Statistica Sinica **20** (2010), 455-468

EVALUATION OF VALUE AT RISK: AN EMPIRICAL LIKELIHOOD APPROACH

Zhenghong Wei and Lixing Zhu

Shenzhen University and Hong Kong Baptist University

Abstract: To evaluate some Value at Risk models, the empirical likelihood approach to martingales is recommended. It turns out that the usual Wilks' theorem still holds in this case under mild conditions, and then it can be performed easily. Simulations were carried out for examining the performance of the new method.

Key words and phrases: Empirical likelihood, martingale, non-nested test, specification test, Value at Risk.

1. Introduction

The market crash in October 1987, recent crises in emerging markets, and disastrous losses resulting from trading activities of institutions such as Orange County, Long-Term Capital Management Fund, and Metallgesellschaft have increased the regulatory demand for reliable quantitative risk management tools. See, for example, Gallati (2003, Chap. 6) for a set of detailed case studies. The value-at-risk (VaR) concept has emerged as the most prominent measure of downside market risk. It places an upper bound on losses in the sense that these will exceed the VaR threshold with only a small target probability, typically chosen between 1% and 5%. In practice, the objective should be to provide a reasonable accurate estimate of risk at a reasonable cost. This involves choosing a method from among the various industry standards that is most appropriate for the portfolio at hand.

Given the importance of VaR estimates to banks and to their regulators, evaluating the accuracy of the models underlying them is a necessary exercise. For this the common approaches are the test of unconditional coverage and the test of conditional coverage proposed by Kupiec (1995) and Christoffersen (1998). In addition, the distribution forecast method proposed by Crnkovic and Drachman (1996) examined whether observed empirical quantiles derived from a VaR model's distribution forecast are independent and uniformly distributed; Lopez (1997) proposes an alternative evaluation method that is based on the probability forecasting framework presented; a dynamic quantile test is proposed by Engle and Manganelli (2004); Fan and Gu (2003) applied mean absolute deviation and

square root absolute deviation criteria; Chen and Wong (2005) introduced an kernel smoothed empirical likelihood for a quantile of weekly depedent processes based on blocks of data. The most related work to our study is Christoffersen, Hahn and Inoue (2001), who proposed specification testing and nonested testing based on the Kullback-Leibler information criterion; some data analysis was conducted for illustration, but there were no simulation results to examine the performance of their tests. Here, roughly speaking, specification testing tests whether the underlying volatility model is really the hypothetical model, and nonnested testing compares the performance of two models. In the next section, we shall have some more details.

In this paper, we investigate how to employ an empirical likelihood method for evaluation of VaR models. The empirical likelihood method was first introduced by Owen (1988, 1990) for constructing confidence regions. Hall and LaScala (1990) summarized some of its advantages over the traditional approaches: the empirical likelihood regions are automatically shaped by the sample, they are Bartlett correctable, range preserving, and transformation respecting, etc. For these reasons, the empirical likelihood has many applications in smooth functions of means (DiCiccio, Hall and Romono (1991)), in nonparametric density and regression function estimation (Chen (1996), Chen and Qin (2000), Xue and Zhu (2007), and Zhu and Xue (2006)), in quantile related estimation (Chen and Hall (1993)), among others. For a more comprehensive review of the empirical likelihood method and its applications, the reader is referred to the monograph by Owen (2001).

For VaR models, we need to deal with the random variables that can form martingale difference sequence. Thus, we first consider the empirical likelihood ratio to martingale difference sequence to show the Wilks' theorem. The result will then be applied to evaluating VaR models.

The paper is arranged as follows. In Section 2, we apply the empirical likelihood method for martingale difference sequence to evaluate some VaR models. In Section 3, we report on simulation studies and a data analysis conducted to examine the performance of our method. We do not give the technical details of the proofs for our theoretical results, they are available from the authors upon request.

2. Empirical Likelihood for Value at Risk Evaluation

Consider VaR modeling first. Let S_t be the price of a portfolio at time t. Let

$$r_t = \log S_t - \log S_{t-1} = \log \frac{S_t}{S_{t-1}}$$

be the observed log return at time t. VaR measures the extreme loss of a portfolio over a predetermined holding period τ with a prescribed confidence level $1-p$. More precisely, VaR_t is defined to be the solution to

$$P(r_t \leq VaR_t|\mathcal{F}_{t-1}) = p, \tag{2.1}$$

where $\mathcal{F}_t = \sigma(S_s, s \leq t)$ represents the historical information available at time t. Clearly, to have p free of t, stationarity of the sequence of r_t is needed. Such a conditional expectation equation indicates that financial returns exhibit nonstandard statistical properties with non-IID property: they are not independently and identically distributed and moreover, and may not be Gaussian. This is reflected by three widely reported stylized facts: (i) volatility clustering, (ii) substantial kurtosis or fat-tailed distributions, (iii) mild skewness of the returns, possibly of a time-varying nature. What we have is often only the above conditional expectation equation (2.1). In order to facilitate these factors, a more general model is given by

$$r_t = \mu_t + \sigma_t \varepsilon_t, \tag{2.2}$$

where μ_t and σ_t are both $\mathcal{F}_{t-1}$-measurable, and ε_t are IID random variables with mean zero and variance one. Here, the distribution of ε_t can be taken to be other distributions than Gaussian, e.g., they could be skewed and/or with fat tails such as the Student's t, and the generalized asymmetric t, see Bollerslev (1986).

Write the conditional distribution function of r_t given $\mathcal{F}_{t-1}$ as $F_t(w) = P(r_t \leq w|\mathcal{F}_{t-1})$. It follows from (2.1) and the location-scale-model that $VaR_t(\beta) = F_t^{-1}(p) := \mu_t + \beta\sigma_t$, where β is the pth quantile of ε_t, i.e., $P(\varepsilon_t \leq \beta) = p$. Note that in some commonly used VaR models (GARCH, Riskmetrics, GJR and History simulation), we mainly consider modelling for volatility. Thus the mean μ_t is fixed when we consider specification and nonnested testing in the following, without loss of generality, $\mu_t = 0$. Then $VaR_t(\beta) = \beta\sigma_t$ and the calculation of VaR consists of two parts: determination of quantiles, β, and estimation of volatilities, σ_t.

Trivially, (2.1) implies that

$$E\Big(I\{r_t \leq VaR_t(\beta)\} - p|\mathcal{F}_{t-1}\Big) = 0. \tag{2.3}$$

If $\{z_{t-1}, z_{t-2}, \ldots\}$ are $\mathcal{F}_{t-1}$-measurable, where $\mathcal{F}_{t-1}$ is the time $t-1$ information set, then (2.3) yields

$$E\bigg\{\Big(I\{r_t \leq VaR_t(\beta)\} - p\Big)k(z_{t-1}, z_{t-2}, \ldots)|\mathcal{F}_{t-1}\bigg\} = 0 \tag{2.4}$$

for every measurable vector-valued function $k(\cdot)$ of $\{z_{t-1}, z_{t-2}, \ldots\}$. The r.v.'s $\{z_{t-1}, z_{t-2}, \ldots\}$ are referred to as instrumental variables. Write

$$Y_t = \Big(I\{r_t \le VaR_t(\beta)\} - p\Big)k(z_{t-1}, z_{t-2}, \ldots) \equiv I_t k(\mathbf{z}_{t-1}),$$

where $\mathbf{z}_{t-1} = (z_{t-1}, z_{t-2}, \ldots) \in \mathcal{F}_{t-1}$. Note that $\{Y_t : -\infty < t \le 0\}$ is a martingale difference sequence. Similar result can be derived when $k(\cdot)$ is a non-constant random vector so that we can choose several functions of instrumental variables. for notational clarity, we write

$$\mathbf{Y_t} = \Big(I\{r_t \le VaR_t(\beta)\} - p\Big)\mathbf{k}(z_{t-1}, z_{t-2}, \ldots) \equiv I_t \mathbf{k}(\mathbf{z}_{t-1}),$$

where $\mathbf{k}(\mathbf{z}_{t-1}) = (k_1(\mathbf{z}_{t-1}), \ldots, k_d(\mathbf{z}_{t-1}))$ and $E\{\mathbf{Y_t}|\mathcal{F}_{t-1}\} = 0$, i.e., $\{\mathbf{Y}_t, t \ge 1\}$ forms a d-dimensional martingale difference sequence.

Here we use this sequence to evaluate VaR model using a specification test constructed by empirical likelihood ratio. More precisely, we assign the probability mass p_t to the point $\mathbf{Y}_t$, and then construct empirical likelihood based on $\mathbf{Y}_t$ as

$$\rho_n := -2\max\Big\{\sum_{t=1}^{n}\log(np_i)\Big| p_t \ge 0,\ \sum_{t=1}^{n} p_t = 1,\ \sum_{t=1}^{n} p_t \mathbf{Y_t} = \mathbf{0}\Big\}.$$

By the Lagrange multiplier method, $\rho_n = 2\sum_{t=1}^{n}\log(1+\lambda\cdot\mathbf{Y_t})$, where λ satisfies $\sum_{t=1}^{n}[\mathbf{Y_t}/(1+\lambda\cdot\mathbf{Y_t})] = 0$. The following theorem is applied to VaR evaluation in the next sections.

Theorem 1. *Assume that the sequence* $\{\mathbf{Y_t} : -\infty < t \le 0\}$ *is strongly stationary and that the third moment of* $\|\mathbf{Y_t}\|$ *is finite, where* $\|\cdot\|$ *is the Euclidian norm. Then* ρ_n *is asymptotically chi-square with* d *degrees of freedom.*

The basic idea to prove this theorem can be easily seen through the proof when $d = 1$. We can show that ρ_n is asymptotically equivalent to

$$\frac{\Big((1/\sqrt{n})\sum_{t=1}^{n} Y_t\Big)^2}{(1/n)\sum_{t=1}^{n} Y_t^2}$$

and the Lindeberg Central Limit Theorem for martingales (see Hall and Heyde (1980)) implies the desired result. We do not present the details of the proof here, it is available from the authors upon request.

When we consider nonnested testing to compare two VaR models, we can use the martingale difference sequence formed by

$$\mathbf{Y}_t = I\Big\{r_t \le VaR_t(\beta)\Big\} - I\Big\{r_t \le VaR_t^1(\beta)\Big\}\mathbf{k}(z_{t-1}, z_{t-2}, \ldots)$$

to construct an empirical likelihood ratio as we did for specification testing. The asymptotic properties are almost identical and we do not give the details here.

3. Simulation Studies and Empirical Analysis

3.1. Estimation of volatilities

Volatilities play an important role in the calculation of VaR. Risk managers have a plethora of volatility measures to choose from when calculating Value-at-Risk measures. We give a brief review of some of these methods below.

In light of the observed volatility clustering, conditionally heteroskedastic parametric models, which allow the scale parameter to be a function of past information, are frequently used. Arguably the most popular formulation is autoregressive conditional heteroskedasticity (ARCH) (Engle (1982)) model and its generalization, GARCH model (Bollerslev (1986)). ARCH relates the error variance to the square of a previous period's error. It is commonly employed in modeling financial time series that exhibit time-varying volatility. On the other hand, GARCH models error variances by an autoregressive moving average model. Some of the approaches given below are related to GARCH models. Since these volatility forecasting models were introduced, there have been many alternatives/modifications proposed to better their use in volatility forecasting. In the following we briefly introduce four models with which we do specification testing and nonnested testing.

1. **GARCH(r, s) model** (Bollerslev (1986)). The conditional variance is

$$r_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = c_0 + \sum_{i=1}^{r} c_i r_{t-i}^2 + \sum_{j=1}^{s} d_j \sigma_{t-j}^2,$$

where the ε_t are IID with mean 0 and variance 1, $c_0 > 0$, $c_i \geq 0$, $d_j \geq 0$, and $\sum_{i=1}^{r} c_i + \sum_{j=1}^{s} d_j < 1$.

2. **RiskMetrics**. The benchmark measure advocated in Morgan (1996) is RiskMetrics. It sets the conditional mean constant, and specifies the variance as $r_t = \sigma_t \varepsilon_t$, $\sigma_t^2 = (1-\lambda) r_{t-1}^2 + \lambda \sigma_{t-1}^2$, where λ is simply set to 0.94 for daily data. By iteration, one can easily show that $\sigma_t^2 = (1-\lambda)(r_{t-1}^2 + \lambda t_{t-2}^2 + \lambda^2 t_{t-3}^2 + \cdots)$, an example of exponential smoothing in the time domain. For this reason, the model is often refereed to as the Exponentially Weighted Moving Average (EWMA) model. It is similar to, but different from, GARCH(1, 1) since, in EWMA, one has $c_0 = 0$, and $\alpha + \beta = 1$. The EWMA is also refereed to as the IGARCH(1, 1) model in the literature.

3. **GJR Models** (Glosten, Jagannathan and Runkle (1993)). Investors usually react differently as the markets move up and down. Typically, the markets

become more volatile as prices move down. To allow for asymmetric effects between positive and negative asset returns, one could consider using GJR models. The general GJR(r, s) model for the conditional variance of the innovations with leverage terms is

$$r_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = c_0 + \sum_{i=1}^{r} c_i r_{t-i}^2 + \sum_{j=1}^{s} d_j \sigma_{t-j}^2 + \sum_{i=1}^{r} L_i r_{t-i}^2 I\{r_{t-i} < 0\}.$$

For the special case $r = s = 1$, it can be seen that, if $L_1 \neq 0$, the model shows asymmetric effects between positive and negative asset returns. If $L_1 > 0$, we say that there is a leverage effect.

4. **History Simulation.** History simulation is the simplest and most transparent method of calculation. One calculates the standard error from some fixed number of past observations before time t as an estimate of the standard error at time t. The benefits of this method are its simplicity, and the fact that it does not assume a normal distribution of asset returns. Drawbacks are the requirement for a large market database, and the computationally intensive calculation.

3.2. Simulation studies

In this section, we report on some simulations conducted to study the performances of the proposed empirical likelihood methods. Consider data generated from the following GARCH(1, 1) model, regarded as the hypothetical model in the testing procedure

$$r_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = 0.0004 + 0.12 r_{t-1}^2 + 0.85 \sigma_{t-1}^2, \tag{3.1}$$

where the ε_t are IID $N(0, 1)$ r.v.'s. All simulation results were based on $B = 2,000$ repetitions. For each, $4,000$ observations were generated from the GARCH(1, 1) model, the first $2,000$ in-sample observations were used to estimate the model while the remaining $2,000$ out-of-sample observations were used for testing purposes.

Four different volatility models were employed for comparisons: Historical Simulations (HistSimu), RiskMetrics, GARCH$(1, 1)$, and GJR$(1, 1)$. For historical simulations, we used the standard error of the past 500 observations before time t to forecast the standard error at time $t + 1$. We used the first lag of the four volatility measures and the return as our instrumental variables.

3.2.1. Specification testing

Apart from the four models mentioned, we also include a few more models for comparison. More precisely, we added some noise terms to the GARCH(1, 1)

Table 3.1. Specification testing by the EL method (no instrumental variables).

VaR measure	$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.25$	average	PR
GARCH(1, 1)	42	46	61	52	39	48	2.4%
HistSimu	431	25	147	284	299	237.2	11.9%
RiskMetrics	121	55	29	20	15	48	2.4%
GJR(1, 1)	41	43	60	50	34	46	2.3%
DISTURB1	42	46	61	52	39	48	2.4%
DISTURB2	42	49	54	52	36	47	2.3%
DISTURB3	1,448	542	197	90	36	46	23%
DISTURB4	2,000	2,000	2,000	2,000	1,803	1,961	98%

Table 3.2. Specification testing by the EL method (one instrumental variable).

VaR measure	$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.25$	average	PR
GARCH(1, 1)	58	16	20	15	22	26.2	1.3%
HistSimu	38	26	30	29	22	29	1.5%
RiskMetrics	35	20	20	12	25	22.4	1.1%
GJR(1, 1)	66	25	24	18	26	31.8	1.6%
DISTURB1	58	16	20	15	22	26.2	1.3%
DISTURB2	59	18	23	16	23	27.8	1.4%
DISTURB3	34	12	15	18	20	19.8	0.99%
DISTURB4	20	20	21	19	14	18.8	0.9%

model (3.1), that can be roughly expressed as follows:

$$DISTURB(i) = GARCH(1,1) + Normal(0, a_i\sigma^2), \qquad i = 1, \ldots, 4,$$

where σ^2 was taken to be the absolute value of the smallest values of volatility coming from GARCH(1, 1), and a_i was chosen to be 0.0001, 0.01, 0.3, and 1.5, respectively. Clearly as a_i gets larger, we have added more noise to the true model.

The simulation results for specification testing are presented in Tables 3.1−3.3 where the nominal level is 0.05. Table 3.1 reports the result without any instrumental variables, while Tables 3.2 and 3.3 include the results with one and five instrumental variables, respectively. For the case with one instrumental variable, we chose the first lag of the return; for the case with five instrumental variables we chose the first lag of the four volatilities of the selected models and the return. The entries in the tables are the number of rejections, with the last two volume being the average and the percentages of rejections(PR) using various methods. The following observations can be made from these tables.

1. When the data come from the true model, we would certainly hope that the number of rejections would be small. This is confirmed from our simulation

Table 3.3. Specification testing by the EL method (five instrumental variables).

VaR measure	$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.25$	average	PR
GARCH(1, 1)	317	57	54	38	33	99.8	4.99%
HistSimu	1,959	1,989	1,989	1,981	1,749	1,933	96.7%
RiskMetrics	1,392	1,356	1,241	1,017	437	1,087	54.4%
GJR(1, 1)	341	69	66	50	37	112.6	5.6%
DISTURB1	318	57	54	38	33	100	5%
DISTURB2	320	52	56	41	31	100	5%
DISTURB3	1,212	320	109	60	25	345.2	17.3%
DISTURB4	2,000	2,000	2,000	2,000	1,446	1,889.2	94.5%

studies. In all the three tables, when the true model was GARCH(1, 1), the percentages of rejections were 2.4%, 1.3%, and 4.99%, respectively. That is, when no or one instrumental variable was included, the tests were conservative, while the test with five instrumental variables maintained the significance level.

2. If we use the other methods to estimate the volatility, we can see the effects of adding instrumental variables. For History Simulation, the percentages of rejections from Tables 3.1−3.3 were 11.9%, 1.5%, and 96.7%, respectively. So adding one instrumental variables is worse than having no instrumental variables, but both were much worse than adding five instrumental variables, which gave power 96.7%.
3. If we compare the four models DISTURB(i) for $i = 1, 2, 3, 4$, we see from Tables 3.1 and 3.3 that the percentages of rejections tended to increase as the noise got larger (There is not much change visible in Table 3.2).
4. For the GJR (1,1) model, we see from Tables 3.1−3.3 that the percentages of rejections were all very low, 2.3%, 1.6%, and 5.6%, respectively. Again when no or one instrumental variable was included, the tests were conservative, while the test with five instrumental variables maintained the significance level. This is because GJR(1,1) is also a GARCH-type model.
5. Intuitively, history simulation is the roughest volatility model. Therefore, the percentages of rejections were usually larger than with such other methods as RiskMetrics, GJR(1, 1). This can be seen from all the three tables.
6. Note that RiskMetrics also belongs to the class of GARCH-type models. Note that the percentages of rejections were usually quite small in Tables 3.1 and 3.2. However, the proportions of rejections are quite high, indicating that there is a great advantage in adding five instrumental variables.
7. In summary, the tests with five instrumental variables seem to perform the best.

Table 3.4. The results with CHI's specification test.

VaR measure	$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.25$	average	PR
GARCH(1, 1)	281	29	20	17	15	72.4	3.6%
HistSimu	1,711	1,886	1,860	1,789	1,275	1,749.4	87.5%
RiskMetrics	733	608	543	434	167	497	24.9%
GJR(1, 1)	293	33	26	20	14	77.2	3.9%
DISTURB1	280	29	20	17	15	72.2	3.6%
DISTURB2	290	32	18	15	15	74	3.7%
DISTURB3	169	36	18	15	15	50.1	2.5%
DISTURB4	2,000	2,000	2,000	1,862	63	1,585	79.3%

Table 3.5. Nonnested testing by the EL method (no instrumental variables).

VaR	0.01	0.05	0.10	0.15	0.25	average	PR
GARCH(1, 1) vs HistSimu	916	488	807	1,093	1,364	933.6	46.7%
GARCH(1, 1) vs RiskMetrics	485	643	577	561	437	540.6	27%
GARCH(1, 1) vs GJR(1, 1)	756	547	440	472	454	533.8	26.7%
HistSimu vs RiskMetrics	588	741	1078	1,275	1,428	1,022	51.1%
HistSimu vs GJR(1, 1)	907	482	804	1,086	1,367	929.2	46.5%
RiskMetrics vs GJR(1, 1)	460	625	584	555	450	534.8	26.7%
GARCH(1, 1) vs DISTURB1	18	42	50	52	32	38.8	1.9%
GARCH(1, 1) vs DISTURB4	2,000	2,000	2,000	2,000	1,812	1,962.4	98.1%

8. Compared with the specification test proposed by Christoffersen, Hahn and Inoue (2001), we can easily see from the results reported in Table 3.4 that the tests with five instrumental variables outperformed it significantly.

3.2.2. Nonnested testing

Recall that the data were generated from GARCH(1, 1). We compared the four volatility models and at the same time, we compared amongst GARCH(1, 1) and DISTURB1, DISTURB4.

The simulation results for nonnested testing are reported in Tables 3.5−3.8. The results of Table 3.5 do not have any instrumental variables while those of Table 3.6 include one instrumental variable which are the first lag of the return, Tables 3.7 for three instrumental variables which are the first lag of the return and the two volatilities compared, and Table 3.8 for five instrumental variables which are the first lag of the return and the four volatilities.

The entries in the tables are the numbers of rejections, the last two columns are the average and the percentages of rejections(PR) using various methods. The following observations can be made.

Table 3.6. Nonnested testing by the EL method(one instrumental variable).

VaR	0.01	0.05	0.10	0.15	0.25	average	PR
GARCH(1, 1) vs HistSimu	42	25	27	20	19	26.6	1.3%
GARCH(1, 1) vs RiskMetrics	95	35	25	39	34	45.6	2.3%
GARCH(1, 1) vs GJR(1, 1)	899	869	835	837	860	860	43%
HistSimu vs RiskMetrics	37	24	27	20	19	25.4	1.3%
HistSimu vs GJR(1, 1)	42	33	34	33	29	34.2	1.7%
RiskMetrics vs GJR(1, 1)	127	55	58	63	60	72.6	3.6%
GARCH(1, 1) vs DISTURB1	18	42	50	52	32	38.8	1.9%
GARCH(1, 1) vs DISTURB4	17	22	20	20	19	19.6	0.99%

Table 3.7. Nonnested testing by the EL method (three instrumental variables).

VaR	0.01	0.05	0.10	0.15	0.25	average	PR
GARCH(1, 1) vs HistSimu	2000	2000	2000	2000	2000	2000	100%
GARCH(1, 1) vs RiskMetrics	940	757	703	734	615	749.8	37.5%
GARCH(1, 1) vs GJR(1, 1)	1243	1440	1450	1431	1438	1400.4	70.9%
HistSimu vs RiskMetrics	2000	2000	2000	2000	2000	2000	100%
HistSimu vs GJR(1, 1)	2000	2000	2000	2000	2000	2000	100%
RiskMetrics vs GJR(1, 1)	930	765	738	783	638	770.8	38.5%
GARCH(1, 1) vs DISTURB1	16	42	50	52	33	39	1.95%
GARCH(1, 1) vs DISTURB4	2000	2000	2000	2000	1588	1917.60	95.9%

Table 3.8. Nonnested testing by the EL method (five instrumental variables).

VaR	0.01	0.05	0.10	0.15	0.25	average	PR
GARCH(1, 1) vs HistSimu	2000	2000	2000	2000	2000	2000	100%
GARCH(1, 1) vs RiskMetrics	2000	2000	2000	2000	2000	2000	100%
GARCH(1, 1) vs GJR(1,1)	1281	1675	1770	1784	1752	1652.4	83.6%
HistSimu vs RiskMetrics	2000	2000	2000	2000	2000	2000	100%
HistSimu vs GJR(1,1)	2000	2000	2000	2000	2000	2000	100%
RiskMetrics vs GJR(1,1)	2000	2000	2000	2000	2000	2000	100%
GARCH(1, 1) vs DISTURB1	18	42	50	52	33	39	1.95%
GARCH(1, 1) vs DISTURB4	2000	2000	2000	2000	1425	1885	94.3%

1. The tests with only one instrumental variable had the lowest rejection rates. In fact, the rejection rates were so low in almost all cases that the tests failed to distinguish among various models. On the other hand, tests with no instrumental variables, or three or five instrumental variables, are all did much better. This indicates that one should be very careful in choosing the instrumental variables if very few of them are being selected.
2. Except for Table 3.6, the largest three rejection rates were between History Simulation and GARCH(1, 1), between History Simulation and GJR(1, 1),

Table 3.9. The results with CHI's Nonnested test.

VaR	0.01	0.05	0.10	0.15	0.25	average	PR
GARCH(1, 1) vs HistSimu	1,000	1,310	1,295	1,030	347	996.4	49.8%
GARCH(1, 1) vs RiskMetrics	175	23	11	4	0	42.6	2.1%
GARCH(1, 1) vs GJR(1, 1)	12	0	0	0	0	2.4	0%
HistSimu vs RiskMetrics	80	0	0	0	0	16	0.8%
HistSimu vs GJR(1, 1)	53	0	0	0	0	10.6	0.5%
RiskMetrics vs GJR(1, 1)	62	0	0	0	0	12.4	0.6%
GARCH(1, 1) vs DISTURB1	0	0	0	0	0	0	0%
GARCH(1, 1) vs DISTURB4	1,981	2,000	1,990	914	0	1,377	68.9%

and between History Simulation and RiskMetrics. One possible explanation is as follows. RiskMetrics and GJR(1, 1) models both belong to the GARCH-type models while History Simulation does not. Therefore, it is not surprising that tests can pick up the differences between History Simulation and the other three models.

3. Although RiskMetrics and GJR(1, 1) models both belong to the GARCH-type models, our tests can still pick up the differences among them. In fact, we were able to tell them apart around 26.7% of the time when no instrumental variables were included, and these rejection rates increased substantially when three or five instrumental variables were included. Similar phenomenon happened between GARCH(1, 1) and RiskMetrics, and between RiskMetrics and GJR(1, 1).
4. If we compare GARCH(1, 1) models and DISTURB(1), we see from Tables 3.5−3.8 that the proportion of rejections was very low, but between GARCH(1, 1) models and DISTURB(4) the proportion of rejections was very high, except for one instrumental variable. This agrees with our intuition since, by construction, the DISTURB(1) model differs only little from the GARCH(1, 1) model, while the DISTURB(4) model differs more from GARCH(1, 1) than DISTURB(1).
5. In summary, we see that tests generally performed very well in all cases except for the case where only one instrumental variable was used. When we chose three or five instrumental variables, the performances seemed to improve a great deal. One should also be careful when selecting very few instrumental variables.
6. We also compared our tests with the one proposed by Christoffersen, Hahn and Inoue (2001). When three or five instrumental variables were chosen, our tests worked much better.

Table 3.10. Specification testing by the EL method (no instrumental variables).

VaR	0.01	0.05	0.10	0.15	0.25
GARCH(1, 1)	14.6	9.2	3.87*	0.30*	0.02*
History Simulation	20.2	4.2*	0.77*	0.65*	1.04*
RiskMetrics	18.8	11.5	5.31*	2.73*	0*

($CV = 6.635$ at 1% level.)

Table 3.11. Specification testing by the EL method (one instrumental variable).

VaR	0.01	0.05	0.10	0.15	0.25
GARCH(1, 1)	1.26*	0.02*	0.21*	0.71*	0.61*
History Simulation	0.01*	0.65*	0.25*	0.04*	0.79*
RiskMetrics	0.83*	1.07*	1.92*	0.80*	2.49*

($CV = 6.635$ at 1% level.)

Table 3.12. Specification testing by the EL method (four instrumental variables).

VaR	0.01	0.05	0.10	0.15	0.25
GARCH(1, 1)	15.2	13.98	8.04*	13.11*	17.47
History Simulation	36.1	39.66	16.59	11.29*	6.67*
RiskMetrics	21.2	13.82	15.47	21.40	26.80

($CV = 13.277$ at 1% level.)

3.3. Empirical analysis

We used some data from the stock markets in the USA to do some comparisons. The data are the S&P500 index from Jan. 1, 1997 to Dec. 1, 2006, obtained from www.finance.com. Recall that daily return is defined as $r_t = \ln(S_t) - \ln(S_{t-1})$, where S_t is the daily closing price at day t. The numerical results are presented in Tables 3.10−3.15. The entries in all tables are the observed value of the test statistics. We considered testing at 1% only. Critical values (CV) at these levels are given with the tables as well. In all tables, we use "$*$" to indicate "do not reject at 1%" .

For specification testing, we make the following remarks about Tables 3.10−3.12.

1. First we look at the Table 3.10. With $p = 0.01$, all models were rejected. When $p \neq 0.01$, all models were not rejected except for GARCH(1, 1). From Table 3.11, since value of statistics are very low, all models were not rejected.
2. Now we look at Table 3.12. With $p = 0.01, 0.05$, all models were rejected. When $p \neq 0.01, 0.05$, the results were mixed.

Table 3.13. nonnested testing by the EL method (no instrumental variables).

VaR	0.01	0.05	0.10	0.15	0.25
GARCH(1, 1) vs History Simulation	277.26	866.43	866.43	589.18	1,143.7
GARCH(1, 1) vs RiskMetrics	7.36	17.45	21.07	11.64	11.6
History Simulation vs RiskMetrics	0*	1.1*	2.04*	0.05*	8.1

($CV = 6.635$ at 1% level.)

Table 3.14. nonnested testing by the EL method (one instrumental variable).

VaR	0.01	0.05	0.10	0.15	0.25
GARCH(1, 1) vs History Simulation	0.30*	0.98*	2.08*	0.01*	7.0
GARCH(1, 1) vs RiskMetrics	1.56*	0.48*	0.34*	2.49*	0*
History Simulation vs RiskMetrics	0.96*	1.69*	1.49*	1.72*	6.3*

($CV = 6.635$ at 1% level.)

Table 3.15. nonnested testing by the EL method (four instrumental variables).

VaR	0.01	0.05	0.10	0.15	0.25
GARCH(1, 1) vs History Simulation	277.23	866.27	866.34	589.15	1143.4
GARCH(1, 1) vs RiskMetrics	346.42	896.04	724.26	827.35	1021
History Simulation vs RiskMetrics	409.94	788.02	616.05	645.61	974.6

($CV = 6.635$ at 1% level.)

For nonnested testing at 1% significance levels, we make the following remarks from Tables 3.13−3.15. First we compare with Table 3.3 and Table 3.14. In Table 3.13, since values of statistics were very low, all null hypothesises were not rejected, but as reported in Table 3.14, all null hypothesises were rejected. At Table 3.15, we conclude that for all p there were significant differences between GARCH(1, 1) and history simulation, between GARCH(1, 1) and RiskMetrics.

Acknowledgement

The research described here was supported by a grant from Research Grants Council of Hong Kong. The authors thank the Editor, an associate editor, and two referees for their constructive suggestions and comments which led to a significant improvement of presentation.

References

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econometrics* **31**, 307-27.

Chen, S. X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **83**, 329-341.

Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.* **21**, 1166-1181.

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

Chen, S. X. and Wong, C. M. (2005). Smoothed block empirical likelihood for quantiles of weekly dependent processes. Manuscript.

Christoffersen, P. (1998). Evaluating interval forecasts. *Internat. Econom. Rev.* **39**, 841-862.

Christoffersen, P., Hahn, J. and Inoue, A. (2001). Testing and comparing value-at-risk measures. *J. Empirical Finance* **8**, 325-342.

Crnkovic, C. and Drachman, J. (1996). Quality control. *Risk* **9**, 139-143.

DiCiccio, T. S., Hall, P. and Romono, J. (1991). Empirical likelihood is Bartlett correctable. *Ann. Statist.* **19**, 1053-1061.

Engle, R. (1982). Autogressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987-1007.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *J. Bus. Econom. Statist.* **22**, 367-381.

Fan, J. and Gu, J. (2003). Semiparametric estimation of value at risk. *Econom. J.* **6**, 261 - 290.

Gallati, R. R. (2003). *Risk Management and Capital Adequacy*. McGraw-Hill, New York.

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

Hall, P. and LaScala, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.* **58**, 109-127.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *J. Derivatives* **3**, 73-84.

Lopez, J. (1997). Regulatory evaluation of value-at-risk models. *J. Risk* **23**, 470-472.

Morgan, J. P. (1996). *Riskmetrics-Technical Document*. 4th edition. Morgan Guaranty Trust Company, New York.

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. *Ann. Statist.* **18**, 90-120.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall, London.

Xue, L. G. and Zhu, L. X. (2007). Empirical likelihood for a varying coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* **102**, 642-654.

Zhu, L. X. and Xue, L. G. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. Roy. Statist. Soc. Ser. B* **68**, 549-570.

College of Mathematics and Computational Science, Shenzhen University, Shenzhen, China.

E-mail: weizhenghong2006@yahoo.com.cn

Department of Mathematics, Hong Kong Baptist University, Hongkong

E-mail: lzhu@hkbu.edu.hk

(Received May 2008; accepted October 2008)