% | −10% | 10% | 20% | 30% |
|---|---|---|---|---|---|---|---|---|
| | | | Power | | | | | |
| Wald 5% ES | .137 | 9.96 | .71 | .52 | .23 | .26 | .45 | .62 |
| Wald 10% ES | .189 | 5.36 | .81 | .60 | .30 | .31 | .54 | .70 |
| EL 5% ES | .066 | 4.36 | 1.00 | .82 | .29 | .14 | .33 | .55 |
| EL 10% ES | .137 | 3.22 | 1.00 | .88 | .43 | .22 | .45 | .67 |
| EL 5% Joint | .078 | 7.00 | 1.00 | .97 | .27 | .24 | .52 | .69 |
| EL 10% Joint | .139 | 5.47 | 1.00 | .99 | .44 | .33 | .66 | .78 |
| EL 5% MXL | .053 | 3.86 | 1.00 | .97 | .25 | .25 | .58 | .75 |
| EL 10% MXL | .104 | 2.78 | 1.00 | .99 | .38 | .37 | .68 | .84 |
| | | | Adjusted Power | | | | | |
| Wald 5% ES | | | .45 | .25 | .12 | .10 | .26 | .40 |
| Wald 10% ES | | | .66 | .47 | .22 | .20 | .39 | .54 |
| EL 5% ES | | | .99 | .81 | .28 | .12 | .28 | .50 |
| EL 10% ES | | | 1.00 | .87 | .42 | .17 | .40 | .61 |

## 6. AN EMPIRICAL EXAMPLE

We consider an empirical application to forecasting risk measures for Canadian stock index returns. The data are daily prices $P_i$ (adjusted for dividends and splits) (with 2,336 observations) of S&P/TSX composite index from 3 January 2000 to 31 March 2009.[14] The log returns are computed as $r_i = \log(P_i/P_{i-1})$, $i = 1, \ldots, 2,335$. We use the lag $r_{i-1}$ as the covariate, so $n = 2,334$. The time plots of the price and the return series are shown in Figure 2(a) and (b) respectively. The plot of $r_i$ versus $r_{i-1}$ is shown in Figure 2(c).

Standard unit root tests (Table 4) provide strong evidence of stationarity of return and squared return series. The weak serial dependence of these two series is confirmed by ACF plots (Figure 4) and Ljung–Box portmanteau test (Figure 5). Although squared returns are much more dependent than the levels, as most asset returns, some evidence against martingale differences is observed. In any event, inference methods ignoring serial dependence in returns might be inappropriate.

To further illustrate our spatial smoothing approach, we examine time series properties for observations in the spatial neighborhood of $r_{i-1} = x$. Considering $x = 0$, Figure 6(c) plots the p-value of the Ljung–Box correlation test at the lag $\log(n)$ for a wide range of bandwidths (applied to squared returns). It shows that the serial correlation for local observations is much weaker than that for the whole

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

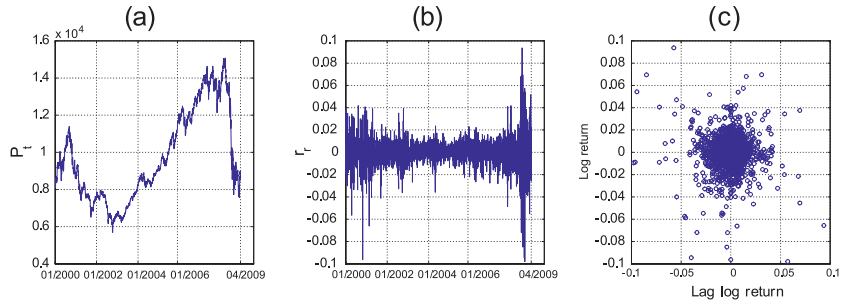



**FIGURE 2.** (a) Daily prices (adjusted for dividends and splits) of S&P/TSX composite index from 3 January 2000 to 31 March 2009; (b) Time series of daily log returns; (c) Log returns versus lagged log returns ($n = 2,334$).

**TABLE 4.** Test statistics of Augmented Dickey–Fuller (ADF) and Elliott, Rothenberg, and Stock (1996) (ERS) DF-GLS unit root tests of S&P/TSX composite index returns. They are all significant at 1% level. For both unit root tests, we use Bayesian information criterion (BIC) to select the numbers of lags

| | ADF Unit Root Tests | | | ERS Unit Root Tests | |
|---|---|---|---|---|---|
| | None | Constant | Trend | Constant | Trend |
| $r_i$ | −37.2 | −37.2 | −37.2 | −4.65 | −8.71 |
| $r_i^2$ | −23.4 | −25.3 | −25.6 | −6.98 | −9.04 |
| 1% critical value | −2.58 | −3.44 | −3.98 | −2.57 | −3.48 |

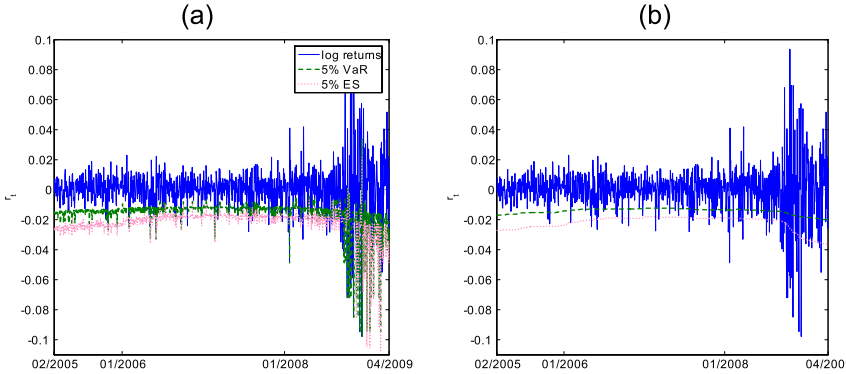



**FIGURE 3.** Nonparametric forecasts of 5% VaR and ES for the last 1,000 observations in the sample, using a rolling-window scheme. (a) Conditional approach; (b) Unconditional approach (historical simulation).

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

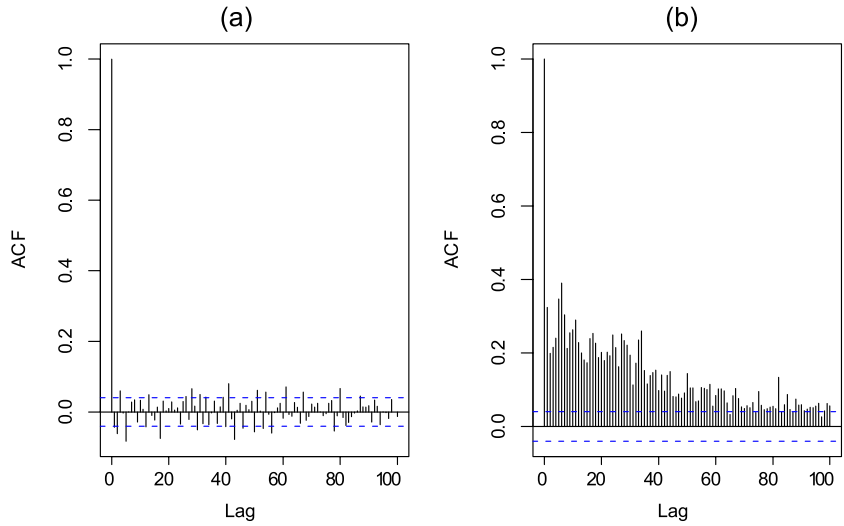



**FIGURE 4.** Autocorrelation functions (ACFs) of S&P/TSX composite index returns (Panel a) and squared returns (Panel b).

history (which corresponds to using a very large bandwidth; cf. Figure 5(b)). The kurtosis for local observations is plotted in Figure 6(b) as a function of bandwidth. Although the unconditional kurtosis is about 12.4, which may raise concerns for heavy tails, the problem is much less for local observations (mostly below 4). These comparisons verify our theoretical assumptions earlier on temporal dependence and moment conditions.

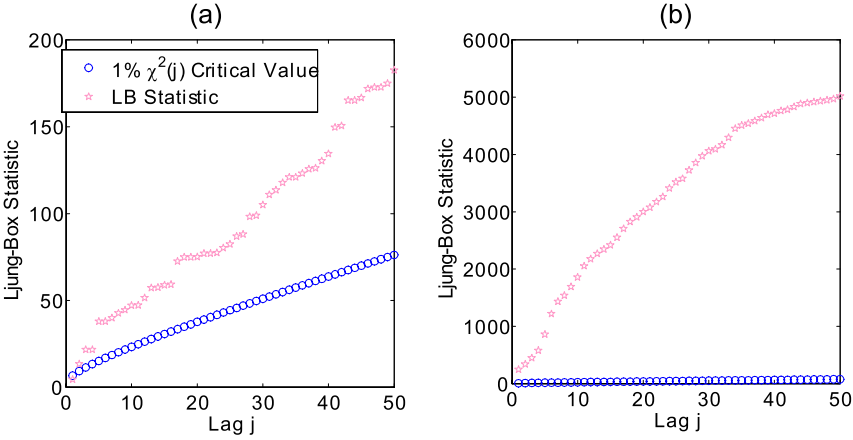



**FIGURE 5.** Ljung–Box portmanteau tests of autocorrelations of S&P/TSX composite index returns (Panel a) and squared returns (Panel b).

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

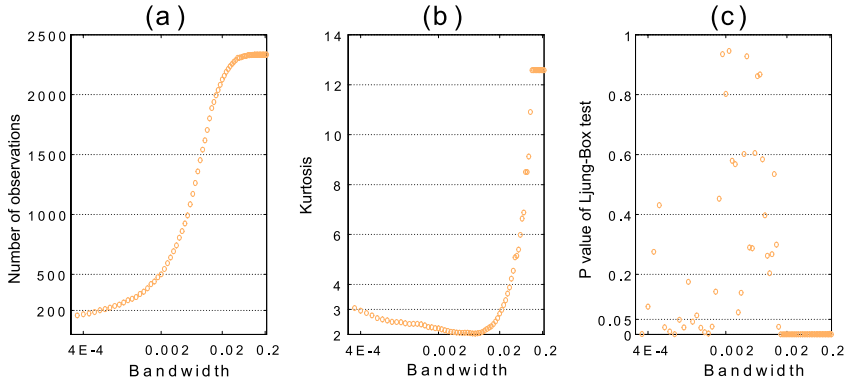



**FIGURE 6.** Time series properties for observations in spatial neighborhoods of $x = 0$. (a) The number of observations; (b) Kurtosis; (c) The p-value of Ljung–Box test (the null: no serial correlation in *squared returns* at the lag $\log(n)$).

## 6.1. Nonparametric Estimates

The point and interval estimates of 5% VaR and ES are reported in Table 5 for a range of the spatial values of the lagged returns $x$ from $-0.025$ to 0.025. The estimated ES curve is plotted in Figure 9 together with empirical likelihood (EL) and Wald confidence intervals (CIs). The curve appears to be asymmetric, having the commonly observed U-shape or "smile"-shape in the region of positive returns and not-too-large losses. The ES is relatively flat when the lag return is less than $-0.010$. It is worth mentioning that the curve is not minimal at zero or mean return but at some positive number around 0.010. The ES is estimated with smaller Wald standard error, which is computed according to Scaillet (2005), when $x$ is close to the middle of the range where the data are relatively rich. The standard error is quite large when $x$ lies relatively far away from zero (e.g., $x = 0.025$) and the associated Wald confidence interval for ES is too wide to be useful.

The curves of nonparametric estimates together with confidence bands in Figure 9 can be used as ES forecasts next day. For instance, the log return at the sample end (31 March 2009) is 0.0143. The next day expected shortfall is estimated as 0.0270 (the VaR is 0.0160) with 90% EL CI [0.0203,0.0339] and Wald CI [0.0159,0.0381]. The forecasts over longer horizons can be built on rolling or expanding windows. The curves can be updated as new data arrive.

One-sided inference can also be made. For example, if a portfolio manager at the sample end is willing to accept portfolios with an ES up to 3.5% of the current capital, he/she may probably be happy with the current asset allocation since the point forecast is less than 3.5%. An EL-based one-sided test tells the manager the probability of the expected shortfall next day larger than 3.5% is 0.026 ($p$-value), while a Wald method gives the probability 0.116.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

**TABLE 5.** The 90% Wald and EL confidence intervals for conditional expected shortfalls (ES's) of S&P/TSX composite index for $-0.025 \leq x \leq 0.025$. The length ratios of two intervals (note: the Wald interval is longer if Len.Ratio$>1$) and the degree of asymmetry of EL intervals (note: the subinterval right to the point estimate is wider if Asym.$>1$) are also reported

| $x$ | VaR | ES | std(ES) | Wald CI | Length | EL CI | Length | Len. Ratio | Asym. |
|---|---|---|---|---|---|---|---|---|---|
| −.025 | .0334 | .0401 | .0197 | [.0077, .0726] | .0649 | [.0323, .0560] | .0237 | 2.74 | 2.04 |
| −.020 | .0315 | .0352 | .0883 | [−.1100, .1803] | .2903 | [.0292, .0359] | .0067 | 43.33 | 0.12 |
| −.015 | .0228 | .0309 | .0094 | [.0154, .0463] | .0309 | [.0261, .0360] | .0099 | 3.12 | 1.06 |
| −.010 | .0228 | .0357 | .0074 | [.0236, .0477] | .0241 | [.0296, .0425] | .0129 | 1.87 | 1.11 |
| −.005 | .0199 | .0321 | .0052 | [.0235, .0408] | .0173 | [.0275, .0399] | .0124 | 1.40 | 1.70 |
| 0 | .0170 | .0285 | .0041 | [.0216, .0353] | .0137 | [.0247, .0354] | .0107 | 1.28 | 1.82 |
| .005 | .0170 | .0267 | .0039 | [.0203, .0331] | .0128 | [.0230, .0308] | .0078 | 1.64 | 1.11 |
| .010 | .0160 | .0246 | .0046 | [.0171, .0321] | .0150 | [.0208, .0287] | .0079 | 1.90 | 1.08 |
| .015 | .0189 | .0282 | .0032 | [.0228, .0335] | .0107 | [.0222, .0354] | .0132 | 0.81 | 1.20 |
| .020 | .0305 | .0388 | .0133 | [.0170, .0606] | .0436 | [.0288, .0472] | .0184 | 2.37 | 0.84 |
| .025 | .0295 | .0375 | 1.9834 | [−3.225, 3.300] | 6.525 | [.0277, .0489] | .0212 | 307.8 | 1.16 |

We observe that compared to Wald method, empirical likelihood generally yields tighter confidence intervals for ES with much more reasonable length and appears to be quite robust to the location of $x$ (see Table 5 and Figure 9). Table 6 reports the point and interval forecasts of the mean excess losses (MXLs). For both ES and MXL, EL confidence intervals are able to account for the asymmetry of estimation risk on the two sides of point estimates (as reported in the last columns of Tables 5 and 6) unlike Wald intervals that are artificially restricted to symmetric. For example, if we observe the log return today is 0.0100 ($x = 0.0100$), the loss that one may expect to suffer tomorrow in the worst 5% percent of possible scenarios is 0.0246 (times the amount of the financial position). EL confidence interval shows that the expected loss would be more likely to be above 0.0246 than to be below it, the information that Wald CI fails to provide.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

**TABLE 6.** The 90% EL confidence intervals for conditional mean excess losses (MXLs) of S&P/TSX composite index for $-0.025 \leq x \leq 0.025$. The degree of asymmetry of EL intervals (note: the subinterval right to the point estimate is wider if Asym.$>1$) is also reported

| $x$ | MXL | EL CI | Length | Asymm. |
|---|---|---|---|---|
| −0.025 | .0067 | [−.0008, .0209] | .0217 | 1.89 |
| −0.020 | .0037 | [.0004, .0096] | .0092 | 1.79 |
| −0.015 | .0081 | [.0030, .0115] | .0085 | 0.67 |
| −0.010 | .0129 | [.0071, .0175] | .0104 | 0.79 |
| −0.005 | .0122 | [.0084, .0190] | .0106 | 1.79 |
| 0 | .0115 | [.0085, .0174] | .0089 | 1.97 |
| 0.005 | .0097 | [.0064, .0124] | .0060 | 0.82 |
| 0.010 | .0086 | [.0046, .0113] | .0067 | 0.68 |
| 0.015 | .0093 | [.0051, .0148] | .0097 | 1.31 |
| 0.020 | .0083 | [.0016, .0171] | .0155 | 1.31 |
| 0.025 | .0080 | [−.0019, .0239] | .0258 | 1.61 |

To further illustrate this, we plot the empirical likelihood statistics (the $\mathcal{L}$ as defined in (11) and (13)) together with the Wald statistics for ES and MXL in Figures 7 and 8 respectively when $x = 0.010$ and three different bandwidths are used $h_j = 1.3^{j-2}\overline{h}$ (with values 0.0043, 0.0056, and 0.0073), where $\overline{h} = 3\widehat{\sigma}_X n^{-1/4}$ as used in simulations. The largest bandwidth $h_3$ generates the tightest confidence interval. Asymmetry of the EL statistic is clear and the associated CIs are tighter on the left sides of the point estimates than the right sides.

Before we conclude this subsection, we compare point estimates of VaR and ES with those based on the *unconditional* method, e.g., historical simulation. Historical simulation uses a quantile of the empirical CDF of all historical observations to forecast the next day VaR. ES is obtained by taking the average of historical losses that are larger than VaR. Using such a method, VaR and ES are then estimated as 0.0203 and 0.0334 respectively. Our estimates for VaR and ES range from 0.0160 to 0.0334, and from 0.0246 to 0.0401 respectively, depending on the values of the covariate $x$. In contrast to historical simulation in which equal weights are assigned to all the historical observations, the *conditional* method we adopt here utilizes the information contained in covariates and puts more weights on observations which bear more resemblance (in terms of the values of covariates) to "the observation next day". To further illustrate this, we hold the last 1,000 observations as the validation sample and use the first $n - 1000$ observations to predict the VaR and ES. Using a rolling-window scheme, we implement the historical simulation approach and our conditional approach. The predictions are shown in Figure 3. The unconditional approach, despite its simplicity in implementation, is relatively less informative in reflecting the time varying features of the VaR and ES; e.g., the associated forecasts miss the periods in which holding the asset is more risky than other periods.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

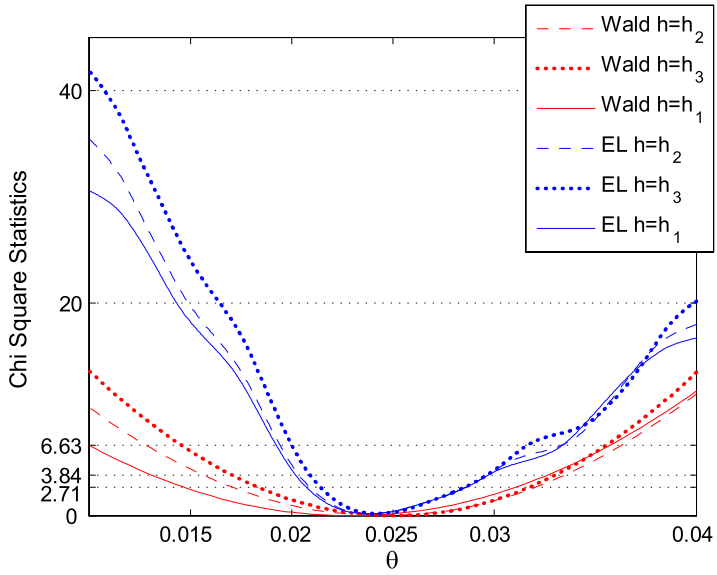



**FIGURE 7.** The Wald test (the lower 3 lines) and EL test statistics (the upper 3 lines) are plotted against $\theta$, the value of the expected shortfall of log returns of S&P/TSX composite index in the null hypothesis when $x = 0.010$. Both test statistics have the $\chi^2(1)$ limit distribution. Three different smoothing bandwidths (in the X direction), $h_j = 1.3^{j-2}\overline{h}$, where $j = 1, 2, 3$ and $\overline{h} = 3\widehat{\sigma}_X n^{-1/4}$, are used when constructing estimating equations. The corresponding point estimates are 0.0229, 0.0246, and 0.0250 for 3 bandwidths. Three critical values for 90%, 95%, and 99% confidence intervals are labeled on the y-axis.

## 6.2. An ES-Based Specification Analysis

We now provide an ES-based specification analysis of several parametric models. We consider a linear mean function and an OLS fit gives

$$r_i = m(r_{i-1}) + \sigma_i \varepsilon_i = \underset{[.921]}{(2.663\text{E}-5)} - \underset{[.036]}{0.043} r_{i-1} + \sigma_i \varepsilon_i$$

with $p$-values in the brackets. The following four models are considered for the volatility and innovation density:

$$\text{AR-ARCH-normal} : \varepsilon_i \sim \mathcal{N}(0,1),\ \sigma_i^2 = (1.142\text{E}-4) + 0.292 r_{i-1}^2,$$

$$\text{AR-ARCH-}t : \varepsilon_i \sim \text{Standardized } t(3.233),\ \sigma_i^2 = (1.203\text{E}-4) + 0.382 r_{i-1}^2,$$

$$\text{AR-ARCH-skewed-}t : \varepsilon_i \sim \text{Skewed } t(4.1, -0.132),\ \sigma_i^2 = (1.004\text{E}-4) + 0.318 r_{i-1}^2,$$

$$\text{AR-ARCH-GJR-}t : \varepsilon_i \sim \text{Standardized } t(3.134), \sigma_i^2 = (1.227\text{E}-4) + 0.292 r_{i-1}^2 + 0.203 r_{i-1}^2 \mathbb{I}_{\{r_{i-1}<0\}},$$

where parameters are estimated by MLE.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

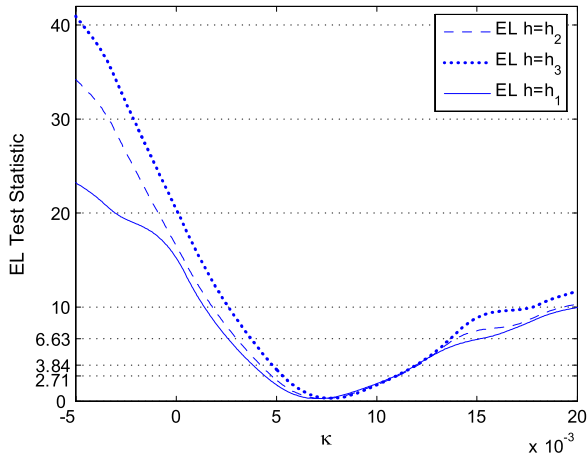



**FIGURE 8.** The EL test statistic is plotted against $\kappa$, the value of the mean excess loss (MXL) of the log returns of S&P/TSX composite index in the null hypothesis when $x = 0.010$. It has the $\chi^2(1)$ limit distribution. Three different smoothing bandwidths (in the X direction), $h_j = 1.3^{j-2}\overline{h}$, where $j = 1, 2, 3$ and $\overline{h} = 3\widehat{\sigma}_X n^{-1/4}$, are used when constructing estimating equations. The corresponding point estimates are 0.0059, 0.0086, and 0.0090 for 3 bandwidths. Three critical values for 90%, 95%, and 99% confidence intervals are labeled on the y-axis.

The sample unconditional (excess) kurtosis is 9.440. The kurtosis is larger than any $t$-distribution that has finite kurtosis, i.e., Student's $t(\upsilon)$, $\upsilon \geq 5$, which indicates a Gaussian model is inadequate. The sample unconditional skewness is $-0.774$, suggesting a skew to the left. It is close to the population skewness of a skewed $t(4.1, -0.132)$ distribution; a simulation based on twenty million draws gives the skewness around $-0.851$. Accommodation of distributional asymmetry reduces the effects of heavy tails and conditional heteroskedasticity. The skewed-$t$ model as a parsimonious vehicle captures stylized features of asset return distribution like time-varying volatility, asymmetry, and heavy tails/leptokurtosis.

In Figure 9 we plot the induced ES curves for parametric models. In particular, the conditional density for the skewed-$t$ model, from which the VaR and ES are derived, is $f(y|r_{i-1}) = \vartheta((y - m(r_{i-1}))/\sigma_i)/\sigma_i$, where $\vartheta$ is the skewed $t$ density with the estimated asymmetry parameter and degree of freedom. All parametric ES curves are quadratic. The Gaussian model underestimates the ES since it does not account for heavy tails in the return distribution. Its ES curve is uniformly below the nonparametrically estimated curve using the unconditional or conditional method. The ES curves based on two Student's $t$ models are closer to the nonparametric fit than the Gaussian model, and they are indistinguishable from each other. This is not very surprising considering that ES mainly depends on the tail behavior of return distribution and is relatively less sensitive to its global shape. We additionally fit an ARCH GJR model (Glosten, Jaganathan, and Runkle, 1993) with $t$ innovations by adding a correction term to allow for

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

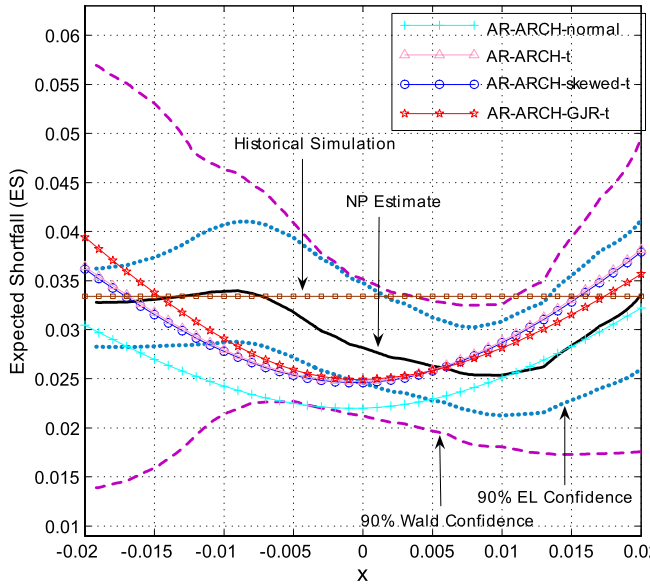



**FIGURE 9.** Nonparametric ES estimates together with 90% EL and Wald confidence bands for the log returns of S&P/TSX composite index. ES curves based on the unconditional method and four estimated parametric models (AR-ARCH model with normal, Student's $t$ and skewed Student's $t$ innovations and AR-GJR-ARCH model with $t$ innovations) are also plotted.

different effects of positive and negative returns on the volatility. The induced ES curve deviates slightly from those based on the two $t$ models and is higher at negative returns than positive returns of the same magnitudes, reflecting negative returns have a big impact on the risk next day. All parametric ES curves lie well within the Wald confidence band. None of them fits in the EL band, although more support is given to three non-Gaussian models.[15]

## 7. CONCLUDING REMARKS AND EXTENSIONS

In this paper we provide a general nonparametric framework to study statistical inference of the financial risk measures including the value at risk, the expected shortfall, and other related measures. It delivers information on a safety margin of the risk measure which is of great practical importance to financial decision making. A novel technique for inference is proposed via empirical likelihood. It avoids the difficulties in obtaining accurate standard errors of the underlying estimators and automatically provides an asymptotically pivotal test statistic. The associated confidence intervals are not restricted to be symmetric unlike those based on the Wald-type statistics. The asymptotic theories are developed in a general framework that nests empirical likelihood as a special case and also allows for weakly dependent time series data.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

## NOTES

1. Chan et al. (2007) noted that the 1996 amendment to the 1988 Basel Accord recommends to use three times VaR to set the market risk capital requirements and argued that the seemingly arbitrary factor of three needs to be justified. The similar concern also applies to the expected shortfall and other risk measures.

2. See Owen (2001) and Kitamura (2006) for recent surveys.

3. In the risk management literature, expected shortfall is also called conditional tail expectation or conditional VaR. In this paper we avoid these references to ES and reserve "conditional" to stand for "conditional on covariates".

4. For example, Tsay (2005) predicted conditional VaR of stock returns using explanatory variables controlling for current volatility and trending behavior of extreme returns, year-end effects, and previous panic trading. Levels of past positive and negative returns can be also included to address asymmetric effects on risk (e.g., Engle and Manganelli, 2004; Chan et al., 2007).

5. Kuan, Yeh, and Hsu (2009) proposed an alternative expectile-based risk measure which is also sensitive to extreme losses. Their risk measure is indexed by the level of "prudentiality" instead of the level of confidence and is argued to be less conservative than expected shortfall.

6. Using a smoothing bandwidth in $Y$-direction is also helpful in estimation of the CDF and conditional quantiles to reduce mean square errors if bandwidths are properly selected (Li and Racine, 2008).

7. Our estimator $\widehat{\theta}$ can be compared with the unconditional ES estimator which is defined as $\widehat{\theta}_{uc} = (n\tau)^{-1}\sum_{i=1}^{n} Y_i G((\widehat{\beta}_{uc} - Y_i)/h_0)$ with the unconditional VaR estimator $\widehat{\beta}_{uc}$ solving $n^{-1}\sum_{i=1}^{n} G((\widehat{\beta}_{uc} - Y_i)/h_0) = \tau$ (see Gourieroux, Scaillet, and Laurent, 2000; Scaillet, 2004; Chen, 2008, inter alia).

8. In the literature of predictive regression or predictive quantile regression (e.g., Campbell and Yogo, 2006), highly persistent explanatory variables such as the dividend yield, book to market ratio, and T-bill rate are commonly used to predict aspects of stock returns. Extension to allow for such persistent covariates will be a worthwhile future research direction.

9. Extension to allow for local observations with infinite variance is possible by using the tail-trimming strategy as in Hill (2014).

10. A GEL test statistic for the conditional VaR can be obtained through the $\mathcal{L}$ in (8) subject to the restrictions (5) and (7) only.

11. Theorem 2 is also useful to provide inferences for other economic or financial quantities which are related to the expected shortfall. For example, as noted in Peracchi and Tanase (2008), the generalized Lorenz curve (the Lorenz curve scaled by the mean income) which is popular in measuring social inequality (Gastwirth, 1971; Shorrocks, 1983; Kleiber, 2008) can be written as $Z_0(x,\tau) = \tau\theta_0(x,\tau)$ for the bottom $100\tau\%$ of households when $Y_i$ represents the household income and $X_i$ contains corresponding predictors. A test or a confidence interval for $Z_0(x,\tau)$ can be constructed based on the statistic $L(\theta)$ in (11).

12. We implement in Matlab using the codes that are written by Bruce Hansen and are available at http://www.ssc.wisc.edu/~bhansen/progs/progs_gmm.html.

13. In such cases, when computing the size and power of the test below we truncate the estimated variance as zero, so the Wald statistic is infinity and the test rejects $H_0'$.

14. Data source: http://ca.finance.yahoo.com.

15. In the analysis we do not allow asymmetry and kurtosis parameters to be time-varying (unlike Hansen, 1994), which simplifies the computation of ES curve. Another important simplification is consideration of only the first lag as the covariate for convenient graphical comparisons with the nonparametric estimator.

## REFERENCES

Acerbi, C. & D. Tasche (2002) On the coherence of expected shortfall. *Journal of Banking and Finance* 26, 1487–1503.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

Andrews, D.W.K. (1987) Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica* 55, 1465–1472.

Artzner, P., F. Delbaen, J.-M. Eber, & D. Heath (1999) Coherent measures of risk. *Mathematical Finance* 9, 203–228.

Baggerly, K.A. (1998) Empirical likelihood as a goodness-of-fit measure. *Biometrika* 85, 535–547.

Basak, S. & A. Shapiro (2001) Value-at-risk based risk management: Optimal policies and asset prices. *Review of Financial Studies* 14, 371–405.

Bassett, G., R. Koenker, & G. Kordas (2004) Pessimistic portfolio allocation and Choquet expected utility. *Journal of Financial Econometrics* 2, 477–492.

Cai, Z. & X. Wang (2008) Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics* 147, 120–130.

Campbell, J.Y. & M. Yogo (2006) Efficient tests of stock return predictability. *Journal of Financial Economics* 81, 27–60.

Chan, N.H., S.J. Deng, L. Peng, & Z. Xia (2007) Interval estimation for the conditional value-at-risk based on GARCH models with heavy tailed innovations. *Journal of Econometrics* 137, 556–576.

Chan, N.H., L. Peng, & D. Zhang (2011) Empirical likelihood based confidence intervals for conditional variance in heteroskedastic regression models. *Econometric Theory* 27, 154–177.

Chaudhuri, P., K. Doksum, & A. Samarov (1997) On average derivative quantile regression. *Annals of Statistics* 25, 715–744.

Chen, S.X. (2008) Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics* 6, 87–107.

Chen, S.X. & Y.S. Qin (2000) Empirical likelihood confidence intervals for local linear smoothers. *Biometrika* 87, 946–953.

Christoffersen, P. & S. Gonçalves (2005) Estimation risk in financial risk management. *Journal of Risk* 7, 1–28.

Corcoran, S.A. (1998) Bartlett adjustment of empirical discrepancy statistics. *Biometrika* 85, 967–972.

Cressie, N. & T. Read (1984) Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society (Series B)* 46, 440–464.

Elliott, G., T.J. Rothenberg, & J.H. Stock (1996) Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–836.

Engle, R.F. & S. Manganelli (2004) CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics* 22, 367–381.

Fan, J. & I. Gijbels (1996) *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.

Fan, J. & Q. Yao (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer.

Gastwirth, J.L. (1971) A general definition of the Lorenz curve. *Econometrica* 39, 1037–1039.

Glosten, L.R., R. Jagannathan, & D.E. Runkle (1993) On the relation between the expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance* 48, 1779–1801.

Gourieroux, C., O. Scaillet, & J.P. Laurent (2000) Sensitivity analysis of values at risk. *Journal of Empirical Finance* 7, 225–245.

Hall, P. & Q. Yao (2005) Approximating conditional distribution functions using dimension reduction. *Annals of Statistics* 33, 1404–1421.

Hansen, B.E. (1994) Autoregressive conditional density estimation. *International Economic Review* 35, 705–730.

Hart, J.D. (1996) Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics* 6, 115–142.

Hill, J. (2014) Expected shortfall estimation and gaussian inference for infinite variance time series. *Journal of Financial Econometrics*, forthcoming.

Kitamura, Y. (2006) Empirical likelihood methods in econometrics: Theory and practice. In R. Blundell, P. Torsten, & W.K. Newey (eds.), *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*. Cambridge University Press.

Kleiber, C. (2008) The Lorenz curve in economics and econometrics. In G. Betti & A. Lemmi (eds.), *Advances on Income Inequality and Concentration Measures: Collected Papers in Memory of C. Gini and M.O. Lorenz*, pp. 225–242. Routledge.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

Komunjer, I. (2007) Asymmetric power distribution: Theory and applications to risk measurement. *Journal of Applied Econometrics* 22, 891–921.
Kuan, C.-M., J.-H. Yeh, & Y.-C. Hsu (2009) Assessing value at risk with CARE: The conditional autoregressive expectile models. *Journal of Econometrics* 150, 261–270.
Li, Q. & J.S. Racine (2008) Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics* 26, 423–434.
Linton, O. & Z. Xiao (2013) Estimation of and inference about the expected shortfall for time series with infinite variance. *Econometric Theory* 29, 771–807.
Masry, E. & J. Fan (1997) Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics* 24, 165–179.
McNeil, A.J., R. Frey, & P. Embrechts (2005) *Quantitative Risk Management*. Princeton University Press.
Mittnik, S. & S.T. Rachev (2000) *Stable Paretian Models in Finance*. Wiley.
Newey, W.K. & R.J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72, 219–255.
Owen, A.B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–249.
Owen, A.B. (1990) Empirical likelihood ratio confidence regions. *Annals of Statistics* 18, 90–120.
Owen, A.B. (2001) *Empirical Likelihood*. Chapman and Hall/CRC.
Peracchi, F. & A.V. Tanase (2008) On estimating the conditional expected shortfall. *Applied Stochastic Models in Business and Industry* 24, 471–493.
Qin, J. & J. Lawless (1994) Empirical likelihood and general estimating equations. *Annals of Statistics* 22, 300–325.
Ruszczyski, A. & R. Vanderbei (2003) Frontiers of stochastically nondominated portfolios. *Econometrica* 71, 1287–1298.
Scaillet, O. (2004) Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance* 14, 115–129.
Scaillet, O. (2005) Nonparametric estimation of conditional expected shortfall. *Insurance and Risk Management Journal* 72(4), 639–660.
Shorrocks, A.F. (1983) Ranking income distributions. *Economica* 50, 3–17.
Smith R.J. (1997) Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal* 107, 503–519.
Taylor, J.W. (2008) Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. *Journal of Financial Econometrics* 6, 382–406.
Tsay, R.S. (2005) *Analysis of Financial Time Series*, 2nd ed. Wiley-Interscience.
Wu, T.Z., K. Yu, & Y. Yu (2010) Single-index quantile regression. *Journal of Multivariate Analysis* 101, 1607–1621.
Xu, K.-L. (2009) Empirical likelihood based inference for recurrent nonparametric diffusions. *Journal of Econometrics* 153, 65–82.
Xu, K.-L. (2013) Nonparametric inference of conditional quantiles of time series. *Econometric Theory* 29(4), 673–698.
Yamai, Y. & T. Yoshiba (2005) Value-at-risk versus expected shortfall: A practical perspective. *Journal of Banking and Finance* 29, 997–1015.

# APPENDIX: Proofs

In this section we provide the proofs of the theorems and some preliminary results. Denote $S$ as the $(p+1)$ by $(p+1)$ matrix with the $(j,k)$-th element $\int u^{j+k-2}K(u)du$, and $S^*$ as the $(p+1)$ by $(p+1)$ matrix with the $(j,k)$-th element $\int u^{j+k-2}K^2(u)du$.

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

LEMMA 1. *Suppose that the assumptions hold.*

*(i)* $\frac{1}{nh}\sum_{i=1}^{n} m_i(\beta_0,\theta_0)m_i(\beta_0,\theta_0)^\top \xrightarrow{p} V$, *where* $V = (V_{jk})_{j,k=1,2} = e_1^\top S^{-1}S^*S^{-1}e_1\overline{V}/\mu(x)$ *with*

$$\overline{V} = \begin{pmatrix} (1-\tau)\tau & (1-\tau)\tau\theta_0 \\ (1-\tau)\tau\theta_0 & (\xi(\beta_0,x)\tau - \tau^2\theta_0^2) \end{pmatrix};$$

*(ii)* $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} m_i(\beta_0,\theta_0) \xrightarrow{d} N(0,V)$.

*(iii)* *Let* $\widehat{\lambda} = \arg\max_{\lambda\in\Lambda_n} 2\big[\sum_{i=1}^{n}\rho(\lambda^\top m_i(\beta_0,\theta_0)) - n\rho(0)\big]$. *Then* $\widehat{\lambda} = O_p((nh)^{-1/2})$ *and* $\max_{1\le i\le n}|\widehat{\lambda}^\top m_i(\beta_0,\theta_0)| \xrightarrow{p} 0$.

*(iv)* *Let* $\widetilde{\beta} = \arg\min_{\beta\in\mathcal{B}} L(\beta,\theta_0)$ *and* $\widetilde{\lambda} = \arg\max_{\lambda\in\Lambda_n} 2\big[\sum_{i=1}^{n}\rho(\lambda^\top m_i(\widetilde{\beta},\theta_0)) - n\rho(0)\big]$. *Then* $\widetilde{\lambda} = O_p\big((nh)^{-1/2}\big)$ *and* $\max_{1\le i\le n}|\widetilde{\lambda}^\top m_i(\widetilde{\beta},\theta_0)| \xrightarrow{p} 0$.

*(v)* $\frac{1}{nh}\sum_{i=1}^{n} m_i(\widetilde{\beta},\theta_0)m_i(\widetilde{\beta},\theta_0)^\top \xrightarrow{p} V$.

*(vi)* *Denote* $\Phi = f(\beta_0|x)\mu(x)(1,\beta_0)^\top$, *where* $f(\cdot|x)$ *is the conditional density function of* $Y_i$ *given* $X_i = x$. *Then*

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\widetilde{m}_i = \left[I_2 - \Phi\left(\Phi^\top V^{-1}\Phi\right)^{-1}\Phi^\top V^{-1}\right]\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} m_i(\beta_0,\theta_0) + o_p(1).$$

**Proof of Lemma 1.** (i) Let $m_i(\beta,\theta) = (m_{1,i}(\beta,\theta), m_{2,i}(\beta,\theta))^\top$ and we will show convergence in probability element-wise. By (21) in Xu (2013), $\frac{1}{nh}\sum_{i=1}^{n} m_{1,i}^2(\beta_0,\theta_0) \xrightarrow{p} V_{11} = \tau(1-\tau)e_1^\top S^{-1}S^*S^{-1}e_1/\mu(x)$. Now consider the convergence to $V_{22}$. Equations (15)–(17) in Xu (2013) show that $\mathrm{E}[G((\beta - Y_i)/h_0) - \mathbb{I}(Y_i \le \beta)] = O(h_0)$ and $\mathrm{Var}[G((\beta - Y_i)/h_0) - \mathbb{I}(Y_i \le \beta)] = O(h_0)$, thus $G((\beta - Y_i)/h_0) = \mathbb{I}(Y_i \le \beta) + o_p(1)$. So by the finiteness of $\mathrm{E}|Y_i|^\varsigma$ for some $\varsigma > 2$ and the Hölder inequality,

$$\begin{aligned}\mathrm{E}(Y_i G((\beta - Y_i)/h_0)|X_i) &= \mathrm{E}(Y_i\mathbb{I}(Y_i \le \beta)|X_i) + o(1) \\ &= \int_{-\infty}^{\beta} y f(y|X_i)dy + o(1), \end{aligned} \tag{A.1}$$

$(= \theta_0\tau$ if $\beta = \beta_0)$ and similarly

$$\begin{aligned}\mathrm{E}([Y_i G((\beta - Y_i)/h_0)]^2|X_i) &= \mathrm{E}(Y_i^2\mathbb{I}(Y_i \le \beta)|X_i) + o(1) \\ &= \int_{-\infty}^{\beta} y^2 f(y|X_i)dy + o(1) := \xi(\beta, X_i)\tau + o(1), \end{aligned} \tag{A.2}$$

where $\xi(\beta, X_i) = \mathrm{E}(Y_i^2|Y_i \le \beta, X_i)$.

By Proposition 6.2 of Fan and Yao (2003),

$$W_p(u) = \mu(x)^{-1}(1 + O_p(h + \sqrt{\log n/(nh)})) \cdot K^*(u), \tag{A.3}$$

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

uniformly in $u$, where $K^*(u) = e_1^\top S^{-1}(1, u, \ldots, u^p)^\top K(u)$. It is easy to see that $\int K^*(u)^2 du = e_1^\top S^{-1} S^* S^{-1} e_1$. Then by (A.1) and (A.2), neglecting the smaller order terms,

$$\begin{aligned}
&\mathrm{E}\left(\frac{1}{h} m_{2,i}^2(\beta_0, \theta_0)\right) \\
&\quad = \mu(x)^{-2} \frac{1}{h} \mathrm{E}\left(K^*((X_i - x)/h)^2 \mathrm{E}([Y_i G((\beta_0 - Y_i)/h_0) - \tau\theta_0]^2 | X_i)\right) \\
&\quad = \mu(x)^{-2} \mathrm{E}\left(K^*((X_i - x)/h)^2 (\xi(\beta_0, X_i)\tau - 2\tau\theta_0 \textstyle\int_{-\infty}^{\beta_0} y f(y|X_i) dy + \tau^2\theta_0^2)/h\right) \\
&\quad = \mu(x)^{-2} \int K^*((z - x)/h)^2 \left(\xi(\beta_0, z)\tau - 2\tau\theta_0 \textstyle\int_{-\infty}^{\beta_0} y f(y|z) dy + \tau^2\theta_0^2\right) \mu(z) dz/h \\
&\quad = \mu(x)^{-2} \int K^*(v)^2 \left(\xi(\beta_0, x + vh)\tau - 2\tau\theta_0 \textstyle\int_{-\infty}^{\beta_0} y f(y|x + vh) dy + \tau^2\theta_0^2\right) \mu(x + vh) dv \\
&\quad \to e_1^\top S^{-1} S^* S^{-1} e_1 (\xi(\beta_0, x)\tau - \tau^2\theta_0^2)/\mu(x) = V_{22},
\end{aligned}$$

and thus by stationarity

$$\mathrm{E}\left(\frac{1}{nh} \sum_{i=1}^n m_{2,i}^2(\beta_0, \theta_0)\right) \to V_{22}. \tag{A.4}$$

Following the similar derivation above,

$$h\mathrm{Var}(h^{-1} m_{2,i}^2(\beta_0, \theta_0)) = \mathrm{E}(h^{-1} m_{2,i}^4(\beta_0, \theta_0)) - h[\mathrm{E}(h^{-1} m_{2,i}^2(\beta_0, \theta_0))]^2 \to \Upsilon,$$

where $\Upsilon = [\int_{-\infty}^{\beta_0} y^4 f(y|x) dy - 4\tau\theta_0 \int_{-\infty}^{\beta_0} y^3 f(y|x) dy + 6\tau^3\theta_0^2 \xi(\beta_0, x) - 3\tau^4\theta_0^4]\widetilde{\Upsilon}/\mu^3(x)$. In the last expression, $\widetilde{\Upsilon}$ is a constant involving functionals of $K(\cdot)$ and it, e.g., equals $\int K^4$ when $p = 0$. Note that

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{nh} \sum_{i=1}^n m_{1,i}^2(\beta_0, \theta_0)\right) \\
&\quad = \mathrm{Var}\left(\frac{1}{h} m_{2,1}^2(\beta_0, \theta_0)\right) \Big/ n + 2 \sum_{l=1}^{n-1} (1 - l/n) \left|\mathrm{Cov}\left(\frac{1}{h} m_{2,1}^2(\beta_0, \theta_0), \frac{1}{h} m_{2,l+1}^2(\beta_0, \theta_0)\right)\right| / n.
\end{aligned}$$

Following the arguments in Fan and Yao (2003, pp. 262–263, see also Masry and Fan, 1997), we can show that the second sum goes to zero. So

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{nh} \sum_{i=1}^n m_{1,i}^2(\beta_0, \theta_0)\right) &= \mathrm{Var}\left(\frac{1}{h} m_{2,1}^2(\beta_0, \theta_0)\right)/n + o(1) \\
&= O(1/(nh)) = o(1),
\end{aligned} \tag{A.5}$$

and therefore by (A.4) and (A.5), $\frac{1}{nh} \sum_{i=1}^n m_{2,i}^2(\beta_0, \theta_0) \xrightarrow{p} V_{22}$.

The convergence to $V_{12}$ can be shown similarly as above and is thus omitted. So Lemma 1(i) holds.

(ii) It suffices to show that for any $\delta \in \mathbb{R}^2$, $\frac{1}{\sqrt{nh}} \sum_{i=1}^n W_p((X_i - x)/h) \delta^\top \overline{m}_i(\beta_0, \theta_0) \xrightarrow{d} \mathcal{N}(0, \delta^\top V \delta)$, where $\overline{m}_i(\beta, \theta) = (G((\beta - Y_i)/h_0) - \tau, Y_i G((\beta - Y_i)/h_0) - \tau\theta)^\top$. Let

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

$\overline{\overline{m}}_i(\beta,\theta) = (\mathbb{I}(Y_i \le \beta) - \tau, Y_i\mathbb{I}(Y_i \le \beta) - \tau\theta)^\top$. Then by Theorem 6.3 in Fan and Yao (2003) (see also Masry and Fan, 1997) and the fact that $\mathrm{Var}(\delta^\top \overline{\overline{m}}_i(\beta_0,\theta_0)) = \delta^\top \overline{V}\delta$, which follows from the proof of Lemma 1(i), we have $\frac{1}{\sqrt{nh}}\sum_{i=1}^n W_p((X_i - x)/h)$ $\delta_i^\top \overline{\overline{m}}(\beta_0,\theta_0) \xrightarrow{d} \mathcal{N}(0, \delta^\top V\delta)$. Thus it remains to show that

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^n W_p((X_i - x)/h)\delta^\top[\overline{m}_i(\beta_0,\theta_0) - \overline{\overline{m}}_i(\beta_0,\theta_0)] = o_p(1). \tag{A.6}$$

Rewrite the left hand side of (A.6) as $\delta_1 Q_{1n} + \delta_2 Q_{2n}$, where $Q_{1n} = \frac{1}{\sqrt{nh}}\sum_{i=1}^n W_p$ $((X_i - x)/h)[G((\beta - Y_i)/h_0) - \mathbb{I}(Y_i \le \beta)]$ and $Q_{2n} = \frac{1}{\sqrt{nh}}\sum_{i=1}^n W_p((X_i - x)/h)$ $[Y_i G((\beta - Y_i)/h_0) - Y_i\mathbb{I}(Y_i \le \beta)]$. By Xu (2013) (treatment of $B_T$, Proof of the Lemma in Appendix), we have $Q_{1n} = o_p(1)$ and similarly $Q_{2n} = o_p(1)$. Thus (A.6) holds and the proof of Lemma 1(ii) is complete.

(iii) and (iv) The conclusion in these parts holds not only for $\beta_0$ and $\widetilde{\beta}$, but also for any $\overline{\beta}$ such that $\overline{\beta} \xrightarrow{p} \beta_0$ and $(nh)^{-1/2}\sum_{i=1}^n m_i(\overline{\beta},\theta_0) = O_p(1)$. These two statements are true for $\beta_0$ given Lemma 1(ii). The case for $\widetilde{\beta}$ can be proved using the arguments in Newey and Smith (2004, Lem. A1–A3 and Thm. 3.1, pp. 239–240) and is thus omitted.

(v) It holds given the uniform convergence of $\frac{1}{nh}\sum_{i=1}^n m_i(\beta,\theta_0)m_i(\beta,\theta_0)^\top$, which follows from Andrews (1987) given the smoothness of $m_i(\cdot,\theta_0)$ and the compactness of $\mathcal{B}$.

(vi) Note that $\widetilde{\lambda}$ satisfies the first-order condition

$$0 = \frac{1}{nh}\sum_{i=1}^n \dot{\rho}(\widetilde{\lambda}^\top \widetilde{m}_i)\widetilde{m}_i. \tag{A.7}$$

Let $\frac{\partial m_i(\beta,\theta_0)}{\partial\beta} = \big(W_p((X_i - x)/h)[g((\beta - Y_i)/h_0)/h_0], W_p((X_i - x)/h)[Y_i g((\beta - Y_i)/h_0)/h_0]\big)^\top \equiv \phi_i(\beta)$, where $g(\cdot) = G'(\cdot)$. Applying the envelope theorem to $\mathcal{L}(\beta,\theta_0) = \sup_{\lambda\in\Lambda_n} 2[\sum_{i=1}^n \rho(\lambda^\top m_i(\beta,\theta_0)) - n\rho(0)]$ (taking differentiation with respect to $\beta$), we have

$$0 = \widehat{\Phi}_1^\top \cdot \widetilde{\lambda}, \tag{A.8}$$

where $\widehat{\Phi}_1 = \frac{1}{nh}\sum_{i=1}^n \dot{\rho}\big(\widetilde{\lambda}^\top \widetilde{m}_i\big)\phi_i\big(\widetilde{\beta}\big)$. On the other hand, an expansion of (A.7) around $\big(\widetilde{\beta},\widetilde{\lambda}\big) = (\beta_0, 0)$ yields

$$0 = \frac{1}{nh}\sum_{i=1}^n \dot{\rho}(0)m_i\,(\beta_0,\theta_0) + \widehat{\Phi}_2\big(\widetilde{\beta} - \beta_0\big) + \overline{\Omega}_3\widetilde{\lambda}, \tag{A.9}$$

where

$$\widehat{\Phi}_2 = \frac{1}{nh}\sum_{i=1}^n \Big(\ddot{\rho}(\overline{\lambda}^\top m_i\,(\overline{\beta},\theta_0))\overline{\lambda}^\top m_i(\overline{\beta},\theta_0) + \dot{\rho}\Big(\overline{\lambda}^\top m_i\,(\overline{\beta},\theta_0)\Big)\Big)\phi_i(\overline{\beta}),$$

$$\overline{\Omega}_3 = \frac{1}{nh}\sum_{i=1}^n \ddot{\rho}(\overline{\lambda}^\top m_i\,(\overline{\beta},\theta_0))m_i\,(\overline{\beta},\theta_0)^\top m_i\,(\overline{\beta},\theta_0),$$

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

and $(\overline{\beta}, \overline{\lambda})$ is a point between $(\beta_0, 0)$ and $(\widetilde{\beta}, \widetilde{\lambda})$. Combining (A.8) and (A.9),

$$0 = \begin{pmatrix} 0 \\ \frac{1}{nh}\sum_{i=1}^{n} m_i(\beta_0, \theta_0) \end{pmatrix} + \widehat{M} \begin{pmatrix} \widetilde{\beta} - \beta_0 \\ \widetilde{\lambda} \end{pmatrix}, \quad \text{where } \widehat{M} = \begin{pmatrix} 0 & \widehat{\Phi}_1^{\top} \\ \widehat{\Phi}_2 & \overline{\Omega}_3 \end{pmatrix}. \tag{A.10}$$

Lemma 1(iv) and (v) implies $\overline{\Omega}_3 \xrightarrow{p} V$ and it is straightforward to show that $\widehat{\Phi}_1 \xrightarrow{p} \Phi$ and $\widehat{\Phi}_2 \xrightarrow{p} \Phi$. Thus, $\widehat{M}$ is invertible w.p.a.1. By solving (A.10) for $\sqrt{nh}(\widetilde{\beta} - \beta_0)$, we have

$$\sqrt{nh}(\widetilde{\beta} - \beta_0) = -\left(\Phi^{\top} V^{-1} \Phi\right)^{-1} \Phi^{\top} V^{-1} \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} m_i(\beta_0, \theta_0) + o_p(1).$$

From this and an expansion of $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} \widetilde{m}_i$ around $\widetilde{\beta} = \beta_0$,

$$\begin{aligned} \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \widetilde{m}_i &= \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} m_i + \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \phi_i(\overline{\overline{\beta}})(\widetilde{\beta} - \beta_0) \\ &= \left[ I_2 - \Phi \left(\Phi^{\top} V^{-1} \Phi\right)^{-1} \Phi^{\top} V^{-1} \right] \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} m_i(\beta_0, \theta_0) + o_p(1), \end{aligned} \tag{A.11}$$

where $\overline{\overline{\beta}}$ is a point between $\beta_0$ and $\widetilde{\beta}$. ■

Now we state the proofs of the main theorems. Let $\widehat{\lambda}$ and $\widetilde{\lambda}$ be as defined in Lemma 1. For expositional simplicity we write $m_i(\beta_0, \theta_0)$ and $m_i(\widetilde{\beta}, \theta_0)$ as $m_i$ and $\widetilde{m}_i$ respectively in what follows.

**Proof of Theorem 1.** The first-order condition of the optimization problem (10) with respect to $\widehat{\lambda}$ can be written as

$$0 = \frac{1}{nh} \sum_{i=1}^{n} \dot{\rho}(\widehat{\lambda}^{\top} m_i) m_i = \frac{\dot{\rho}(0)}{nh} \sum_{i=1}^{n} m_i - \widehat{\Omega}_1 \widehat{\lambda}, \tag{A.12}$$

where $\widehat{\Omega}_1 = -\frac{1}{nh}\sum_{i=1}^{n} \ddot{\rho}(\dot{\lambda}^{\top} m_i) m_i m_i^{\top}$, and the second equality of (A.12) follows from a Taylor expansion of $\dot{\rho}$ at $\widehat{\lambda} = 0$, and $\dot{\lambda}$ is a point between 0 and $\widehat{\lambda}$. The probability limit of $\widehat{\Omega}_1$ is straightforward to obtain. Indeed, $\widehat{\Omega}_1 = -\frac{1}{nh}\sum_{i=1}^{n} [\ddot{\rho}(\dot{\lambda}^{\top} m_i) + 1] m_i m_i^{\top} + \frac{1}{nh}\sum_{i=1}^{n} m_i m_i^{\top} := T_1 + T_2$. The term $T_2 \xrightarrow{p} V$ by Lemma 1(i) and now we examine the term $T_1$. Note that $|T_1| \le \max_{1 \le i \le n} |\ddot{\rho}(\dot{\lambda}^{\top} m_i) + 1| \cdot \frac{1}{nh}\sum_{i=1}^{n} |m_i m_i^{\top}|$, where the first factor on the right hand side $\xrightarrow{p} 0$ by Lemma 1(iii) and the assumption that $\ddot{\rho}(0) = -1$, and the second factor goes to a bounded matrix in probability by Lemma 1(i) and an application of Cauchy inequality to the nondiagonal elements. Thus $T_1 \xrightarrow{p} 0$, $\widehat{\Omega}_1 \xrightarrow{p} V$ and $\widehat{\Omega}_1$ is invertible w.p.a.1.

By (A.12) $\widehat{\lambda}$ can be written as $\widehat{\lambda} = \widehat{\Omega}_1^{-1} \frac{\dot{\rho}(0)}{nh} \sum_{i=1}^{n} m_i$. A second-order Taylor expansion of $\mathcal{L}(\beta_0, \theta_0) = 2\left(\sum_{i=1}^{n} \rho(\widehat{\lambda}^{\top} m_i) - n\rho(0)\right)$ at $\widehat{\lambda} = 0$ yields

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

$$\begin{aligned}\mathcal{L}(\beta_0,\theta_0) &= 2\dot{\rho}(0)\widehat{\lambda}^\top \sum_{i=1}^n m_i - \sqrt{nh}\widehat{\lambda}^\top \widehat{\Omega}_2 \cdot \sqrt{nh}\widehat{\lambda} \\ &= 2\dot{\rho}(0)\frac{\dot{\rho}(0)}{\sqrt{nh}}\sum_{i=1}^n m_i^\top \widehat{\Omega}_1^{-1}\frac{1}{\sqrt{nh}}\sum_{i=1}^n m_i \\ &\quad - \frac{\dot{\rho}(0)}{\sqrt{nh}}\sum_{i=1}^n m_i^\top \widehat{\Omega}_1^{-1}\widehat{\Omega}_2\widehat{\Omega}_1^{-1}\cdot\frac{\dot{\rho}(0)}{\sqrt{nh}}\sum_{i=1}^n m_i \\ &= \left(\frac{\dot{\rho}(0)}{\sqrt{nh}}\sum_{i=1}^n m_i\right)^\top \left[2\widehat{\Omega}_1^{-1} - \widehat{\Omega}_1^{-1}\widehat{\Omega}_2\widehat{\Omega}_1^{-1}\right]\left(\frac{\dot{\rho}(0)}{\sqrt{nh}}\sum_{i=1}^n m_i\right), \end{aligned} \tag{A.13}$$

where $\widehat{\Omega}_2 = -\frac{1}{nh}\sum_{i=1}^n \ddot{\rho}(\ddot{\lambda}^\top m_i)m_i m_i^\top \xrightarrow{p} V$ by the similar arguments for $\widehat{\Omega}_1$ and $\ddot{\lambda}$ is a point between 0 and $\widehat{\lambda}$. So we have $2\widehat{\Omega}_1^{-1} - \widehat{\Omega}_1^{-1}\widehat{\Omega}_2\widehat{\Omega}_1^{-1} \xrightarrow{p} V^{-1}$. Therefore Theorem 1 follows from the assumption that $\dot{\rho}(0) = -1$ and Lemma 1(ii). ■

**Proof of Theorem 2.** Let $\widetilde{\lambda} = \arg\max_{\lambda\in\Lambda_n} 2[\sum_{i=1}^n \rho(\lambda^\top m_i(\widetilde{\beta},\theta_0)) - n\rho(0)]$. By Lemma 1(iv) and the similar arguments used to derive (A.13) we obtain

$$\mathcal{L}(\theta_0) = \left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n \widetilde{m}_i\right)^\top \left[2\widetilde{\Omega}_1^{-1} - \widetilde{\Omega}_1^{-1}\widetilde{\Omega}_2\widetilde{\Omega}_1^{-1}\right]\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n \widetilde{m}_i\right), \tag{A.14}$$

where $\widetilde{\Omega}_1$ and $\widetilde{\Omega}_2$ are defined like $\widehat{\Omega}_1$ and $\widehat{\Omega}_2$ except that $\beta_0$ is replaced by $\widetilde{\beta}$. Similar to the arguments in the proof of Theorem 1, $2\widetilde{\Omega}_1^{-1} - \widetilde{\Omega}_1^{-1}\widetilde{\Omega}_2\widetilde{\Omega}_1^{-1} \xrightarrow{p} V^{-1}$ by Lemma 1(iv) and (v). Note that the asymptotic expansion of $\frac{1}{\sqrt{nh}}\sum_{i=1}^n \widetilde{m}_i$ is given in Lemma 1(vi).

By (A.14) and Lemma 1(vi) and (ii),

$$\begin{aligned}\mathcal{L}(\theta_0) &= \left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n m_i\right)^\top \left[I_2 - \Phi\left(\Phi^\top V^{-1}\Phi\right)^{-1}\Phi^\top V^{-1}\right]^\top V^{-1}\cdot \\ &\quad \cdot\left[I_2 - \Phi\left(\Phi^\top V^{-1}\Phi\right)^{-1}\Phi^\top V^{-1}\right]\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n m_i\right) + o_p(1) \\ &\xrightarrow{d} \xi^\top V^{1/2}\left[I_2 - \Phi\left(\Phi^\top V^{-1}\Phi\right)^{-1}\Phi^\top V^{-1}\right]^\top \\ &\quad \times V^{-1}\left[I_2 - \Phi\left(\Phi^\top V^{-1}\Phi\right)^{-1}\Phi^\top V^{-1}\right]V^{1/2}\xi \\ &= \xi^\top\left[I_2 - P\left(P^\top P\right)^{-1}P^\top\right]\xi \sim \chi^2(1), \end{aligned}$$

where $\xi \sim \mathcal{N}(0, I_2)$, $P = V^{-1/2}\Phi$ and we have used the fact that $I_2 - P\left(P^\top P\right)^{-1}P^\top$ is idempotent with rank one. Therefore the proof of Theorem 2 is complete. ■

**Proof of Theorem 3.** It can be similarly proved as for Theorem 2. ■

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

**Proof of Theorem 4.** Let $Z_i^\top \gamma_0 = X_i$ and $z^\top \gamma_0 = x$. Denote $\breve{m}_i(\beta,\theta)$ similarly as in $m_i(\beta,\theta)$ except that $(Z_i - z)^\top \widehat{\gamma}/h$ is used in place of $(X_i - x)/h$. Similar notations are generated (using the decoration $\breve{}$) on other quantities in the proofs of Theorems 1 and 2 whenever $(X_i - x)/h$ appears. These generated notations include $\breve{W}_p(u)$, $\breve{\breve{S}}$, etc. The proof is completed by showing that the steps (essentially Lemma 1(i)–(vi)) still go through with the generated notations. We will only show the case for Lemma 1(vi), that is,

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\breve{m}_i(\widetilde{\beta},\theta_0) - \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} m_i(\widetilde{\beta},\theta_0) = o_p(1), \tag{A.15}$$

and others follow similarly.

(A.15) is equivalent to

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\big[W_p((X_i - x)/h) - \breve{W}_p((Z_i - z)^\top \widehat{\gamma}/h)\big]\overline{m}_i(\widetilde{\beta},\theta_0) = o_p(1). \tag{A.16}$$

By Lemma 2, the equality (A.3), which holds for $W_p(u)$, also holds for $\breve{W}_p(u)$. To show (A.16), we will simply show that

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\big[K((X_i - x)/h) - K((Z_i - z)^\top \widehat{\gamma}/h)\big]\overline{m}_i(\widetilde{\beta},\theta_0) = o_p(1), \tag{A.17}$$

and other entries in (A.16) follow similarly.

Note that by the Taylor expansion, (A.17) holds if $(nh)^{-1/2}(\widehat{\gamma} - \gamma_0)^\top \sum_{i=1}^{n}[(Z_i - z)/h]K'((Z_i - z)^\top \overline{\gamma}/h)\overline{m}_i(\widetilde{\beta},\theta_0) = o_p(1)$, for some $\overline{\gamma}$ between $\widehat{\gamma}$ and $\gamma_0$, which holds if $\widehat{\gamma} - \gamma_0 = O_p((nh)^{-1/2})$ and

$$\sum_{i=1}^{n}\big[(Z_i - z)/h\big]K'\big((Z_i - z)^\top \overline{\gamma}/h\big)\overline{m}_i(\widetilde{\beta},\theta_0) = o_p(nh). \tag{A.18}$$

Note that by a standard weak LLN,

$$(nh)^{-1}\sum_{i=1}^{n}[(Z_i - z)/h]K'((Z_i - z)^\top \gamma_0/h)\overline{m}_i(\beta_0,\theta_0) = o_p(1),$$

since $\mathrm{E}(\overline{m}_i(\beta_0,\theta_0)|Z_i) = o_p(1)$ as $h_0 \to 0$. Then (A.18) follows from a uniform convergence of $(nh)^{-1}\sum_{i=1}^{n}[(Z_i - z)/h]K'((Z_i - z)^\top \gamma/h)\overline{m}_i(\beta,\theta_0)$ in $(\gamma,\beta)$, which holds given the compactness of $\Gamma \times \mathcal{B}$, using the arguments in showing (A.22). Thus (A.17) holds and the proof is complete. ■

LEMMA 2. *Let $\breve{W}_p(u)$ be denoted as in the proof of Theorem 4. Then* $\sup_{u\in\mathbb{R}}|W_p(u) - \breve{W}_p(u)| = o_p(1)$.

**Proof of Lemma 2.** We only need to show that

$$\widehat{S} - \breve{\breve{S}} = o_p(1). \tag{A.19}$$

(A.19) can be proved element by element, and we will show that for $0 \le k \le 2p$,

$$\breve{\breve{S}}_k = n^{-1}\sum_{i=1}^{n}[(Z_i - z)^\top \widehat{\gamma}/h]^k K((Z_i - z)^\top \widehat{\gamma}/h)/h \xrightarrow{p} f(z^\top \gamma_0)S_k. \tag{A.20}$$

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press

(A.20) follows from the probability limit of $\widehat{S}_k$ and

$$\sup_{\gamma\in\Gamma}|n^{-1}\sum_{i=1}^{n}[(Z_i-z)^\top\gamma/h]^k K((Z_i-z)^\top\gamma/h)/h - f(z^\top\gamma_0)S_k| = o_p(1). \tag{A.21}$$

We will now show (A.21) for $k=0$:

$$\sup_{\gamma\in\Gamma}|n^{-1}\sum_{i=1}^{n}K((Z_i-z)^\top\gamma/h)/h - f(z^\top\gamma)| = o_p(1). \tag{A.22}$$

Other cases follow similarly. (A.22) follows from (A.23) and (A.24):

$$\sup_{\gamma\in\Gamma}|n^{-1}\sum_{i=1}^{n}K((Z_i-z)^\top\gamma/h)/h - \mathrm{E}K((Z_i-z)^\top\gamma/h)/h| = o_p(1), \tag{A.23}$$

$$\sup_{\gamma\in\Gamma}|\mathrm{E}K((Z_i-z)^\top\gamma/h)/h - f(z^\top\gamma)| = o(1). \tag{A.24}$$

The uniform convergence in (A.24) is the one for a sequence of deterministic functions, so it follows (applying the theorem of dominance convergence) from the pointwise convergence and

$$\mathrm{E}K((Z_i-z)^\top\gamma/h)/h \overset{\substack{X_i\equiv Z_i^\top\gamma\\ x\equiv z^\top\gamma}}{=} \int_{\mathbb{R}} K((u-x)/h)f_{X_i}(u)du/h$$
$$\overset{v\equiv(u-x)/h}{=} \int_{-1}^{1} K(v)f_{X_i}(vh+x)dv < C,$$

given that for any $\gamma$, $f_{Z_i^\top\gamma} < C$. Now consider (A.23). It follows from the compactness of $\Gamma$, a pointwise weak LLN, and (A.25) and (A.26) (Andrews, 1987, Assump. A1–A3), applying the Theorem in Andrews (1987, p. 1466), where for any $n$ (thus any $h$)

$$\lim_{\rho\to 0}\sup_{\widetilde{\gamma}\in\Gamma,|\widetilde{\gamma}-\gamma|<\rho}\mathrm{E}K((Z_i-z)^\top\widetilde{\gamma}/h)/h = \mathrm{E}K((Z_i-z)^\top\gamma/h)/h, \tag{A.25}$$

$$\lim_{\rho\to 0}\inf_{\widetilde{\gamma}\in\Gamma,|\widetilde{\gamma}-\gamma|<\rho}\mathrm{E}K((Z_i-z)^\top\widetilde{\gamma}/h)/h = \mathrm{E}K((Z_i-z)^\top\gamma/h)/h. \tag{A.26}$$

(A.25) and (A.26) follow from the continuity of $\mathrm{E}K((Z_i-z)^\top\widetilde{\gamma}/h)/h$ in $\widetilde{\gamma}$ for any $h$, even though the function $K(\cdot)$ has discontinuities, given the assumption that $f_{Z_i^\top\widetilde{\gamma}}(\cdot)$ is continuous for any $\widetilde{\gamma}$. Thus (A.23) holds. So (A.22) holds. ■

https://doi.org/10.1017/S0266466614000802 Published online by Cambridge University Press