The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

2229-17

School and Workshop on Market Microstructure: Design, Efficiency and Statistical Regularities

21 - 25 March 2011

Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails

Cees DIKS

University of Amsterdam, Dept of Economics
Roetersstraat 11, NL-1018 WB
Amsterdam
THE NETHERLANDS

Strada Costiera 11, 34151 Trieste, Italy - Tel.+39 040 2240 111; Fax +39 040 224 163 - sci_info@ictp.it

Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails

Cees Diks[1] Valentyn Panchenko[2] Dick van Dijk[3]

[1]Universiteit van Amsterdam

[2]University of New South Wales

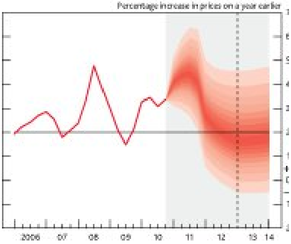
[3]Erasmus University, Rotterdam

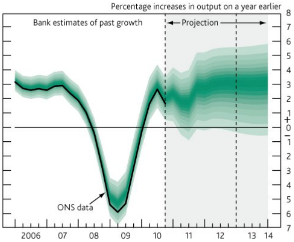
Trieste, March 24, 2011

Motivation

Macroeconomics: density forecasts of output and inflation from

- Statistical time series models (Clements & Smith, 2000)
- Professional forecasters (Diebold *et al.*, 1999)
- Central banks producing ‘fan charts’ (Mitchell & Hall, 2005)





Finance: Basis for risk management:

- Value-at-risk (VaR)
- Expedcted shortfall (ES)

Density forecast evaluation

One or more available density forecast(s) for sequence of random variables $\{Y_t\}$

Example: one-step-ahead predictive densities

- $\{Y_t\}$ is a scalar time series process
- Predictive pdf $\widehat{f}_t(y)$ of Y_{t+1}

How to evaluate such predictive densities if the true densities are never revealed?

Well-known measure of predictive ability: *mean squared prediction error*. However, suitable for point predictors only.

Approaches

1. *Probability integral transforms* (PITs)

   $$\widehat{U}_{t+1} := \widehat{F}_t(Y_{t+1})$$

   should be a sequence of independent UNIF(0,1) random variables for a correct specification. (Diebold *et al.*, 1998, 1999)

   Also for the multivariate case (Rosenblatt, 1952)

2. *Scoring rules:* assign a score to the predictive density for each realised value Y_{t+1}, high (low) if $\widehat{f}_t(Y_{t+1})$ is high (low).

   The average score is a measure for the quality of the predictive densities.

Tests for equal predictive ability

Giacomini & White (2006): score difference

$$d^*_{t+1} = S^*(\hat{f}_t; y_{t+1}) - S^*(\hat{g}_t; y_{t+1}),$$

Null hypothesis of equal scores, on average:

$$H_0: \quad \mathrm{E}(d^*_{t+1}) = 0, \qquad \text{for all } t = m, m+1, \ldots, T-1.$$

Mean score difference:

$$\overline{d}^*_{m,n} = \frac{1}{n} \sum_{t=m}^{T-1} d^*_{t+1} \quad \text{with} \quad n = T - m$$

Diebold-Mariano (1995) type test statistic:

$$t_{m,n} = \frac{\overline{d}^*_{m,n}}{\sqrt{\hat{\sigma}^2_{m,n}/n}} \xrightarrow{d} N(0,1)$$

Properness

Rational users would prefer p_t over any incorrect density forecast (Diebold *et al.*, 1998; Granger and Pesaran, 2000)

$\Rightarrow$ Natural to focus on *proper* scoring rules:

$$\mathrm{E}_t\left(S^*(\hat{f}_t; Y_{t+1})\right) \leq \mathrm{E}_t\left(S^*(p_t; Y_{t+1})\right), \qquad \text{for all } t.$$

Logarithmic scoring rule

Log-likelihood score:

$$S^{\ell}(\widehat{f}_t; y_{t+1}) = \log \widehat{f}_t(y_{t+1})$$

Based on a sequence of n density forecasts and realisations, $\widehat{f}$ and $\widehat{g}$ can be ranked according to average scores

$$\frac{1}{n}\sum_t \log \widehat{f}_t(y_{t+1}) \quad \text{and} \quad \frac{1}{n}\sum_t \log \widehat{g}_t(y_{t+1}).$$

Test of equal predictive ability

$$H_0 : E\left(d_t^{\ell}\right) = 0,$$

where

$$d_t^{\ell} = S^{\ell}(\widehat{f}_t; y_{t+1}) - S^{\ell}(\widehat{g}_t; y_{t+1}) = \log \widehat{f}_t(y_{t+1}) - \log \widehat{g}_t(y_{t+1}).$$

Kullback Leibler information criterion

$E\left(\bar{d}^{\ell}\right)$ can be interpreted as a difference of the distance of $\widehat{f}$ and $\widehat{g}$ to the true model.

Kullback-Leibler information criterion (KLIC)

$$\begin{aligned}
\mathrm{KLIC}(\widehat{f}_t) &= \int p_t(y_{t+1}) \log\left(\frac{p_t(y_{t+1})}{\widehat{f}(y_{t+1})}\right) \mathrm{d}y_{t+1} \\
&= E_t\left(\log p_t(Y_{t+1}) - \log \widehat{f}_t(Y_{t+1})\right) \\
&\geq 0
\end{aligned}$$

measures divergence between $\widehat{f}_t$ and the true conditional density p_t.

Weighted logarithmic scoring rules

Amisano and Giacomini (2007) suggest weighted logarithmic (WL) score

$$S^{wl}(\widehat{f}_t; y_{t+1}) = w(y_{t+1}) \log \widehat{f}_t(y_{t+1}).$$

and

$$d_t^{wl} = w(y_{t+1}) \left(\log \widehat{f}_t(y_{t+1}) - \log \widehat{g}_t(y_{t+1}) \right).$$

For financial applications (VaR, ES, ...) accuracy of the predictive density in the lower tail is of particular importance

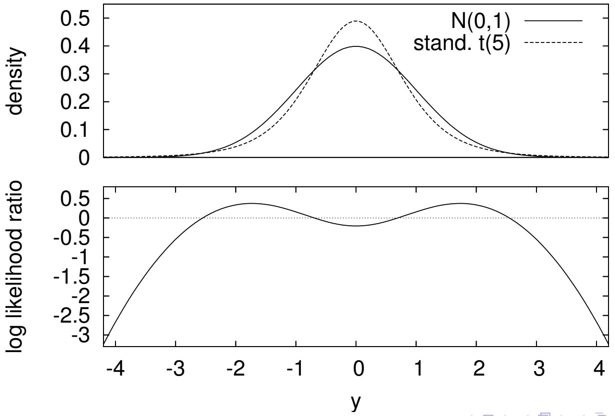
$\Rightarrow$ put most weight in left tail, e.g. choose

$$w_t(y) = \mathrm{I}(y \leq r_t)$$

Example: normal v.s. fat-tailed

Competing densities: $N(0,1)$ and standardised $t(5)$

$$f(y) = (2\pi)^{-\frac{1}{2}} \exp(-y^2/2), \qquad g(y) = 8(1+y^2/3)^{-3}/(3\pi\sqrt{3})$$



Weighted probability scores (Gneiting & Ranjan, 2008)

Continuous ranked probability score:

$$\mathrm{CRPS}(\hat{f}_t, y_{t+1}) = \int_{-\infty}^{\infty} \mathrm{PS}(\hat{F}_t(r), \mathrm{I}(y_{t+1} \le r))\, dr,$$

where

$$\mathrm{PS}(\hat{F}_t(r), \mathrm{I}(y_{t+1} \le r)) = (\mathrm{I}(y_{t+1} \le r) - \hat{F}_t(r))^2$$

(Brier probability score for the forecast)

Weighted version:

$$\mathrm{CRPS}(\hat{f}_t, y_{t+1}) = \int_{-\infty}^{\infty} w_t(r)\mathrm{PS}(\hat{F}_t(r), \mathrm{I}(y_{t+1} \le r))\, dr,$$

Conditional and censored likelihood

Idea: require scores to have an interpretation as a log-likelihood

Why? Likelihood-based scores are well-adapted to model comparison.

Expected score difference have an interpretation as a KLIC.

A correct forecast will receive higher average score than any competing model (properness)

Scoring rules based on conditional and censored likelihood

Region of interest: A_t

Conditional log-likelihood:

$$S^{cl}(\hat{f}_t; y_{t+1}) = \mathrm{I}(y_{t+1} \in A_t) \log \left(\frac{\hat{f}_t(y_{t+1})}{\int_{A_t} \hat{f}_t(s)\mathrm{d}s} \right)$$

Censored log-likelihood:

$$\begin{aligned} S^{csl}(\hat{f}_t; y_{t+1}) \quad &= \quad \mathrm{I}(y_{t+1} \in A_t) \log \hat{f}_t(y_{t+1}) \\ &\quad + \mathrm{I}(y_{t+1} \in A_t^c) \log \left(\int_{A_t^c} \hat{f}_t(s)\mathrm{d}s \right) \end{aligned}$$

Smooth scoring rules

Conditional log-likelihood

$$S^{cl}(\hat{f}_t; y_{t+1}) = w_t(y_{t+1}) \log \left(\frac{\hat{f}_t(y_{t+1})}{\int w_t(s)\hat{f}_t(s)\mathrm{d}s} \right)$$

Censored log-likelihood:

$$\begin{aligned} S^{csl}(\hat{f}_t; y_{t+1}) \quad = \quad & w_t(y_{t+1}) \log \hat{f}_t(y_{t+1}) \\ & +(1 - w_t(y_{t+1})) \log \left(1 - \int w_t(s)\hat{f}_t(s)\mathrm{d}s\right). \end{aligned}$$

Properness of the new scoring rules

Assumption 1: The density forecasts $\hat{f}_t$ and $\hat{g}_t$ satisfy $\text{KLIC}(\hat{f}_t) < \infty$ and $\text{KLIC}(\hat{g}_t) < \infty$, where $\text{KLIC}(h_t) = \int p_t(y) \log\left(p_t(y)/h_t(y)\right) \mathrm{d}y$ is the Kullback-Leibler divergence between the density forecast h_t and the true conditional density p_t.

Assumption 2: The weight function $w_t(y)$ is such that (a) it is determined by the information available at time t, and hence a function of $\mathcal{F}_t$, (b) $0 \leq w_t(y) \leq 1$, and (c) $\int w_t(y) p_t(y) \, \mathrm{d}y > 0$.

Lemma 1: Under Assumptions 1 and 2, the generalized conditional likelihood scoring rule and the generalized censored likelihood scoring rule are proper.

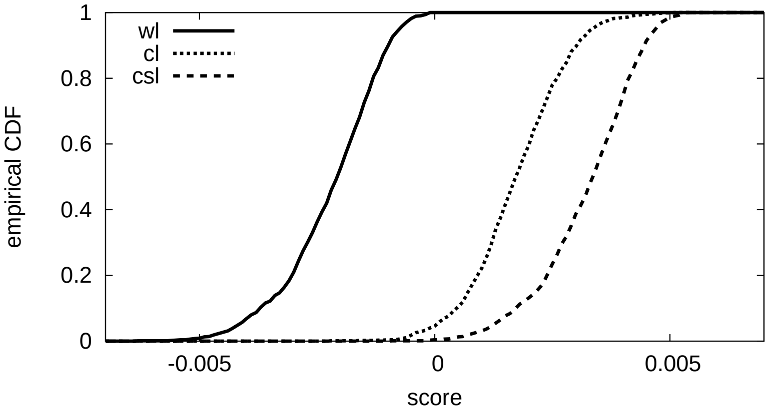
Proof of Lemma 1

Define $P_t \equiv \int w_t(s)p_t(s)\,\mathrm{d}s$ and $\hat{F}_t \equiv \int w_t(s)\hat{f}_t(s)\,\mathrm{d}s$

$$
\begin{aligned}
\mathrm{E}_t\left(d^{cl}_{t+1}(p_t,\hat{f}_t)\right) &= \int p_t(y)\left(w_t(y)\log\left(\frac{p_t(y)}{P_t}\right)\right) dy \\
&\qquad - \int p_t(y)\left(w_t(y)\log\left(\frac{\hat{f}_t(y)}{\hat{F}_t}\right)\right) dy \\
&= P_t \int \frac{w_t(y)p_t(y)}{P_t}\log\left(\frac{w_t(y)p_t(y)/P_t}{w_t(y)\hat{f}_t(y)/\hat{F}_t}\right)\mathrm{d}y \\
&= P_t \cdot K\left(\frac{w_t(y)p_t(y)}{P_t}, \frac{w_t(y)\hat{f}_t(y)}{\hat{F}_t}\right) \geq 0,
\end{aligned}
$$

Example: normal v.s. fat-tailed (continued)

Simulated score differences for $N(0,1)$ and standardised $t(5)$



Empirical CDFs of scores under the threshold weight function $w(y) = \mathrm{I}(y \leq -2.5)$. $n = 1000$, 1000 replications, $Y_t \sim N(0,1)$

Simulations for smooth weight functions

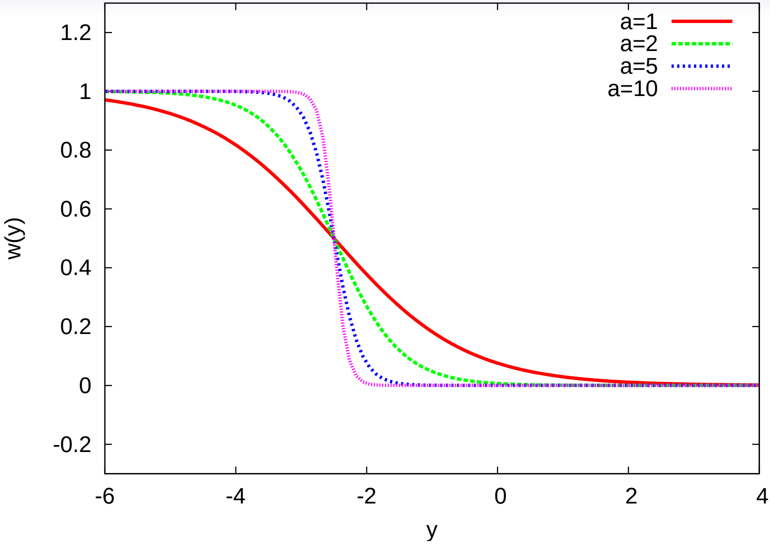
Weight functions of the form

$$w(y) = 1/(1 + \exp(a(y - r))).$$

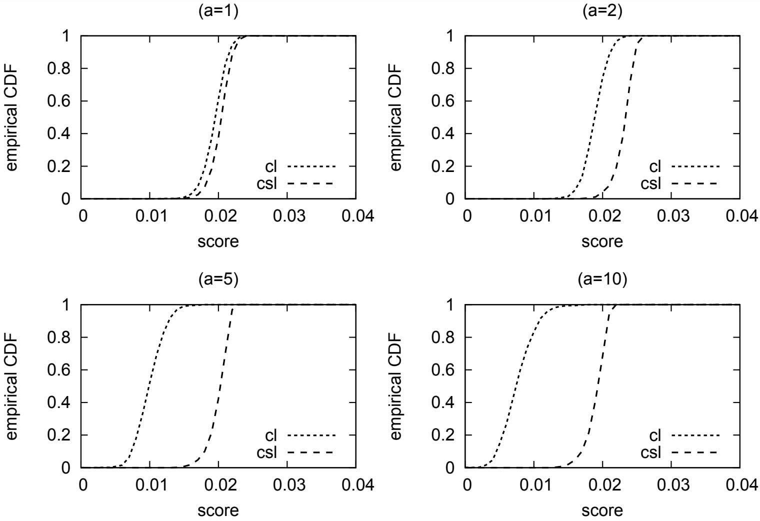
Sigmoidal function of y with center r and slope parameter a.

Here r is fixed at $r = -2.5$. The slope parameter a varies. For $a \to \infty$ the threshold weight function is recovered.

Integrals $\hat{F}_t = \int \hat{f}_t(x) w_t(x) \mathrm{d}x$ and $\hat{G}_t = \int \hat{g}_t(x) w_t(x) \mathrm{d}x$ determined numerically



Weight functions for increasing smoothing parameter



Score distributions under the two smooth weighting schemes

Monte Carlo simulations for size/power

Properties of test statistics for each score: HAC-estimators of the standard error of the sample mean score

E.g. for the type l scoring rule, the test statistic is $\widehat{Q}_n^l = \sqrt{n}\bar{d}^l/\widehat{\sigma}_n^l$, where

$$\widehat{\sigma}_n^{2,l} = \widehat{\gamma}_0 + 2\sum_{k=1}^{K-1} a_k \widehat{\gamma}_k$$

where $\widehat{\gamma}_k$ denotes the lag-k sample covariance of the sequence $\{d_t^l\}$. The weights are taken as $a_k = 1 - k/K$ with $K = \lfloor n^{-1/4} \rfloor$.

Under the null hypothesis of equal predictive ability each of the test statistics is asymptotically standard normally distributed

Size, simulation setup

Data generating process: $Y_t \sim N(0,1)$ IID

Competing forecasts:

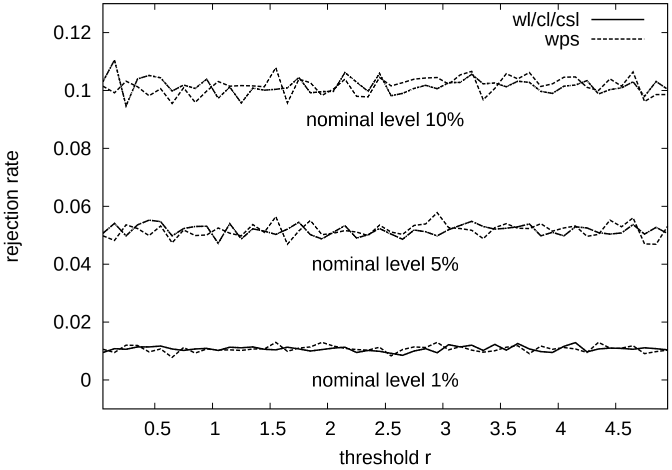
$$N(0,-0,2) \qquad \text{versus} \qquad N(0,0.2)$$

Weight function:

$$w_t(y) = \mathrm{I}(-r \leq y \leq r)$$

Size, results

DGP: IID $N(0, 1)$, forecasts IID $N(-0.2, 1)$, $N(0.2, 1)$
Weight function $w_t(y) = \mathrm{I}(-r \leq y \leq r)$



Power: Simulation Setup

Building on motivating example:

DGP: IID $N(0,1)$ or IID standardised $t(5)$.

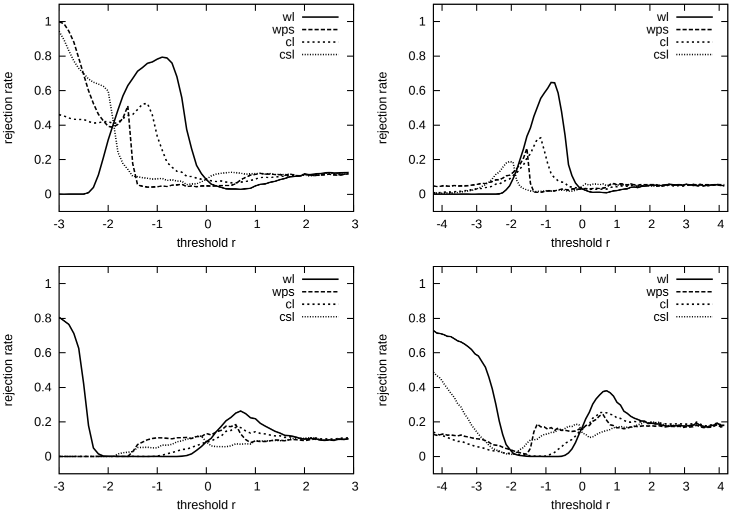
Test equal predictive ability, versus two alternatives:

- $N(0,1)$ forecast outperforms $t(5)$
- $t(5)$ forecast outperforms $t(5)$

Weight function: $w_t(y) = \mathrm{I}(y \leq r)$

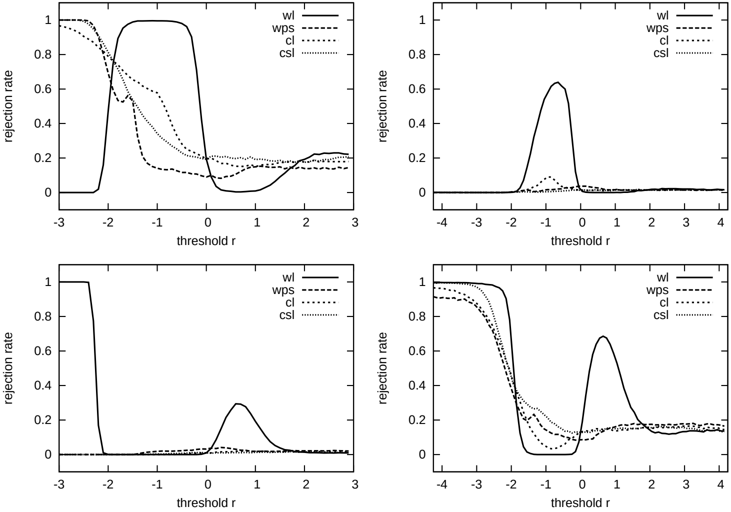
To control for loss of power in the tails, the expected number of observations in the left tail, c, is fixed.

Power, results for $c = 5$

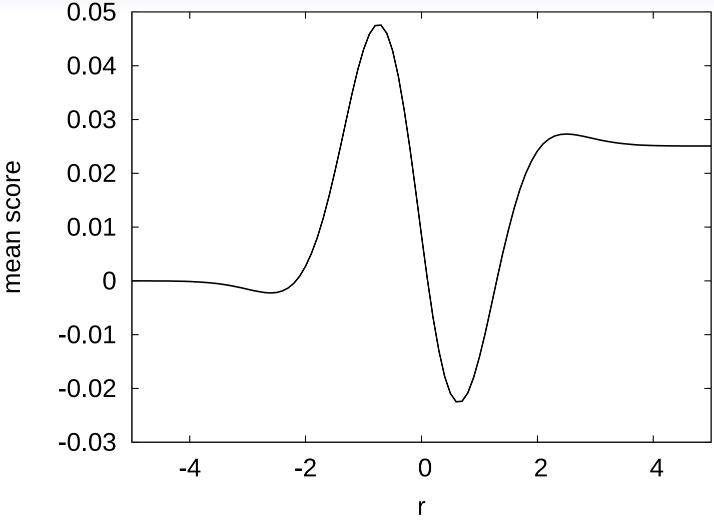


Left: DGP IID $N(0.1)$, Right: DGP IID std. $t(5)$. Top: test of std. $t(5)$ against $N(0, 1)$, bottom: test of $N(0, 1)$ against std. $t(5)$.

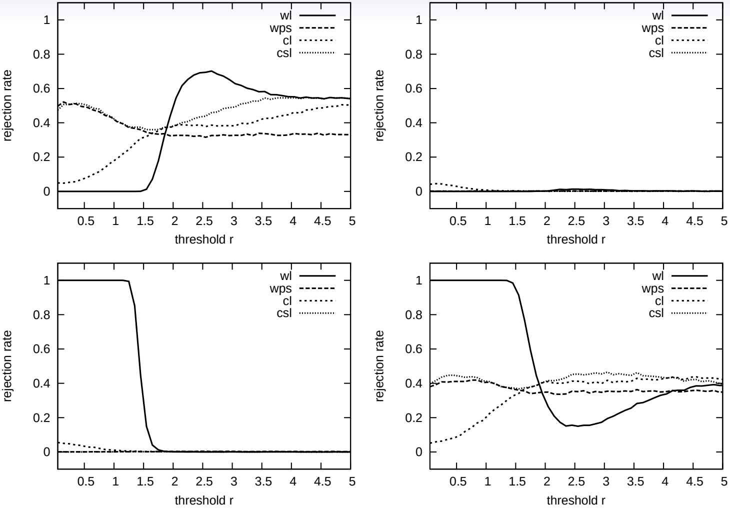
Power, results for $c = 40$



Left: DGP $N(0.1)$, right: DGP std. $t(5)$. Top: std. $t(5)$ against $N(0,1)$, bottom: $N(0,1)$ against std. $t(5)$. $c = 40$.



Mean WL score $E\left(d_t^{wl}\right)$ as a function of the threshold value r, for the standard normal DGP.



Symmetric case, $c = 200$. Left: DGP: i.i.d. $N(0,1)$. Competing densities $N(0,1)$ and std. $t(5)$, weight function $\mathrm{I}(-r \le y \le r)$.

Parameter estimation uncertainty

	m	100	250	500	1000	2500	5000
H_a :	$E(d^l_{t+1}) > 0$	0.000	0.000	0.024	0.134	0.339	0.463
H_a :	$E(d^l_{t+1}) < 0$	0.982	0.239	0.026	0.004	0.001	0.000

One-sided rejection rates, $w_t(y) = 1$

DGP: AR(2): $Y_t = 0.8Y_{t-1} + 0.05Y_{t-2} + \varepsilon_t$

Score differences: log-scores for $AR(2)$, minus log-scores for $AR(1)$

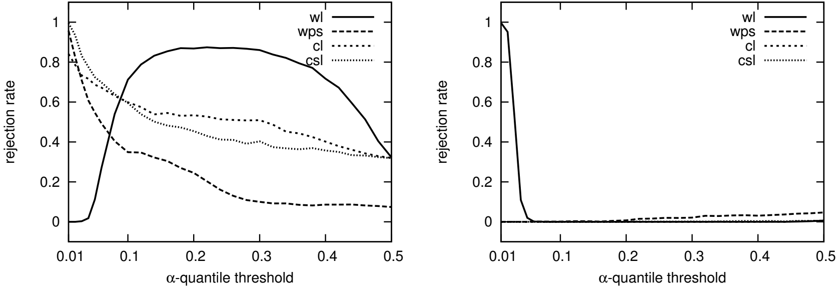
(Correct specification versus a more parsimonious incorrectly specified model)

Time-varying weights

GARCH(1,1)-model, $c = 40$, weight function

$$w_t(y) = \mathrm{I}(y \leq r_t) \quad \text{with} \quad r_t = \hat{y}_\alpha^t$$

(empirical α-quantile)



Correct $N(0,1)$ innovations versus std. $t(5)$ innovations
Left: power, right: spurious power
DGP: GARCH(1,1)

Comparing two models for log-returns

Data: daily S&P 500 daily log-returns $X_t = \log(P_t/P_{t-1})$, period Jan. 1, 1980 – March 14, 2008 (7115 observations)

Comparison of two models, one of which is restricted

$$\begin{aligned} X_t &= \mu_t + h_t\varepsilon_t, \qquad \varepsilon_t \sim t(\nu), \\ \mu_t &= \rho_0 + \textstyle\sum_{\ell=1}^{5} \rho_\ell X_{t-\ell}, \\ h_t &= c + \alpha(X_{t-1} - \mu_{t-1})^2 + \beta h_{t-1}. \end{aligned}$$

Excess kurtosis $6/(\nu - 4)$

Alternative innovation distribution: Laplace (excess kurt. = 3)

Aim: compare one-step-ahead predictive densities
Estimation window $m = 2000$

Average score differences

Testing Laplace versus $t(\nu)$ innovations

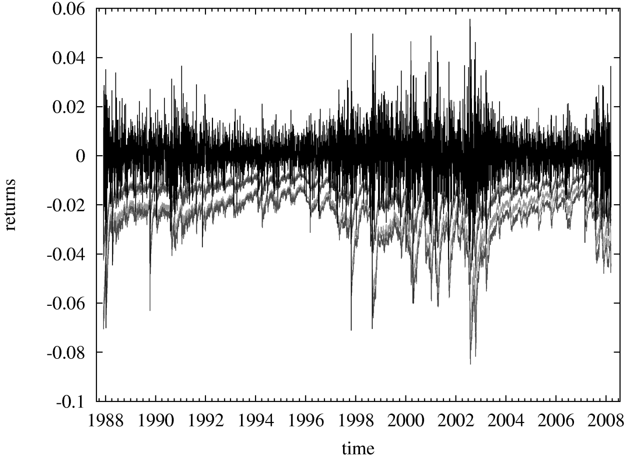
$$d_t^* = t\text{-score} - \text{Laplace score}$$

$$w_t(y) = \mathrm{I}(y \leq r_t) \quad \text{with} \quad r_t = \hat{y}_t^{\alpha}$$

Scoring rule	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	$\bar{d}^*$	Test stat.	$\bar{d}^*$	Test stat.	$\bar{d}^*$	Test stat.
			Threshold weight function			
wl	−0.000169	−0.14	−0.00512	−4.74	−0.0032	−3.75
wps	0.000000429	0.69	0.000000775	1.56	0.000000868	4.28
cl	0.00147	1.48	0.00158	2.32	0.000778	1.81
csl	0.00221	1.89	0.00163	1.53	0.00116	1.35

Average score differences and tests of equal predictive accuracy

Daily S&P 500 log-returns (black) and out-of-sample 95% and 99% VaR forecasts



Forcasts from AR(5)-GARCH(1,1) specification with Student-*t* innovations (light gray) and Laplace innovations (dark gray)

VaR and ES characteristics

	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	$t(\nu)$	Laplace	$t(\nu)$	Laplace	$t(\nu)$	Laplace
Average VaR	-0.0110	-0.0112	-0.0149	-0.0162	-0.0243	-0.0279
Coverage ($y_t \leq \mathrm{VaR}_t$)	0.1056	0.1001	0.0530	0.0405	0.0104	0.0055
CUC (p-value)	0.1876	0.9814	0.3324	0.0012	0.7961	0.0004
IND (p-value)	0.1082	0.2315	0.0465	0.3658	0.5809	0.5788
CCC (p-value)	0.1156	0.4887	0.0861	0.0036	0.8304	0.0015
Average ES	-0.0168	-0.0185	-0.0209	-0.0235	-0.0312	-0.0351
McNeil-Frey (test stat.)	-0.7538	3.1164	-0.8504	0.3639	-1.1899	-2.3174
McNeil-Frey (p-value)	0.4510	0.0018	0.3951	0.7159	0.2341	0.0205

Coverage: observed fraction of returns below VaR

Summary

- Existing weighing schemes for scoring rules have demonstrable shortcomings
- Proposed new scoring rules based on partial likelihood
- Properness of the new scoring rules could be proved
- Numerical study showed correct behaviour for new scoring rules
- Illustrated with an empirical application