On Backtesting Risk Measurement Models

Hideatsu Tsukahara
Department of Economics, Seijo University
e-mail address: tsukahar@seijo.ac.jp

1 Introduction

In general, the purpose of backtesting is twofold: to monitor the performance of the model and estimation methods for risk measurement, and to compare relative performance of the models and methods. It is a tool for the validation process which is indispensable for adequate financial risk management. According to Consultative Document of Basel Committee on Banking Supervision (October 2013),

> "Move from Value-at-Risk (VaR) to Expected Shortfall (ES): A number of weaknesses have been identified with using VaR for determining regulatory capital requirements, including its inability to capture "tail risk". For this reason, the Committee proposed in May 2012 to *replace VaR with ES*. ES measures the riskiness of a position by considering both the size and the likelihood of losses above a certain confidence level. The Committee has agreed to use a 97.5% ES for the internal models-based approach and has also used that approach to calibrate capital requirements under the revised market risk standardised approach"

However, in the same document, the Committee requires backtesting VaR, *not ES*, in *Revised Models-based Approach*

> "In addition to P&L attribution, the performance of a trading desk's risk management models will be evaluated through daily backtesting. **Backtesting requirements would be based on comparing each desk's 1-day static value-at-risk measure at both the 97.5th percentile and the 99th percentile to actual P&L outcomes**, using at least one year of current observations of the desk's one-day actual and theoretical P&L. The backtesting assessment would be run at each trading desk as well as for the global (bank-wide) level."

This seemingly self-contradictory requirement is perhaps based on the following fact: many people believe that it is easier to backtest VaR than Expected Shortfall (ES) and other risk measures because (i) the existing tests for ES are based on parametric assumptions for the null distribution; (ii) asymptotic approximation is needed for the null distribution of the test statistics. In the case of Value-at-Risk (VaR), a popular procedure for backtesting depends on the number of VaR violations (Campbell [1]), which is *distribution-free* with finite samples and intuitively appealing.

Thus the 'backtestability 'of a risk measure apparently means that it can be backtested in a distribution-free manner with finite samples. Recently, the concept called "elicitability" has recently attracted much attention in order to claim the superiority of VaR in terms of backtesting. Roughly speaking, a statistical functional is called *elicitable* if it is a unique

minimizer of some expected loss (Gneiting [7]). While VaR is easily seen to be elicitable, it has been proved that ES and the distortion risk measures fails to satisfy this condition (Wang and Ziegel [11], Ziegel [13]). We claim that while elicitability is certainly useful for comparing and ranking models/procedures, there seems to be no clear connection with monitoring the performance of a model/procedure in use. We will illustrate this with simple examples, and also examine the problem from the decision-theoretic perspective including the prequential principle by Dawid [4, 5, 6], recently argued in Davis [3].

Finally, some backtesting procedures for distortion risk measures are suggested, and we check its effectiveness in a simulation study using ES and also proportional odds distortion risk measure.

2 Risk Measurement Models

We follow McNeil et al. [10] for the purpose of risk measurement; based on historical observations and given a specific model, a statistical estimate of the distribution of the future loss of a position, or one of its functionals, is to be calculated. Let $L_1, L_2, \ldots, L_n, \ldots$ be loss variables with values in $\mathbb{R}$, and let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n, \ldots$ be $\mathbb{R}^d$-valued covariates. Let $\mathscr{F}_n := \sigma(L_k, \boldsymbol{X}_k, 1 \leq k \leq n)$ be a filtration generated by the loss variables and covariates. Then the risk measurement model specifies a family of probability distributions on $(\mathbb{R}^{d+1})^\infty$ for the sequence $(L_n, \boldsymbol{X}_n, n \in \mathbb{N})$

As nicely explained in McNeil et al. [10], there are two approaches to risk measurement. Suppose that the loss process $(L_n)_{n\in\mathbb{N}}$ is a stationary time series with a stationary distribution function F. At time $n-1$, we have two options:

Unconditional Approach: Look at the unconditional distribution function $F(x) = \mathrm{P}(L_n \leq l)$ and its functionals. This is considered to be suitable for the credit risk and insurance with a large time horizon.

Conditional Approach: Look at the conditional distribution function $F_n(l) := \mathrm{P}(L_n \leq l \,|\, \mathscr{F}_{n-1})$ and its functinals. This is considered to be suitable for the market risk with a short time horizon./

There are pros and cons with these two approaches from the theoretical, not conceptual point of view:

(i) Under the assumption of stationarity, nonparametric estimation of many functionals is possible (indeed, we can use the empirical process theory)

(ii) We need to fit very specific (parametric) models for the computation of the conditional distribution and its functionals.

(iii) For the backtesting purpose, the conditional approach is preferred since the distribution theory is theoretically clear and correct (as will be seen below).

(iv) From the Bayesian perspective, the conditional approach would definitely be preferable.

In what follows, we take the conditional approach.

Next we discuss backtesting procedures in a general setup. Let $L_1, \dots, L_N$ be the entire observations, and set the estimation window size equal to m. Let ρ be a risk measure of one's choice. Statistically, a backtesing procedure is just a form of cross validation; the ex ante risk measure forecasts from the model is compared with the ex post realized portfolio loss. Namely, at step k, use the sample $L_k, \dots, L_{k+m-1}$ to estimate $\rho(L_{k+m})$, and using the realized loss L_{k+m} to measure the goodness-of-fit of your risk measurement model in terms of the estimation of the risk measure.

data	estimand(risk measure)	realized loss
$L_1, \dots, L_m$	$\rho(L_{m+1})$	L_{m+1}
$L_2, \dots, L_{m+1}$	$\rho(L_{m+2})$	L_{m+2}
$\vdots$	$\vdots$	$\vdots$
$L_{N-m}, \dots, L_{N-1}$	$\rho(L_N)$	L_N

In the VaR case, the conditional VaR, denoted by VaR_α^n, satisfies

$$\mathrm{E}\big(\mathbf{1}\{L_n > \mathrm{VaR}_\alpha^n\} \,\big|\, \mathscr{F}_{n-1}\big) = \alpha$$

By Lemma 4.29 of McNeil et al. [10], if (Y_n) is a sequence of Bernoulli random variables adapted to $(\mathscr{F}_n)$ and if $\mathrm{E}(Y_{n+1} \,|\, \mathscr{F}_n) = p > 0$, then (Y_n) must be i.i.d. Therefore, $\mathbf{1}\{L_n > \mathrm{VaR}_\alpha^n\}$, $n = m+1, \dots, N$ are i.i.d. Bernoulli random variables, and this gives the grounds for backtesting using $\mathbf{1}\{L_n > \widehat{\mathrm{VaR}}_\alpha^n\}$, where $\widehat{\mathrm{VaR}}_\alpha^n$ is an estimate of the VaR associated with the conditional distribution function $F_n(l) = \mathrm{P}(L_n \le l \,|\, \mathscr{F}_{n-1})$. Namely,

(i) Test $\displaystyle\sum_{n=m+1}^{N} \mathbf{1}\{L_n > \widehat{\mathrm{VaR}}_\alpha^n\} \sim \mathrm{Bin}(N-m, \alpha)$

(ii) Test independence of $\mathbf{1}\{L_n > \widehat{\mathrm{VaR}}_\alpha^n\}$, $n = m+1, \dots, N$ (e.g., runs test)

This is intuitively appealing, and also distribution-free with finite samples in the sense that the null distributions of the test statistics do not depend on the underlying loss distribution F.

On the other hand, many researchers claim that it is more difficult to backtest a procedure for calculating ES than it is to backtest a procedure for calculating VaR (Yamai and Yoshiba [12], Hull [8], Daníelsson [2], among others). It is because the existing tests for ES are based on parametric assumptions for the null distribution and some asymptotic approximation for the null distribution.

Recently the concept called elicitability is called for to support the claim that the expected shortfall (and distortion risk measures) cannot be backtested. A statistical functional $T(F)$ is called *elicitable* relative to $\mathscr{F}$ if $T(F)$ is a unique minimizer of $t \mapsto E^F[S(t, Y)]$ for some scoring function S, $\forall F \in \mathscr{F}$. Examples include $\mathrm{VaR}_\theta(F) = F^{-1}(1-\theta)$ and the mean functional $T(F) = \int y \, \mathrm{d}F(y)$. It is useful when one wants to compare and rank several estimation procedures: With forecasts x_i and realizations y_i, use

$$\frac{1}{N} \sum_{i=1}^{N} S(x_i, y_i)$$

as a performance evaluation criterion. But there seems to be no clear connection with monitoring purpose. For example, the mean cannot be backtested nonparametrically based

on the sum of squared errors without invoking asymptotic approximation or assuming parametric distribution. The expectile has emerged as the only law invariant and coherent risk measure that is elicitable, but it lacks intuitive interpretation as a risk measure.

2.1 Consistency of Mark Davis

We use the following terminologies:

Model: $\mathscr{P} = \{P^z : z \in \mathscr{Z}\}$ is a family of probability distributions on $(\mathbb{R}^{d+1})^\infty$ for the sequence $(L_n, \boldsymbol{X}_n)_{n\in\mathbb{N}}$;

Calibration function $\phi : \mathbb{R}^2 \to \mathbb{R}$

Normalizing sequence $b = (b_n)$: strictly increasing predictable sequence such that

$$\lim_{n\to\infty} b_n = \infty, \quad P^z\text{-a.s., } \forall z \in \mathscr{Z}$$

Regular conditional distribution functions

$$F_1^z(l) := P^z(L_1 \le l), \qquad F_n^z(l) := P^z(L_n \le l \,|\, \mathscr{F}_{n-1}),\ n = 2, 3, \dots$$

According to Davis [3], a statistical functional T is called $(\phi, b, \mathscr{P})$-*consistent* if

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} \phi(L_k, T(F_k^z)) = 0, \quad P^z\text{-a.s., } \forall z \in \mathscr{Z}$$

For instance, VaR_α^k satisfies the above condition with

$$\phi(l, t) = \mathbf{1}_{[t,\infty)}(l) - \alpha, \quad b_n = n$$

because of the SLLN.

In practice, given observations $(L_1, \boldsymbol{X}_1), \dots, (L_{k-1}, \boldsymbol{X}_{k-1})$, we produce a forecast τ_k for $T(F_k^z)$ based on some algorithm. And one can evaluate the performance of this forecast by calculating

$$J_n(L_1, \dots, L_n, \tau_1, \dots, \tau_n) = \frac{1}{b_n} \sum_{k=1}^{n} \phi(L_k, \tau_k).$$

If $J_n(\tau)$ is sufficiently close to 0 (for large n), then our forecast algorithm may be considered to be fine. This criterion satisfies Dawid's *prequential principle*: "Any validity criterion should be calculated knowing only the realized losses and the actual forecasts issued".

Davis shows that the consistency of VaR holds under very general conditions, while for mean-type functionals, significant conditions must be imposed to ensure their consistency. Then he concludes that verifying the validity of forecasts for mean-type functionals is essentially more problematic than that for quantile-type functionals. However, since the consistency is an asymptotic requirement, it does not give us a totally satisfactory answer.

2.2 Backtesting Predictive Distributions

Recall the following result on so-called Rosenblatt transform:

Theorem 2.1 *Define*

$$U_n := F_n^z(L_n), \quad n = 1, 2, \dots$$

Then, assuming the a.s.-continuity of all F_n^z, $U_1, U_2, \dots$, are i.i.d. $U(0,1)$ random variables under P^z.

Suppose we have a sequence of forecasts $\widehat{F}_k$ of F_k^z, $k = 1, 2, \dots, n$. One can verify the validity of $(\widehat{F}_k)$ by testing

$$H_0 \colon \widehat{F}_1(L_1), \dots, \widehat{F}_n(L_n) \overset{\text{i.i.d.}}{\sim} U(0,1)$$

We can extend Davis's Consistency to maps with range being a general metric space. Let $\mathscr{G}$ be the space of distribution functions on $\mathbb{R}$ with the topology of weak convergence, and let $\mathscr{S}$ be a Polish space. Consider a map $T \colon \mathscr{G} \to \mathscr{S}$, extending a statistical functional. A calibration function ϕ is now a map from $(\mathbb{R}, \mathscr{S}) \to \mathbb{R}$

Definition 2.2 A map $T \colon \mathscr{G} \to \mathscr{S}$ is called $(\phi, b, \mathscr{P})$-*consistent* if

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} \phi(L_k, T(F_k^z)) = 0, \quad P^z\text{-a.s., } \forall z \in \mathscr{Z}$$

For the predictive distributions, take $b_n = n$, $\mathscr{S} = \mathscr{G}$, $T(F) = F$, and

$$\phi_q(l, F) := \mathbf{1}\{F(l) \le q\} - q, \quad q \in \mathbb{Q} \cap [0,1]$$

Then by SLLN, we have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} \phi_q(L_k, T(F_k^z)) &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \Big[\mathbf{1}\{F_k^z(L_k) \le q\} - q \Big] \\ &= 0, \ P^z\text{-a.s., } \forall z \in \mathscr{Z} \end{aligned}$$

for all $q \in \mathbb{Q} \cap [0,1]$ with the same conditions as the VaR case. This shows that the VaR and the predictive distribution are consistent under the same condition, while the mean is consistent under much more stringent conditions. This is intuitively an odd conclusion.

3 Backtesting Distortion Risk Measures

We can use the same approach as in McNeil and Frey [9] to devise a backtesting procedure for the distortion risk measure. Write $\rho_n(L_n)$ for a distortion risk measure with a distortion D for the conditional distribution function $F_n(l) := \mathrm{P}(L_n \le l \,|\, \mathscr{F}_{n-1})$, $\mathscr{F}_n := \sigma(L_k, \boldsymbol{X}_k \colon 1 \le k \le n)$:

$$\rho_n(L_n) := \int_{[0,1]} F_n^{-1}(u) \, \mathrm{d}D(u)$$

Suppose that for $\mathscr{F}_{n-1}$-measurable μ_n and σ_n,

$$L_n = \mu_n + \sigma_n Z_n,$$

where (Z_n) is i.i.d. with finite 2nd moment. For example, the following popular ARMA(p_1, q_1) model with GARCH(p_2, q_2) errors satisfies the above assumptions:

$$\begin{aligned}
L_n &= \mu_n + \sigma_n Z_n, \\
\mu_n &= \mu + \sum_{i=1}^{p_1} \phi_i (L_{n-i} - \mu) + \sum_{j=1}^{q_1} \theta_j (L_{n-j} - \mu_{n-j}), \\
\sigma_n^2 &= \alpha_0 + \sum_{i=1}^{p_2} \alpha_i (L_{n-i} - \mu_{n-i})^2 + \sum_{j=1}^{q_2} \beta_j \sigma_{n-j}^2,
\end{aligned}$$

where Z_n's are i.i.d. with finite second moment, $\alpha_0 > 0$, $\alpha_i \geq 0$, $i = 1, \ldots, p_2$, $\beta_j \geq 0$, $j = 1, \ldots, q_2$. Usually, it is assumed that (L_n) is covariance stationary, and $\sum_{i=1}^{p_2} \alpha_i + \sum_{j=1}^{q_2} \beta_j < 1$.

By the (conditional version of) translation invariance and positive homogeneity, we have

$$\rho_n(L_n) = \mu_n + \sigma_n \rho(Z)$$

where Z is a generic random variable with the same distribution function G as Z_n's.

(i) If G is a known df, $\rho(Z)$ is a known number. We need to estimate μ_k and σ_k based on $L_{k-n}, \ldots, X_{k-1}$ using some specific model and method (e.g., ARMA with GARCH errors using QML). Then the risk measure estimate is given by

$$\widehat{\rho}_n(L_n) := \widehat{\mu}_n + \widehat{\sigma}_n \rho(Z)$$

Observe that $\rho(Z) = \mathrm{E}\left[Z_n d(G(Z_n))\right]$ implies $\mathrm{E}\left[(Z_n - \rho(Z)) d(G(Z_n))\right] = 0$. Defining

$$R_n := Z_n - \rho(Z) = \frac{L_n - \rho_n(L_n)}{\sigma_n}$$

one sees that $(R_n d(G(Z_n)))_{t \in \mathbb{Z}}$ is i.i.d. This suggests that in practice, we may perform backtesting by examining mean-zero behavior of by examining mean-zero behavior of $\widehat{R}_k d(G(\widehat{Z}_k))$, $k = n+1, \ldots, N$, where

$$\widehat{R}_k := \frac{L_k - \widehat{\rho}_k(L_k)}{\widehat{\sigma}_k}$$

and

$$\widehat{Z}_k = \frac{L_k - \widehat{\mu}_k}{\widehat{\sigma}_k} = \widehat{R}_k + \rho(Z)$$

The bootstrap test can be employed for a formal test.

(ii) When G is unknown, we need to estimate G in addition to μ_k and σ_k. In ARMA with GARCH errors model, we could use the empirical distribution function of the residuals $\widetilde{Z}_m$'s: for $m = k-n, \ldots, k-1$, let

$$\widetilde{Z}_m = \widetilde{\varepsilon}_m / \widetilde{\sigma}_m,$$

where $\widetilde{\varepsilon}_m$ is the residuals from ARMA part, and

$$\widetilde{\sigma}_m^2 = \widehat{\alpha}_0 + \sum_{i=1}^{p_2} \widehat{\alpha}_i \widetilde{\varepsilon}_{m-i}^2 + \sum_{j=1}^{q_2} \widehat{\beta}_j \widehat{\sigma}_{m-j}^2.$$

Then one can estimate G by

$$\widetilde{G}_k(z) = \frac{1}{n} \sum_{m=k-n}^{k-1} \mathbf{1}\{\widetilde{Z}_m \leq z\}.$$

Simulation study

We simulate the following GARCH(1,1) process.

$$L_k = \sigma_k Z_k, \quad Z_k \sim N(0,1) \text{ i.i.d.}$$
$$\sigma_k^2 = 0.01 + 0.9\sigma_{k-1}^2 + 0.08L_{k-1}^2$$

where the Z_t are i.i.d. standard normal random variables. We set $T = 1000$, $n = 500$ and $\theta = 0.05$, and for $t = n+1, \ldots, T$, plot

(i) $X_t d(\widehat{\mathbb{F}}_{t-n:t-1}(X_t)) - \widehat{\rho}_{(t-n:t-1)}$ (historical, unconditional)

(ii) $\widehat{R}_t d(G(\widehat{Z}_t))$ (normal-GARCH based, conditional)

The results are displayed in Figures 1 and 2. (i) mean $= -0.0286$, std $= 2.073$ and (ii) mean $= -0.0185$, std $= 1.019$

References

[1] S. D. Campbell. A review of backtesting and backtesting procedures. *Journal of Risk*, 9:1–17, 2007.

[2] J. Daníelsson. *Financial Risk Forecasting*. John Wiley & Sons, 2011.

[3] M. H. A. Davis. Verification of internal risk measure estimates. *Statistics & Risk Modeling*, 33:67–93, 2016.

[4] A. P. Dawid. Present position and potential developments: some personal views. statistical theory. the prequential approach (with discussion). *Journal of the Royal Statistical Society (A)*, 147:278–292, 1984.

[5] A. P. Dawid. Prequential data analysis. In M. Ghosh and P. K. Pathak, editors, *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, IMS Lecture Notes–Monograph Series, vol. 17, pages 113–126, 1992.

[6] A. P. Dawid. Prequential analysis. In C. B. Read S. Kotz and D. L. Banks, editors, *Encyclopedia of Statistical Sciences, Update Volume 1*, pages 464–470. Wiley-Interscience, 1997.

[7] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011.

[8] J. C. Hull. *Risk Management and Financial Institutions*. John Wiley & Sons, New York, fourth edition, 2015.

[9] A. J. McNeil and R. Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7:271–300, 2000.

[10] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press, Princeton, New Jersey, second edition, 2015.

[11] R. Wang and J. F. Ziegel. Elicitable distortion risk measures: A concise proof. *Statistics & Probability Letters*, 100:172–175, 2015.

[12] Y. Yamai and T. Yoshiba. Value-at-risk versus expected shortfall: A practical perspective. *Journal of Banking & Finance*, 29:997–1015, 2005.

[13] J. F. Ziegel. Coherence and elicitability. *Mathematical Finance*, 26:901–918, 2016.

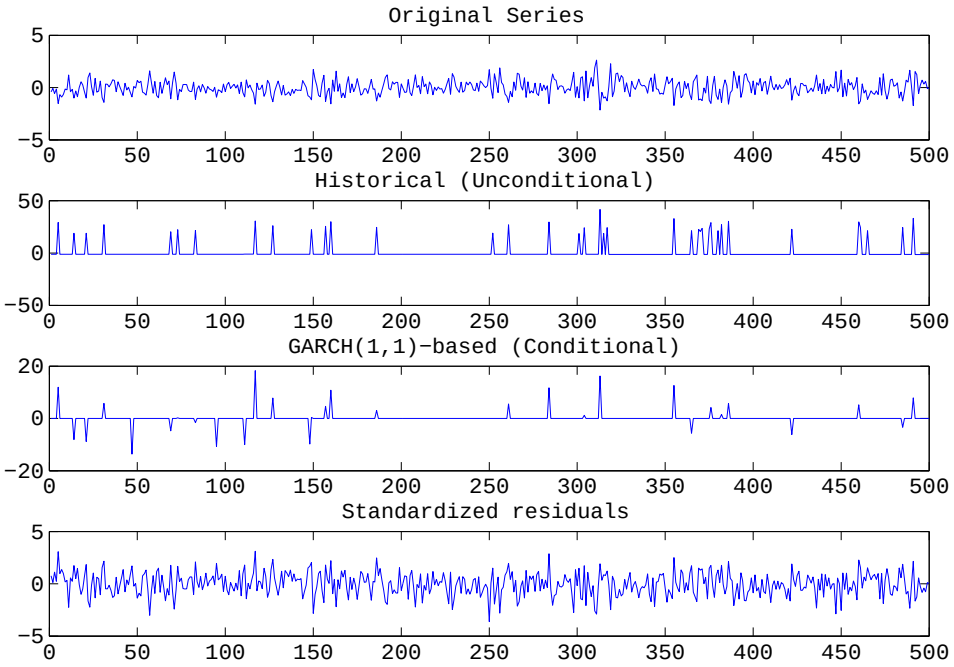

Figure 1: Backtesting results for expected shortfall ($\theta = 0.05$)

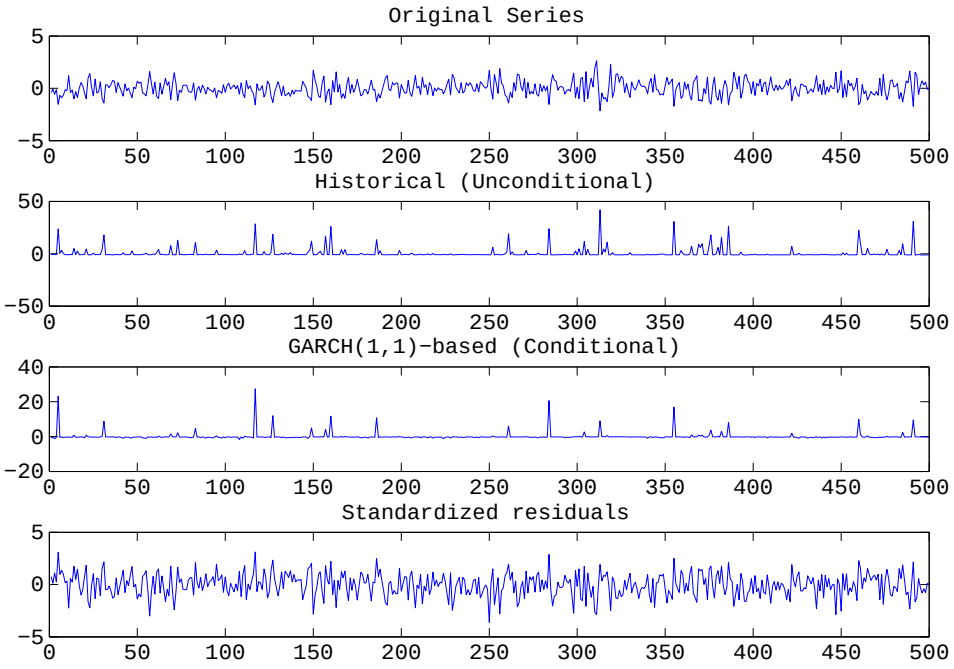

Figure 2: Backtesting results for proportional odds distortion ($\theta = 0.05$)