The Predictive Power of Value-at-Risk Models in Commodity Futures Markets

Roland Füss
Department of Finance, Accounting & Real Estate, European Business School (EBS), International University Schloss Reichartshausen, Rheingaustrasse 1, D-65375 Oestrich-Winkel, Germany
E-mail: roland.fuess@ebs.edu

Dr. Roland Füss is a professor at the Department of Finance, Accounting and Real Estate at the European Business School (EBS), International University Schloss Reichartshausen in Oestrich-Winkel, Germany. He holds a diploma in Economics from the University of Freiburg, Germany, where he also obtained his Ph.D. and his Habilitation. His research focuses on applied econometrics, risk management and alternative investments as well as on politics and financial markets. Roland has authored numerous articles in finance journals as well as book chapters. He is a member of the American Finance Association, the German Finance Association, the American Real Estate Society and the Verein für Socialpolitik.

Zeno Adams
Department of Finance, Accounting & Real Estate, European Business School (EBS), International University Schloss Reichartshausen, Rheingaustrasse 1, D-65375 Oestrich-Winkel, Germany
E-mail: zeno.adams@ebs.edu

Zeno Adams is a research assistant at the European Business School (EBS), International University Schloss Reichartshausen in Oestrich-Winkel. He obtained his diploma in Economics at the University of Freiburg, Germany. Zeno has authored articles in finance journals as well as book chapters. His research focuses on risk modelling, and the econometrics of financial markets.

Dieter G. Kaiser
Feri Institutional Advisors GmbH, Haus am Park, Rathausplatz 8-10, 61348 Bad Homburg, Germany. Tel.: +49 (6172) 916 3712. E-mail: dieter.kaiser@feri.de

Dr. Dieter Kaiser is Director Hedge Funds at Feri Institutional Advisors GmbH in Bad Homburg, Germany where he is responsible for managing hedge funds portfolios. From 2003 to 2007 he was responsible for institutional research at Benchmark Alternative Strategies GmbH in Frankfurt, Germany. Dieter Kaiser has written numerous articles on alternative investments that have been published in both, academic and professional journals and is the author and editor of seven books. Dieter received a BA in Business Administration from the University of Applied Sciences Offenburg, an MA in Banking and Finance from the Frankfurt School of Finance and Management, and a PhD in Finance from the Chemnitz University of Technology. On the academic side, he is a research fellow at the Centre for Practical Quantitative Finance of the Frankfurt School of Finance and Management.

The Predictive Power of Value-at-Risk Models in Commodity Futures Markets

Abstract

Applying standard value-at-risk (VaR) models to assets with non-normally distributed returns can lead to an underestimation of the true risk. Commodity futures returns are driven by continuous supply and demand shocks that lead to a distinct pattern of time-varying volatility. Due to these specific risk characteristics, commodity returns create the ideal environment for testing the accuracy of VaR models. Therefore, this paper examines the in- and out-of-sample performance of various VaR approaches for commodity futures investments. Our results suggest that dynamic VaR models such as the CAViaR and the GARCH-type VaR generally outperform traditional VaRs. These models can adequately incorporate the time-varying volatility of commodity returns and are sensitive to significant changes in the series of commodity returns. This has important implications for the risk management of portfolios involving passive long-only commodity futures positions. Risk managers willing to familiarize themselves with these complex models are rewarded with a VaR that shows the adequate level of risk even under extreme and rapidly changing market conditions as well as under calm market periods, during which excessive capital reserves would lead to unnecessary opportunity costs.

Keywords: *commodities, risk management, value-at-risk (VaR), GARCH modelling, conditional autoregressive value-at-risk (CAViaR), quantile regression*

JEL classification: *C*14, *C*22, *G*11, *G*13

The authors are grateful to the editor Stephen Satchell, and also to Jonathan Batten, Tom Aabo, Raphael Paschke, and the participants of the 2008 European Financial Management Association conference in Athens for helpful comments and suggestions. We would also like to thank Simone Manganelli from the European Central Bank for providing the Matlab codes for the CAViaR models. We bear responsibility for all remaining errors.

Introduction

Recent literature has established that commodity futures can serve as diversification instruments in conventional portfolios because of their low correlations with equities and bonds. It is thus possible to enhance portfolio diversification by adding commodity futures, certificates, structured notes, swaps, and/or futures, as well as options on commodity futures indices, to standard portfolios to reduce value at risk and improve overall performance.

Prior studies have found that using passive investable commodity futures indices as proxies for direct commodity investments show good stand-alone performance. In addition, adding commodity futures to a portfolio of stocks and bonds can substantially reduce portfolio risk (Anson, 2004 and Jensen *et al* 2000, 2002).

Commodity futures investors, however, typically face margin requirements, which must be sufficient to cover daily fluctuations in the value of their positions. The initial margin is usually only 5% to 15% of the contract's value. Hence, if at the settlement date the spot price deviates strongly from the futures price, investors will find themselves with highly leveraged positions. The leveraged nature of such investments will increase volatility and intensify the effect of price fluctuations, creating larger profits and losses than we would normally expect for traditional assets. An inability to maintain the futures position in the face of collateral calls or internal risk limits can then result in liquidation, generating a loss.

In contrast, if the cash position is too high, investors face opportunity costs because capital is not allocated to more profitable investments for the duration of the contract. Furthermore, long-term investments in passive long-only commodity indices exhibit higher short-term volatility and higher risk during commodity price downturns than during price upturns. Losses are intensified by a contangoed term structure, so if the term structure remains in contango for consecutive periods, the roll losses increase dramatically.

According to Samuelson's hypothesis, futures price volatility increases with the closeness to expiration date because of the mean-reverting behaviour of the spot price.[1] Thin trading is also likely to occur shortly before maturity, when many market participants are not interested in actual delivery of the underlying commodity. Generally, adding commodity futures to conventional portfolios provides diversification benefits, but it also poses risks that require accurate measures and appropriate risk management.

Commodities exhibit certain risk characteristics that are different from traditional assets like stocks and bonds. A good value-at-risk (VaR) model must capture those characteristics. First, many commodities, such as agricultural products, are not storable. Others, such as livestock and energy products like electricity, can only be stored at very high costs. As a result, supply or demand changes are translated immediately into price changes, which lead to higher volatility of commodity investments compared to traditional assets.

Second, supply and demand shocks occur more frequently and on a larger scale. Drought or frost can lead to an unexpected decrease in the supply of agricultural products and a subsequent sharp increase in prices. Natural disasters can also affect commodity prices. For example, when Hurricane Katrina hit New Orleans in 2005, the damage to petroleum production and refinery capabilities led to an increase in crude oil prices.[2]

Political instability in oil-exporting countries accounts for the additional variation in oil prices. Demand shocks occur less frequently and are generally on a smaller scale, but can

---

[1] See Samuelson (1965). The evidence for this effect is, however, mixed, and has been found mainly for agricultural futures (see, e.g., Galloway and Kolb (1996) and Bessembinder et al. (1996)).

[2] The price increase was partly offset when the Bush Administration tapped the strategic oil reserves a few days later.

significantly affect commodity prices. Increased ethanol production, for example, has led to an increase in corn futures prices. The rapid expansion of the Chinese economy has led to higher volatility in the energy and industrial metals sectors.

The third source of commodity price variation is the lack of governmental control. While central banks can influence stock and bond markets, they cannot compensate for supply shocks or changes in commodity prices.[3] The risk characteristics mentioned above are reflected in the return-generating process, which makes commodities the ideal time series for testing value-at-risk models: Long tranquil periods alternate with sudden volatility spikes or long periods of high volatility. Some commodity series do not show mean-reverting behavior, but rather exhibit shifts in volatility. Others switch between positive and negative skewness depending on the time period under investigation. These sudden changes in the return distribution pose a challenge to every VaR model.

Over the years, the uniform value-at-risk (VaR) measure has become the standard methodology for predicting market risk for financial institutions due to its conceptual simplicity and regulatory importance in quantifying market risk subject to the Basel II Accord. VaR has also received a great deal of attention from the academic side, because its measurement involves challenging statistical problems.

The VaR measure is defined as the loss a portfolio is not expected to exceed with a given probability over a predefined time horizon. The main difference among the many existing VaR approaches is the estimation of the distribution of portfolio returns. For example, daily financial data show unconditional return distributions, which often exhibit leptokurtosis and to some extent skewness. Hence, the assumption of normally distributed returns does not hold, and estimation of the normal VaR may be biased.

Although the Cornish-Fisher (CF) VaR adjusts the critical values of the standard normal distribution, both normal VaR and CF-VaR do not react sufficiently to changes in the return process, which can be problematic for forward-looking investment decisions (see Erb and Harvey, 2006). In contrast, the property of significantly autocorrelated absolute and squared financial returns allows the use of parametric models such as GARCH-type VaR (Engle, 1982, Bollerslev, 1986), and RiskMetrics (1996). Thus, volatility clustering enables us to consider market volatilities as quasi-stable, i.e., changing over the long run but remaining stable in the short run (see Manganelli and Engle, 2001).

The advantage of GARCH VaR and RiskMetrics compared to the more simple VaR measures is the incorporation of time-varying conditional volatility. The main drawback of these two models is again the assumption of standardized residuals following a specific analytical distribution that must be independently and identically distributed (*i.i.d.*) to estimate the unknown parameters.[4] Both approaches thus tend to underestimate (or overestimate) market risk if the distribution of standardized residuals does not follow the respective distribution.

To overcome these shortcomings, Engle and Manganelli (2004) introduced a conditional autoregressive specification of VaR based on a quantile regression framework. This approach

---

[3] Value-at-risk methods model negative price changes but the examples above discuss unexpected positive price changes. However, positive price shocks lead to higher volatility in general, which in turn leads to large negative returns.

[4] We estimated GARCH models using conditional normal distributions, *t*-distributions or generalized error distributions. The estimated degrees of freedom of the *t*-distribution were around 7, and the GED parameters were lower than 2, which suggests that a conditional distribution with heavier than normal tails should be preferred over a normal distribution. In this study, we use conditional *t*-distributions for the GARCH models, although the choice of the distribution only changed parameter estimates by around 0.01.

allows us to model a specific conditional quantile instead of the whole return distribution and is generally of the form $VaR_t = f\left(VaR_{t-1}, r_{t-1}\right)$. No explicit distributional assumptions for the time series are necessary for this model and by including lagged VaRs, $VaR_{t\text{-}1}$, as explanatory variables, the conditional autoregressive value-at-risk (CAViaR) model adapts to serial dependence in shortfall volatility.

Recent studies have evaluated the predictive performance of various VaR models. Kuester *et al* (2006) examine the fit of various parametric and non-parametric VaR models. They also cover the semi-parametric approaches of extreme value theory (Danielson and De Vries, 2000), and modified CAViaR specifications. The authors used daily Nasdaq composite index returns over thirty years, and found that most models performed poorly, except for conditionally heteroskedastic models.

Bao *et al* (2006) also compare various VaR approaches by using expected quantile loss functions and White's (2000) reality check test. They study the comparative risk forecast for five Asian emerging markets before, during, and after the 1997-1998 financial crisis. The results of these stress tests indicate that the RiskMetrics model works reasonably well during tranquil periods, while, not surprisingly, some EVT models do better during the crisis period.

Applications of VaR models to financial futures include Brooks *et al* (2005), as well as Cotter (2005). These authors compare different extreme value models for three LIFFE futures contracts. The empirical results show that the semi-nonparametric techniques yield superior results.

Giot and Laurent (2003) investigate commodity futures by focusing on applications of RiskMetrics, and various asymmetric GARCH VaR models. The results suggest that both skewed Student APARCH and skewed Student ARCH models provide good results for cash prices and nearby futures contracts.[5]

In contrast to Bao *et al* (2006) and Kuester *et al* (2006), we propose a new performance measure that combines familiar concepts such as the hit ratio or the correlation between a VaR model and squared returns with new elements such as a loss function that evaluates the discrepancy between the level of risk estimated by the VaR and the actually observed returns. Our measure accounts for the flexibility of VaR models, as well as the individual preferences of the risk manager. In line with previous results, we consider the following five approaches here:[6] normal VaR, CF-VaR, GARCH VaR, RiskMetrics, and CAViaR models.[7]

The remainder of this paper is organized as follows. The next section details the commodity futures asset class, which is analyzed empirically and reports basic summary statistics. Section 3 describes the methodology and the advantages and disadvantages of the various VaR approaches. Section 4 uses several performance measures to compare the in-sample performance and out-of-sample predictive ability of the VaR models. Section 5 gives our conclusions.

---

[5] The last three references also consider short positions on financial or commodity futures. The return data in this study consists of mainly negative skewness and extreme negative returns, so we confine ourselves to long-only positions.

[6] Our focus is on risk measurement at the sector level (Till, 2006), as portfolio allocations are individually different. Thus we leave the estimation of portfolio VaR to the individual investor.

[7] The EVT is not discussed here because it tends to provide reasonable results only for very extreme quantiles such as the 1% quantile. It is thus less appropriate for our 5% VaR. See Danileson and De Vries (2000) and Engle and Manganelli (2004).

Commodity Futures Indices

In this study, we use the S&P GSCI long-only passive excess return indices for five sectors: agricultural, energy, industrial metals, livestock, and precious metals. We obtain all data in daily frequency from the Thomson Financial Datastream database. We use S&P GSCI indices because of the quality of their reputation and their high level of open interest in the market for commodity futures, which serve as underlyings for derivatives and passive investment products (e.g., exchange-traded funds, certificates, etc.).[8]

Our sample period ranges from January 1, 1991, to December 31, 2006, and covers 4,175 observations. We use the last 523 observations as an out-of-sample period of one-step-ahead forecasts. From those 523, we use the last 23 for a 23-step-ahead forecast in order to evaluate the predictive performance over a monthly time span. Using daily data ensures the presence of ARCH effects (i.e., leptokurtosis and volatility clustering) in the return series, and allows us to include short-term volatility in the VaR.

We denote P_t as the price of the excess return index of a particular commodity sector at time t. We can thus calculate the continuous return for period t as follows:

$$r_t \equiv \ln\left(\frac{P_t}{P_{t-1}}\right)\cdot 100 \tag{1}$$

Table 1 gives an overview of the risk/return characteristics of the five commodity excess return subindices, denominated in U.S. dollars. The means and standard deviations are in annualized values.[9]

---

[8] The futures return can be decomposed into a spot and a roll return component, and is referred to as excess return. The spot return describes the percentage change in the near-month futures contract, since daily spot price data for commodities are not always available. In order to obtain continuous exposure to commodities and maintain a long future position, the futures contract must be rolled before expiration by selling the near-month futures and reinvesting the proceeds into the next-month futures contract. When the market is in backwardation, the roll return is positive; when it is in contango, the roll return is negative. Thus, the roll return stems from rolling up or down the term structure of futures prices. In comparison to the total return index as a fully collateralized commodity futures investment, the excess return measures the return of uncollateralized futures, which can be considered a leveraged spot position.

[9] The annualized average return is calculated by multiplying daily average returns by 250 (trading days). The volatility results from multiplying the daily standard deviation by $\sqrt{250}$.

Table 1 Descriptive statistics

Sectors	Mean (in %)	Volatility (in %)	Min (in %)	Max (in %)	Skewness	Kurtosis	Jarque-Bera test
Agriculture	-4.458	14.671	-3.694	4.457	0.157 (4.144)	4.661 (21.966)	494.711***
Energy	2.655	28.870	-14.407	10.545	-0.198 (-5.227)	6.185 (42.103)	1,782.7***
Industrial Metals	3.524	16.741	-9.018	7.576	-0.209 (-5.517)	8.510 (72.823)	5,282.6***
Livestock	-2.371	12.975	-4.251	3.243	-0.156 (-4.118)	4.013 (13.404)	194.369***
Precious Metals	0.399	13.969	-8.253	8.545	-0.175 (-4.620)	12.158 (121.02)	14,530***

This table presents summary statistics for daily excess returns of the S&P GSCI indices (4,153 observations). The sample period is 01/01/1991–11/29/2006. ***, **, and * denote significance at the 99%, 95%, and 90% confidence levels, respectively (rejection of the normal distribution). The test results of statistically significant deviations from 0 for skewness coefficients and from 3 (normal distribution) for the kurtosis coefficients according to Urzua (1996) are reported in parentheses. The critical values for the coefficient tests at the 1%, 5%, and 10% significance levels are 2.58, 1.96, and 1.65, respectively (standard normal distribution).

Note that the return properties of the different commodity sector indices vary substantially over our sample period. The highest return is 3.52 for industrial metals; the lowest is -4.46 for agriculture. The higher energy excess returns are probably due to the high demand for oil, which makes the highest contribution to this index. Furthermore, the particularly high demand for metals from emerging market economies like China and India causes an increase in the industrial metals index. Except for energy, which has much higher volatility, the indices exhibit similar volatility values of around 13%-17% per year. Precious metals and livestock have the lowest volatilities.

Note that only agriculture excess returns show significantly positive skewness subject to the Urzua test. In contrast, the theory suggests that non-storable commodities such as energy exhibit positive skewness. This is because, in the case of a fixed supply, a negative supply shock has a particularly strong impact on prices. Storable commodities can respond to negative supply shocks as long as supply is not depleted. The property of positive skewness that is often found in the literature usually refers to monthly returns, but cannot be considered a general property of commodity returns.[10]

All indices exhibit significant excess kurtosis, i.e., fat tails, which leads to a clear rejection of the null hypothesis of normally distributed returns according to the Jarque-Bera test. These high values for kurtosis seem to be common in commodity return distributions because of their occasionally aggressive price swings (this is particularly true for the energy and metals sectors).

[10] Kat and Oomen (2006) find the same results for daily futures returns. The authors also note that this contrasts sharply with the statement that futures returns are usually positively skewed because commodities have positive exposure to supply shocks. Their results also provide empirical evidence that spot and futures returns do not necessarily need to have the same properties.

In the case of the energy sector, these price spikes are driven mostly by unexpected increases on the demand side, e.g., during unexpectedly cold winters. Conversely, down spikes may occur in the summer when stores are near full capacity, or generally during periods of high volatility.

Comparison of Value-at-Risk Methodologies

In this section and in the next, we compare and test eight VaR models over three time periods. We use a long in-sample period from 01/01/1991 to 12/28/2004, which consists of 3,652 observations. We aim to provide a general comparison of the respective VaR characteristics over time, using information from the whole in-sample period.

The second period ranges from 12/29/2004 to 11/29/2006, and covers 500 daily one-step-ahead forecasts. This period presents a realistic situation in which an investor uses the preceding 250 days as an observation period to predict tomorrow's value at risk. The window is rolled over the next day and estimations are rerun, taking the new information into account. We present bootstrapped confidence intervals in order to show how safe investors should feel with their VaR forecasts. We use this period to evaluate the VaR models under more practical considerations.

The third period is a 23-step-ahead forecast over December 2006. We use this short time period to provide an indication of whether VaR models are useful for forecasts over a monthly period.

Conventional value-at-risk and the Cornish-Fisher expansion

Value-at-risk (VaR) is a downside risk measurement widely used by financial institutions for internal and external purposes. It has the appealing property of expressing risk in only one figure, and is the estimated loss of an asset that, within a given period (usually one to ten days), will only be exceeded by a certain small probability θ (usually 1% or 5%). Thus, the one-day 5% VaR shows the negative return that will not be exceeded within this day with a 95% probability:[11]

$$prob\left[return_t < -VaR_t \left| \Omega_t \right.\right] = 5\% \tag{2}$$

where Ω_t denotes the information set available at time t.

Statistically speaking, the 5% quantile of the probability density function of asset returns is considered. Assuming returns are normally distributed under a one-year observation period (250 trading days), VaR can be calculated as the deviation of the value from the distribution function η (-1.65), times the standard deviation σ from its mean μ:

$$VaR = \mu + \eta \cdot \sigma \tag{3}$$

In general, a higher probability value and a longer time period increase VaR. In this paper, we assume a period of one day and a probability of 95%.

Conventional VaR is based on the assumption of normally distributed returns. As the descriptive statistics in Table 1 show, however, financial asset and especially commodity returns often feature excess kurtosis and negative skewness. This makes the probability of

[11] Regulators require banks to compute a VaR at the 99% confidence level. However, many banks choose the 95% level for internal purposes of backtesting and model validation. See Jorion (2007), p. 147.

extreme returns more likely than under a normal distribution. Thus, applying VaR under the normality assumption can lead to a systematic underestimation of actual risk.

Favre and Galeano (2002) apply a Cornish-Fisher (CF) expansion by adjusting the critical value according to the distribution function η:

$$\eta_{CF} = \eta + \frac{1}{6}\left(\eta^{2} - 1\right)S + \frac{1}{24}\left(\eta^{3} - 3\eta\right)K - \frac{1}{36}\left(2\eta^{3} - 5\eta\right)S^{2} \tag{4}$$

where S is the skewness and K is the excess kurtosis of the empirical distribution, respectively. If the return distribution is normal, i.e., S and K equal 0, $\eta_{CF} = \eta$ according to Equation (4).

In contrast, for non-normally distributed commodity returns, we expect the VaR threshold obtained from the CF expansion to be more accurate in comparison to the estimated threshold from the conventional VaR. Figure 1 gives the 5% conventional VaR and CF-VaR for the industrial metals index.[12] Note that the development of CF-VaR over time is very similar to normal VaR. Although CF-VaR allows for the non-normality characteristics of the return process, it does not take volatility clustering into account. This leads to the kind of response inertia seen in Figure 1, which in turn results in consecutive hits in the case of volatility clusters of extreme negative returns.

[12] We chose the industrial metals index because of its volatile out-of-sample period that is used to demonstrate the characteristics of the individual VaR model.

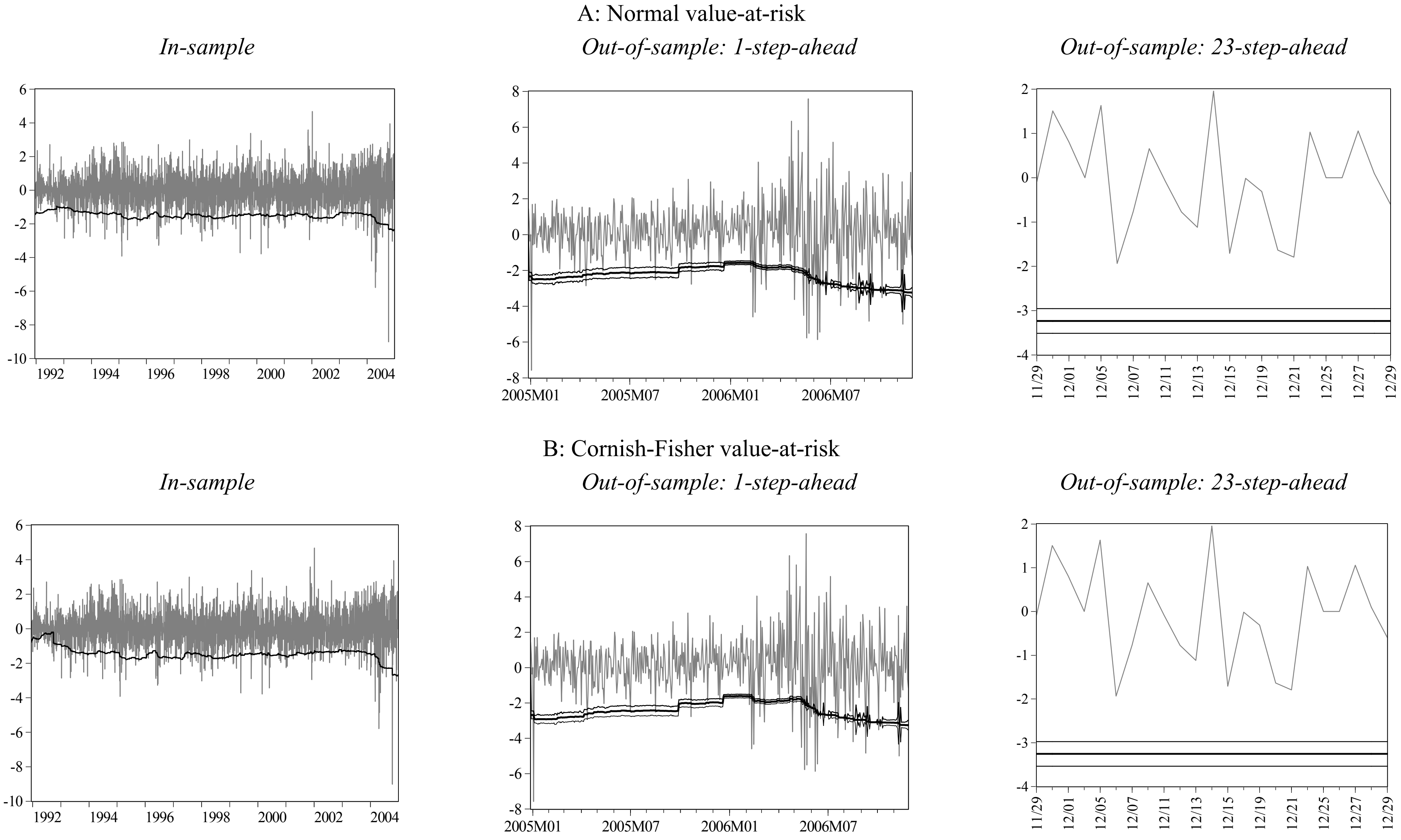


Figure 1 Normal and CF-VaR for the industrial metals index

The graphs in panel A show the variation of the normal value-at-risk over time, while panel B shows the development of the Cornish-Fisher VaR. The left-hand side boxes show the in-sample period from 01/01/1991 to 12/28/2004 (3,652 observations). The boxes in the middle present the out-of-sample period of one-step-ahead forecasts (12/28/2004 to 11/28/2006, 500 observations). The right-hand side boxes plot the 23-step-ahead forecasts from 1/29/2006 to 12/29/2006. Bootstrapped confidence intervals are computed using parametric bootstrap with 500 repetitions.

The one-step-ahead forecasts for both VaR models are the values from the preceding day, and the 23-step-ahead forecasts are simply straight lines, as commodity returns cannot be forecasted in an efficient market.[13] The confidence bands are very narrow, which is not surprising considering the low sensibility of the VaR models. To allow for volatility clustering, the next section presents VaR-type models that incorporate conditional price volatility, such as RiskMetrics and GARCH models.

GARCH-type VaR and RiskMetrics™

With ARCH effects, one bad day with highly negative returns makes a consecutive bad day more likely than without ARCH effects. Thus, our finding of ARCH effects in all five indices suggests that risk is systematically underestimated using normal or CF-VaR. In contrast to CF-VaR, GARCH-type VaR does not adjust the quantile of the distribution function. Rather, it replaces the rolling unconditional standard deviation by a more accurate conditional standard deviation that responds more quickly and more strongly to changes in the return process:

$$VaR_t = \mu_t + \eta \cdot \sqrt{h_t} \tag{5}$$

where $\sqrt{h_t}$ is the conditional standard deviation of a GARCH(1,1) model assuming a conditional *t*-distribution: $\sqrt{h_t} = \sqrt{\omega + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}}$.

Table 2 shows the estimated GARCH coefficients for the in-sample period. All coefficients are highly significant, so ARCH effects in the return series exist.

[13] We estimated ARIMA models for the returns (not shown here). Because of market efficiency, the estimated parameters were either insignificant or very small. Thus, the forecasted normal or CF-VaR remained basically unchanged.

Table 2 GARCH(1,1) model for GS commodity futures indices

Parameter	Agriculture	Energy	Industrial Metals	Livestock	Precious Metals
$\hat{c}$	-0.024*	0.019	-0.015	-0.000	-0.017*
	(0.013)	(0.025)	(0.014)	(0.012)	(0.010)
$\hat{\omega}$	0.033***	0.028***	0.022***	0.008***	0.005***
	(0.008)	(0.009)	(0.006)	(0.003)	(0.002)
$\hat{\alpha}$	0.068***	0.045***	0.051***	0.045***	0.052***
	(0.011)	(0.007)	(0.009)	(0.007)	(0.008)
$\hat{\beta}$	0.893***	0.947***	0.925***	0.945***	0.944***
	(0.017)	(0.008)	(0.013)	(0.009)	(0.007)
$\hat{\nu}$	7.462***	6.485***	6.734***	9.594***	3.999***
	(0.927)	(0.744)	(0.676)	(1.716)	(0.275)
AIC	2.532	3.878	2.608	2.347	2.146
SIC	2.541	3.887	2.616	2.355	2.155
LogL	-4,619.3	-7,077.9	-4,757.3	-4,280.6	-3914.4
Q(5)	14.087**	1.519	28.601*	8.857	1.369
$Q^2(5)$	6.633	16.676***	1.474	4.306	6.147
ARCH-LM test	0.680	0.002	0.115	0.238	3.806*
J.B. test	263.75***	391.15***	924.34***	75.32***	11,304***
$\hat{\alpha}+\hat{\beta}$	0.961	0.992	0.975	0.989	0.997
HLP	17.42	86.29	27.38	62.66	230.7
$\bar{\sigma}_{ann.}$ (in %)	14.544	29.580	14.832	13.484	20.412

This table reports GARCH(1,1) estimates based on daily continuously compounded commodity futures returns for the period 01/01/1991 to 12/28/2004 (3,652 observations). ***, **, and * denote significance of the GARCH coefficients at the 99%, 95%, and 90% confidence levels, respectively. Standard errors in parentheses. Conditional *t*-distribution with estimated degrees of freedom $\hat{\nu}$. The annualized volatility ($\bar{\sigma}_{ann.}$) and the half-life period (HLP) are computed as $\sqrt{\hat{\omega}/\left(1-\hat{\alpha}-\hat{\beta}\right)\cdot 250}$ and $log(0.5)/[log(\hat{\alpha}+\hat{\beta})]$, respectively.

The estimated degrees of freedom of the conditional *t*-distribution indicate the presence of fat tails. This indicates that the standardized residuals are not normally distributed even after taking GARCH effects into account. However, the Ljung-Box tests for standardized and squared standardized residuals and the ARCH-LM(1) tests show there are no GARCH effects left after estimation, although the evidence for energy and precious metals is not very strong. The sum of $\hat{\alpha}+\hat{\beta}$ is close to unity for all indices.

This high level of persistence in volatility can be intuitively described by the half-life period, which shows the number of days until half the volatility, generated by a price innovation, is decomposed. The half-life period, computed as $HLP = \log(0.5)/[\log(\hat{\alpha}+\hat{\beta})]$, takes approximately two to three months for the energy and the livestock indices. Thus, volatility shocks in commodity futures markets tend to persist for a very long time.

The precious metals index has an exceptionally high value of 240 days, which could call into question the property of mean reversion for this index. The unconditional volatility from the GARCH model, computed as $\sqrt{\frac{\hat{\omega}}{(1-\hat{\alpha}-\hat{\beta})}\cdot 250}$, is much higher (20.412) than the sample volatility in Table 1 (13.969). This is not the case for the other indices.[14] This behaviour is captured by the RiskMetrics model described below.

For a GARCH(1,1) model, the forecast for $T+1$ is $h_{T+1} = \hat{\omega} + \hat{\alpha}\varepsilon_T^2 + \hat{\beta}h_T$, where ε_T^2 and h_T are the last observations of the squared residuals and the conditional variance at the end of the 250-observation period. Compared to a general *k*-step-ahead forecast, the one-step-ahead forecasts contain only the estimation risk of the unknown GARCH parameters. But they are certain in the sense that the squared residuals and the conditional variance for period *t* are known.

We estimated the out-of-sample GARCH models in a rolling window of 250 days. We frequently encountered convergence problems of the maximum likelihood function, however, because in some cases no GARCH effects could be found for this short time period.[15] In those cases, we took the in-sample period, reflected in slightly altered confidence bands. We computed the confidence bands using a parametric bootstrap.

The 23-step-ahead forecasts are calculated as $h_{T+x} = \hat{\omega} + \left(\hat{\alpha} + \hat{\beta}\right)h_{T+x-1}$. Although better than the straight lines of the normal or CF-VaR, the forecasts do not include the unpredictable squared residuals ε_t^2, and are therefore less accurate than the one-step-ahead forecasts.

Table 2 shows the sum of the estimated GARCH parameters $\alpha + \beta$. For precious metals, it is 0.997, however, for the other indices it is also very high. The long half-life periods suggest that some indices may not show mean reverting behavior, but may be characterized by permanent volatility changes instead. If this is the case, volatility can be modelled more accurately by RiskMetrics™. RiskMetrics was introduced by J.P. Morgan and assumes that returns follow a conditional normal distribution, as follows:

$$r_t|\Omega_{t-1} \sim N\left(\mu_t, \sigma_t^2\right) \tag{6}$$

where μ_t is the conditional mean, σ_t^2 is the conditional variance, and Ω_{t-1} is the information set available at time *t* - 1. In RiskMetrics, μ_t is set to 0, and the return process can be expressed as:

$$r_t = \sigma_t \varepsilon_t, \text{ with } \varepsilon_t \sim i.i.d. \tag{7}$$

The variance is then modelled as:

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2, \text{ with } 0 < \lambda < 1 \tag{8}$$

The parameter λ is often set to values between 0.9 and 1. For daily data, we use λ = 0.94 as RiskMetrics (1996) suggests, because it delivers satisfying results for most financial time series. Because the parameters sum to unity and the conditional mean is set to 0, the return

[14] Furthermore, volatilities from the GARCH models are also slightly different for the other commodity indices because the time period in Table 1 encompasses our whole data range, while Table 2 refers only to the in-sample period.

[15] We tried different lengths of observation periods. Convergence problems tend to disappear with periods as long as 500 observations. The drawback, however, is that longer time periods make the VaRs less flexible and shocks must be carried through a much longer window (see Hendricks, 1996). For the sake of model comparison, we decided to stay at the shorter 250-day time period.

process $r_t = \sigma_t \varepsilon_t$ has a unit root and follows an integrated GARCH(1,1) model without a drift. Thus, shocks to the return process that increase volatility have a permanent effect on future volatility.

To calculate 5% VaR using RiskMetrics, the one-sided 5% quantile of the normal distribution with a mean of 0 and standard deviation σ_t is:

$$VaR_t = -1.65 \cdot \sigma_t \tag{9}$$

The RiskMetrics approach is thus easy and fast, making it convenient for VaR calculation.

For the RiskMetrics model, we calculate the one-step-ahead forecasts as $\sigma^2_{T+1} = \lambda \sigma^2_T + (1-\lambda) r^2_T$. Similar to the GARCH model, σ^2_T and r^2_T are known and allow for reasonable forecasts. In the general case, $T + x$, volatility is calculated as $\sigma_{T+x} = \sqrt{x} \cdot \sigma_{T+1}$, so it follows the square root of time rule. This leads to increasing forecasted volatility over time.

RiskMetrics has some drawbacks, however. For example, the parameter λ is predetermined and not estimated. In contrast, GARCH modelling estimates parameters directly using the maximum likelihood technique. Furthermore, RiskMetrics assumes the mean of the returns to be 0. However, for most financial returns the means are observed to be different from 0, and are often found to be statistically significant when applying GARCH models. In this case, the assumption of the square root of time rule fails, and the forecast equations cannot be applied. However, to determine whether GARCH VaR is superior to RiskMetrics, we must use performance measures, as discussed below.

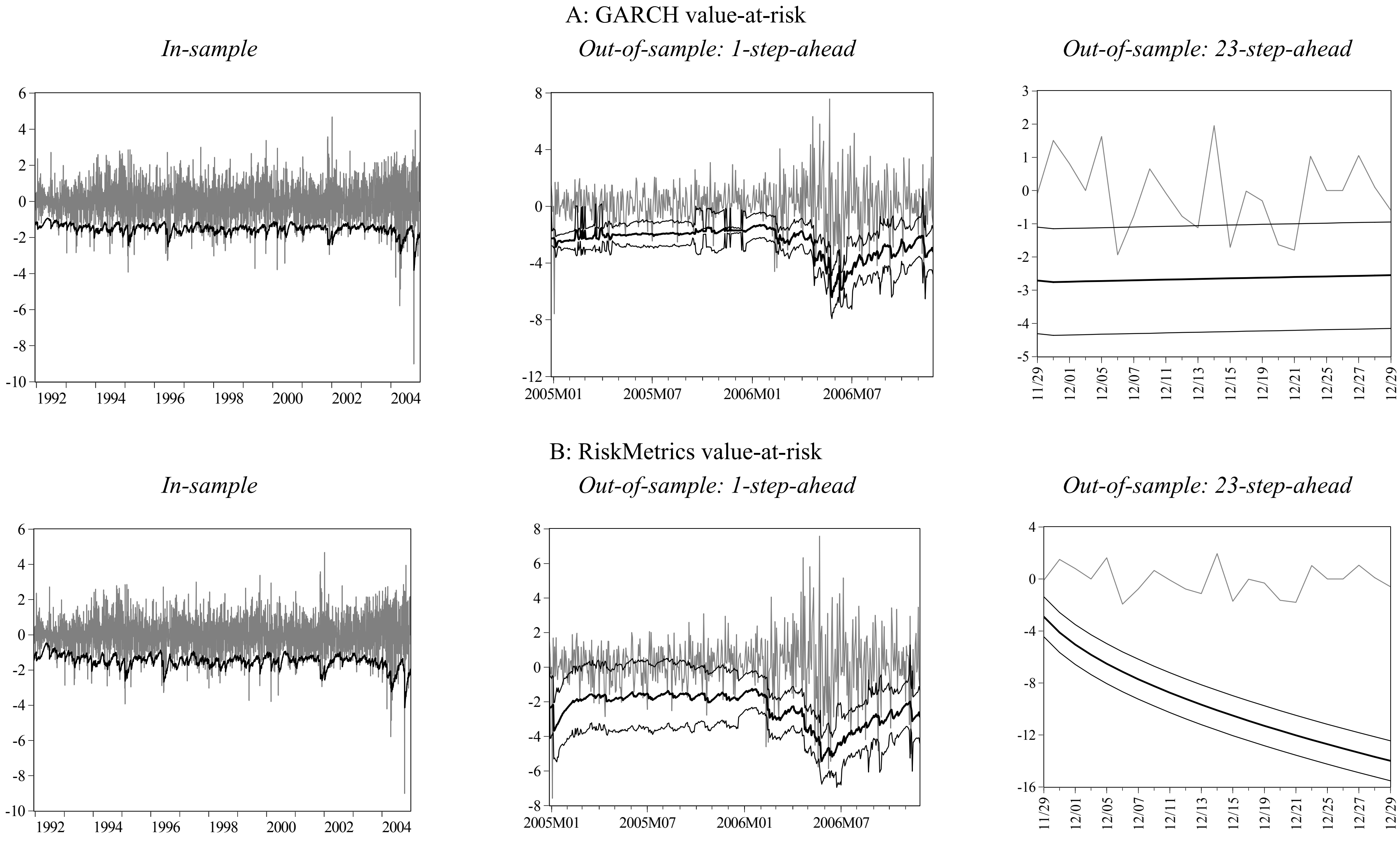


Figure 2 GARCH- and RiskMetrics-VaR for industrial metals index

The graphs in panel A show the variation of the GARCH VaR over time, while panel B shows the development of the RiskMetrics VaR. The left-hand side boxes show the in-sample period from 01/01/1991 to 12/28/2004 (3,652 observations). The boxes in the middle present the out-of-sample period of one-step-ahead forecasts (12/28/2004 to 11/28/2006, 500 observations). The right-hand side boxes plot the 23-step-ahead forecasts from 1/29/2006 to 12/29/2006. Bootstrapped confidence intervals are computed using parametric bootstrap with 500 repetitions.

Figure 2 shows the GARCH-type and RiskMetrics VaR for the industrial metals index. In contrast to the normal and the CF-VaR, both models show strong reactions to extreme return events. The conditional volatilities are an important contribution to the risk model, because the value of the hits is much lower than in the unconditional case. A strongly decreasing VaR gives the risk manager a warning. When hits occur, the capital requirements have increased and the risk manager is much better prepared. During tranquil periods, the VaRs are generally lower than either the normal or the CF-VaR, which allows investors to allocate reserves to more profitable investments. The normal and CF-VaR would lead to opportunity costs over the same period.

Comparing the GARCH and RiskMetrics models for the other commodity indices (not shown here) reveals that RiskMetrics is somewhat more fluctuating. It increases by more than the GARCH-VaR in turbulent periods, but it also decreases strongly during relatively tranquil periods.

The out-of-sample period shows larger confidence bands, because both VaRs are generally more variable. Returns that break through the upper confidence band indicate higher risk in terms of penetrating a risk-prone region, and should lead the risk manager to act with caution.

The 23-step-ahead forecasts are very different for the two models. In the GARCH model, the forecasted standard deviation returns to the long-run unconditional standard deviation. But the forecasted volatility of the RiskMetrics model does not have the mean reversion property; rather, it leads to a decreasing VaR over time due to the integrated structure of the model. It is obvious that the RiskMetrics model is inappropriate for general *k*-step-ahead forecasts.

Conditional autoregressive value at risk (CAViaR) models

The previous parametric models for VaR estimation were based on certain assumptions about the return distribution. They were clearly violated by the conventional VaR and CF-VaR, but also describe a possible weakness of the RiskMetrics model and the GARCH-VaR.

The conditional autoregressive value at risk (CAViaR) model developed by Engle and Manganelli (2004) uses a quantile regression framework to model the quantile directly instead of the whole return distribution.[16] This semi-parametric approach does not assume any distributional properties, and is thus more appropriate for estimating the VaR of commodity returns when the analytical distributions deviate from the population distribution of returns. In fact, the general formulation of CAViaR models allows for time-varying changes in the probability density of the error terms and the volatility.

We assume that VaR displays similar behavior to the volatility clustering seen in financial asset returns, because it is tightly linked to the standard deviation of the distribution. To illustrate this characteristic, Engle and Manganelli (2004) apply an autoregressive specification. If the probability associated with VaR is defined as θ, and r_t describes a vector of time t observable asset returns (e.g., lagged returns), the time t θ-quantile of the distribution of portfolio returns formed at $t-1$ is:

$$f_t(\beta) = f(r_{t-1}, \beta_\theta) \qquad (10)$$

[16] Koenker and Bassett (1978, 1982) introduced the methodology of quantile regression. For details on application in the context of VaR, see Engle and Manganelli (2004), who use a least absolute deviation (LAD) model that is more robust than OLS whenever the errors have a fat-tailed distribution, as they do here.

where β_θ is a p-vector of unknown parameters, and $f_t(\beta)$ denotes the θ% VaR, usually 5% or 1%.[17] A general type of CAViaR specification can be described as follows:

$$f_t(\beta)=\gamma_0+\sum_{i=1}^{q}\gamma_i f_{t-i}(\beta)+\sum_{i=1}^{p}\alpha_i l(r_{t-i},\varphi) \tag{11}$$

where $\beta'=(\alpha',\gamma',\varphi')$ and l are functions of a finite number of lagged values of observations. The autoregressive terms $\gamma_i f_{t-i}(\beta)$ and $i=1,\ldots,q$ allow for "smooth" changes of the quantile over time. On the other hand, the term $l(r_{t-i},\varphi)$ links $f_t(\beta)$ to observable variables that belong to the information set, and fulfill the function of the news impact curve in the context of GARCH modelling.

Engle and Manganelli (2004) propose four CAViaR specifications. Each differs slightly in how it includes past return observations, lagged VaR values, and asymmetric information in the model. The simplest is the *adaptive model*, where the VaR increases when a hit occurs in the last observation, and decreases slightly otherwise:

$$f_t(\beta_1)=f_{t-1}(\beta_1)+\beta_1\left\{\left[1+\exp\left(G\left[r_{t-1}-f_{t-1}(\beta_1)\right]\right)\right]^{-1}-\theta\right\} \tag{12}$$

For $G\to\infty$ and considering that $f_t(\beta)$ is the θ% VaR, Equation (12) can be expressed in a simpler form as:

$$VaR_t=VaR_{t-1}+\beta_1\left(hit_{t-1}\right)$$

and

$$hit_t=I\left(r_t<-VaR_t\right)-\theta \tag{12'}$$

where $I(\cdot)$ is the indicator function and θ is the VaR probability. The adaptive model considers only the hits, not the returns that are close to VaR.

The *symmetric absolute value model* in Equation (13) also accounts for the absolute value of the lagged return. It estimates the parameter on the lagged VaR instead of restricting it to 1:

$$VaR_t=\beta_0+\beta_1 VaR_{t-1}+\beta_2\left|r_{t-1}\right| \tag{13}$$

The third model, the *asymmetric slope model*, assumes that positive and negative past returns affect the VaR differently:

$$VaR_t=\beta_0+\beta_1 VaR_{t-1}+\beta_2\left|r_{t-1}{}^{+}\right|+\beta_3\left|r_{t-1}{}^{-}\right| \tag{14}$$

where $\left(r^{+}\right)=\max(r,0)$, and $\left(r^{-}\right)=-\min(r,0)$. Assuming the tails of the distribution follow a different dynamic than the whole distribution but can still be modelled by GARCH, we apply an *indirect GARCH(1,1) model*:

$$VaR_t=k\left(\beta_0+\beta_1\left(\frac{VaR_{t-1}}{k}\right)^2+\beta_2 r_{t-1}^2\right)^{1/2} \tag{15}$$

where k is usually set to 1. The vector of parameters β of the CAViaR model is estimated with non-linear autoregressive quantile regression. Quantile regression was introduced by Koenker and Bassett (1978), and has been extended to many linear and non-linear cases (for

[17] In the following derivations, the subscript from β_θ is neglected for notational convenience.

examples, see Bloomfield and Steiger (1983) for the linear case and White (1994) for the non-linear case). Consider the following model:

$$r_t = f\left(y_{t-1}, x_{t-1}, \ldots, y_1, x_1; \beta^0\right) + \varepsilon_{t\theta} = f_t\left(\beta^0\right) + \varepsilon_{t\theta}, \text{ for } t = 1, \ldots, T \tag{16}$$

$$Quant_\theta\left(\varepsilon_{t\theta} \middle| \Omega_t\right) = 0$$

where x is a vector of regressors, and $Quant_\theta\left(\varepsilon_{t\theta} \middle| \Omega_t\right)$ is the θ-quantile of $\varepsilon_{t\theta}$ conditional on the available information set Ω_t. The θ^{th} regression quantile is the estimated vector of parameters $\hat{\beta}$ that solves the following minimization problem:

$$\min \frac{1}{T}\left[\theta - I\left(r_t < f_t\left(\beta^0\right)\right)\right]\left[r_t - f_t\left(\beta^0\right)\right] \tag{17}$$

Instead of estimating the least squared residuals, the regression quantile objective function estimates the least absolute deviations (LAD) when the return is even more negative than the VaR. Thus the indicator function becomes 1. When the return is not as negative as the VaR, the indicator function is 0.

Koenker and Bassett (1978) showed that the variance of the mean is slightly lower in OLS regression when the distribution is normal, but is higher than the variance of the median in LAD regression under all other distributions. (For the exact mathematical derivation, see Engle and Manganelli, 2004, who show that the quantile estimator $\hat{\beta}$ is consistent and asymptotically normal.)

Table 3 shows the estimated $\hat{\beta}$ coefficients for the four CAViaR models.

Table 3 Estimation results of the 5% CAViaR specifications for S&P GSCI commodity futures indices

Parameters	Agriculture	Energy	Industrial Metals	Livestock	Precious Metals
	Symmetric Absolute Value				
$\hat{\beta}_0$	0.042*** (0.010)	0.076*** (0.017)	0.059*** (0.019)	0.085*** (0.027)	0.035*** (0.006)
$\hat{\beta}_1$	0.955*** (0.009)	0.943*** (0.012)	0.944*** (0.021)	0.913*** (0.035)	0.952*** (0.008)
$\hat{\beta}_2$	0.091*** (0.010)	0.100*** (0.020)	0.076*** (0.021)	0.148*** (0.039)	0.117*** (0.010)
$\hat{\beta}_3$	-	-	-	-	-
	Indirect GARCH				
$\hat{\beta}_0$	0.015** (0.008)	0.098** (0.044)	0.077*** (0.031)	0.061*** (0.025)	0.017** (0.008)
$\hat{\beta}_1$	0.955*** (0.004)	0.936*** (0.007)	0.925*** (0.016)	0.894*** (0.019)	0.947*** (0.007)
$\hat{\beta}_2$	0.103** (0.062)	0.143*** (0.044)	0.095*** (0.032)	0.214*** (0.072)	0.095 (0.127)
$\hat{\beta}_3$	-	-	-	-	-
	Asymmetric Slope				
$\hat{\beta}_0$	0.012 (0.009)	0.052** (0.028)	0.046* (0.031)	0.032** (0.018)	0.013** (0.006)
$\hat{\beta}_1$	0.960*** (0.010)	0.934*** (0.016)	0.930*** (0.033)	0.935*** (0.022)	0.948*** (0.008)
$\hat{\beta}_2$	0.086*** (0.015)	0.098*** (0.021)	0.051** (0.025)	0.023* (0.016)	0.104*** (0.008)
$\hat{\beta}_3$	0.051*** (0.019)	0.114*** (0.037)	0.110*** (0.036)	0.158*** (0.035)	0.076*** (0.027)
	Adaptive				
$\hat{\beta}_0$	-	-	-	-	-
$\hat{\beta}_1$	0.143*** (0.024)	0.191*** (0.033)	0.095*** (0.027)	0.182*** (0.034)	0.081*** (0.025)
$\hat{\beta}_2$	-	-	-	-	-
$\hat{\beta}_3$	-	-	-	-	-

This table shows the coefficient estimates of the CAViaR models according to equation (12) to (15). ***, **, and * denote significant coefficients at the 1%, 5%, and 10% significance levels, respectively. Calculations are based on daily data for the time period 01/01/1991 to 12/29/2006 (4,175 observations). Standard errors are in parentheses.

Note that the parameter of the lagged VaR, $\hat{\beta}_1$, is always significant. This implies volatility clustering is also relevant for the tails of the return distribution. Furthermore, the parameter $\hat{\beta}_3$ from the asymmetric slope model is also significant, which means that past negative returns have a stronger effect on the VaR than positive returns. The significant parameters of the indirect GARCH model show that GARCH effects also apply to the tails of the return distribution. The four CAViaR models for the industrial metals index are shown in Figure 3.

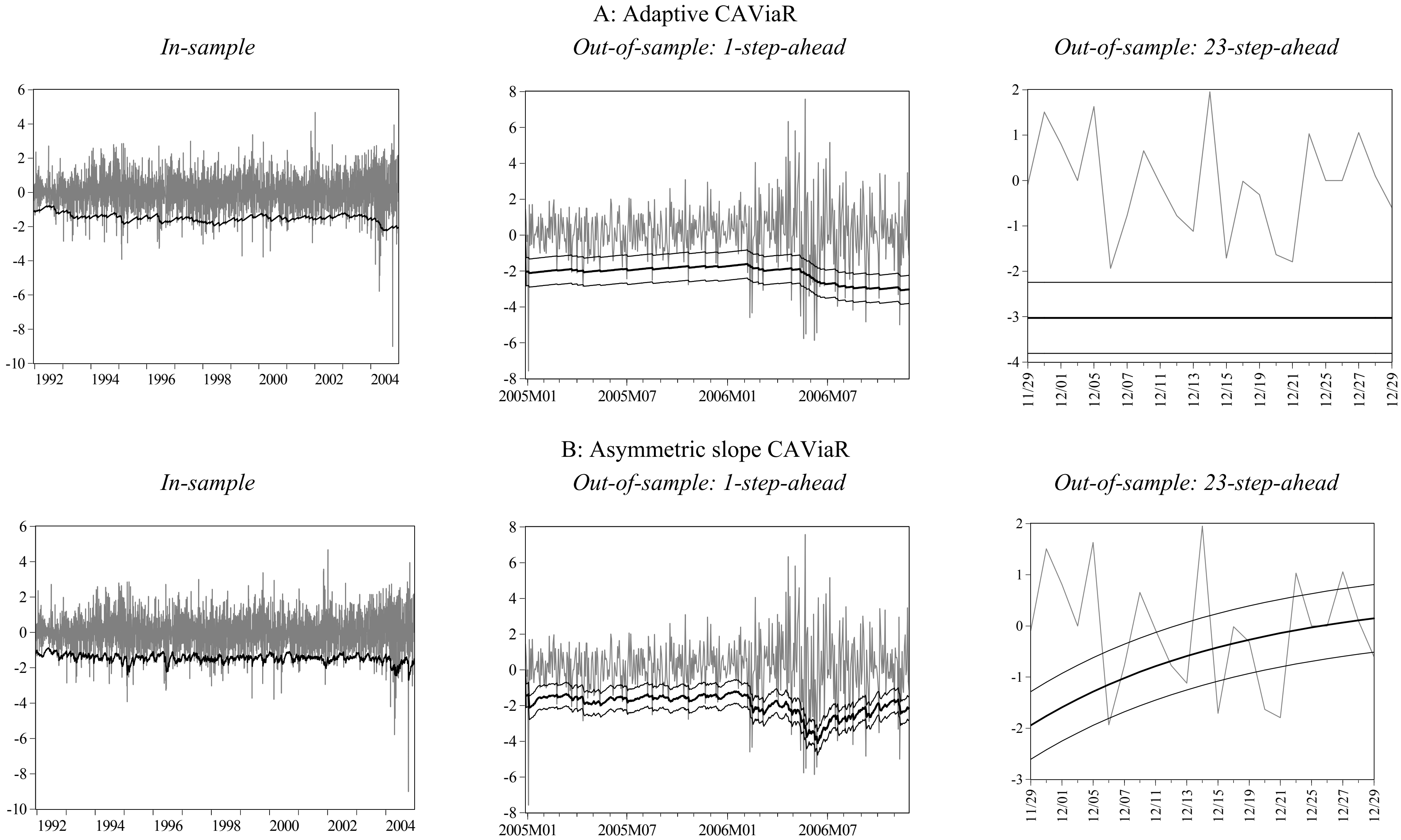


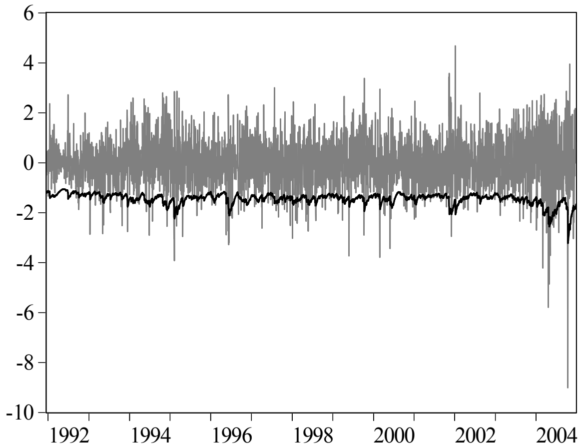
Figure 3 CAViaR models: adaptive, asymmetric slope, indirect GARCH, and symmetric absolute value for industrial metals index

The graphs in panel A show the development of the Adaptive CAViaR over time, while panel B shows the variation of the Asymmetric CAViaR model. Panel C and D present the indirect GARCH CAViaR and the Symmetric Absolute Value model, respectively. The left-hand side boxes show the in-sample period from 01/01/1991 to 12/28/2004 (3,652 observations). The boxes in the middle present the out-of-sample period of one-step-ahead forecasts (12/28/2004 to 11/28/2006, 500 observations).

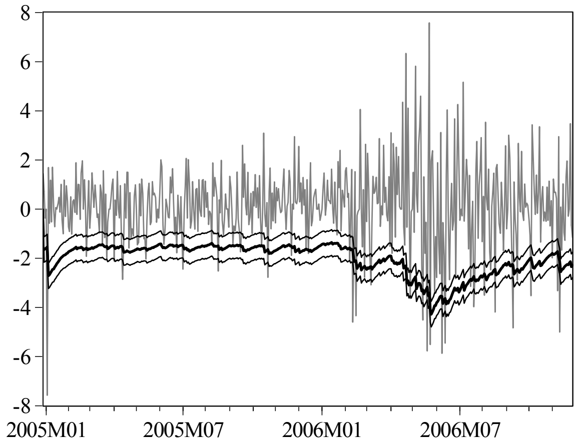
The right-hand side boxes plot the 23-step-ahead forecasts from 1/29/2006 to 12/29/2006. Bootstrapped confidence intervals are computed using parametric bootstrap with 500 repetitions.

C: Indirect GARCH CAViaR

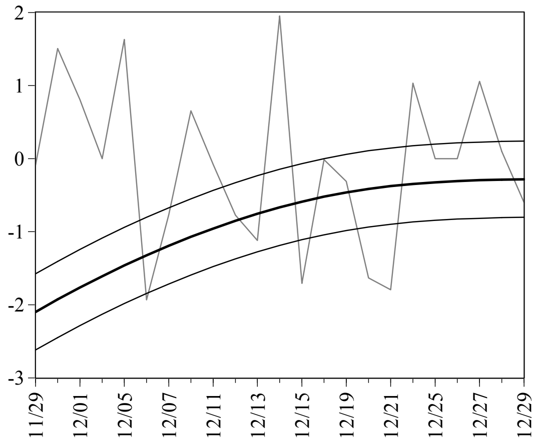
In-sample



Out-of-sample: 1-step-ahead

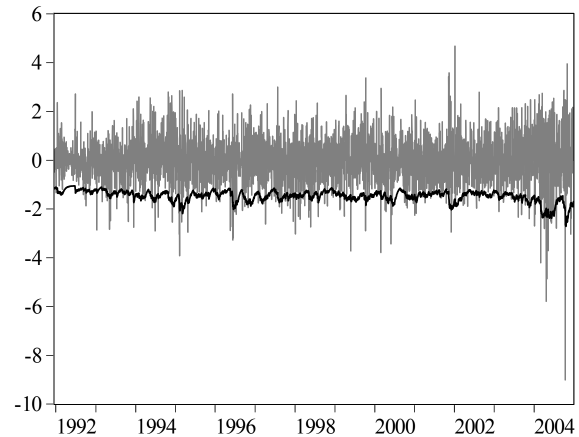


Out-of-sample: 23-step-ahead

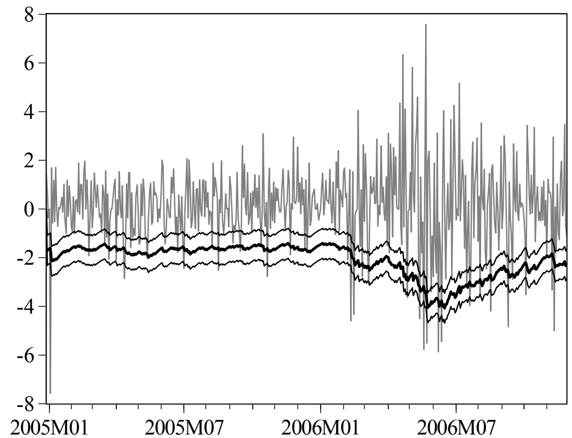


D: Symmetric Absolute Value CAViaR

In-sample



Out-of-sample: 1-step-ahead



Out-of-sample: 23-step-ahead

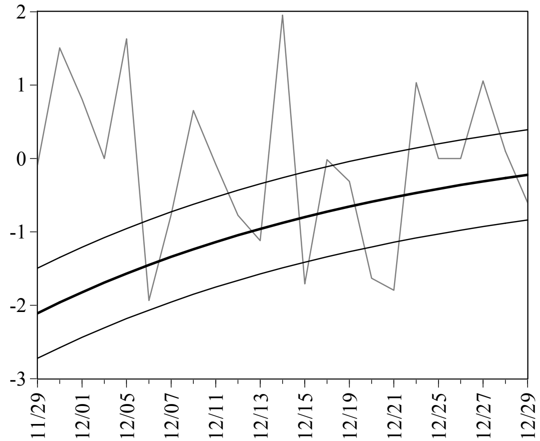


Figure 3 Continued.

Except for the adaptive model, all models show strong reactions to changes in return volatility. The adaptive model only reacts when hits occur, so it is not very sensitive to changes in the return process. Although this behavior could also be achieved by the GARCH-type VaRs, those models had much wider confidence bands and consequently more returns in the risk-prone region. Interestingly, the adaptive model has wider confidence bands than the other CAViaR models, but it is much less sensitive to changes in returns.[18]

The 23-step-ahead forecasts require lagged values of the VaR, and in some cases lagged positive and negative returns or hits. However, future returns or even hits are not available, so the forecasts must rely on constants and past VaR values.

Performance Evaluations

The previous section gave some indication of which VaR model might perform best in terms of distribution modeling and performance quality. This section uses three performance measures to compare the VaR model in more detail. We need to determine 1) which model is most sensitive to changes in the returns (correlation comparison), 2) in which model hits occur only 5% of the time, and 3) which models best manage the trade-off between sufficient reserves and efficient capital allocation (VaR performance criterion).

A VaR model capable of indicating the correct level of risk even during periods of high volatility should be able to respond quickly to changes in the return process. Figure 4 shows the correlations between squared returns and the respective VaR model. We used squared returns because both positive and negative return changes lead to higher volatility in general, and thus influence the VaR. Moreover, it is more important that a VaR model can capture large return fluctuations that indicate increased financial or even default risk than small return changes.

As expected, we find the strongest negative correlations with the GARCH-type VaRs and the CAViaR models, except for the adaptive VaR. The normal, CF, and adaptive VaRs remain more or less unchanged over short time periods, which is reflected in their lower correlation values. The out-of-sample period shows that the correlation structure, although similar to the in-sample period, covers correlation coefficients ranging from -0.31 to +0.07. This shows that the flexibility of the VaR models also depends on the underlying return series. The risk manager should therefore compare several VaR models to decide which one fits the underlying data best.

[18] Engle and Manganelli (2004) show that the CAViaR parameters are asymptotically normally distributed. The confidence bands were consequently estimated using a parametric bootstrap with fifty repetitions.

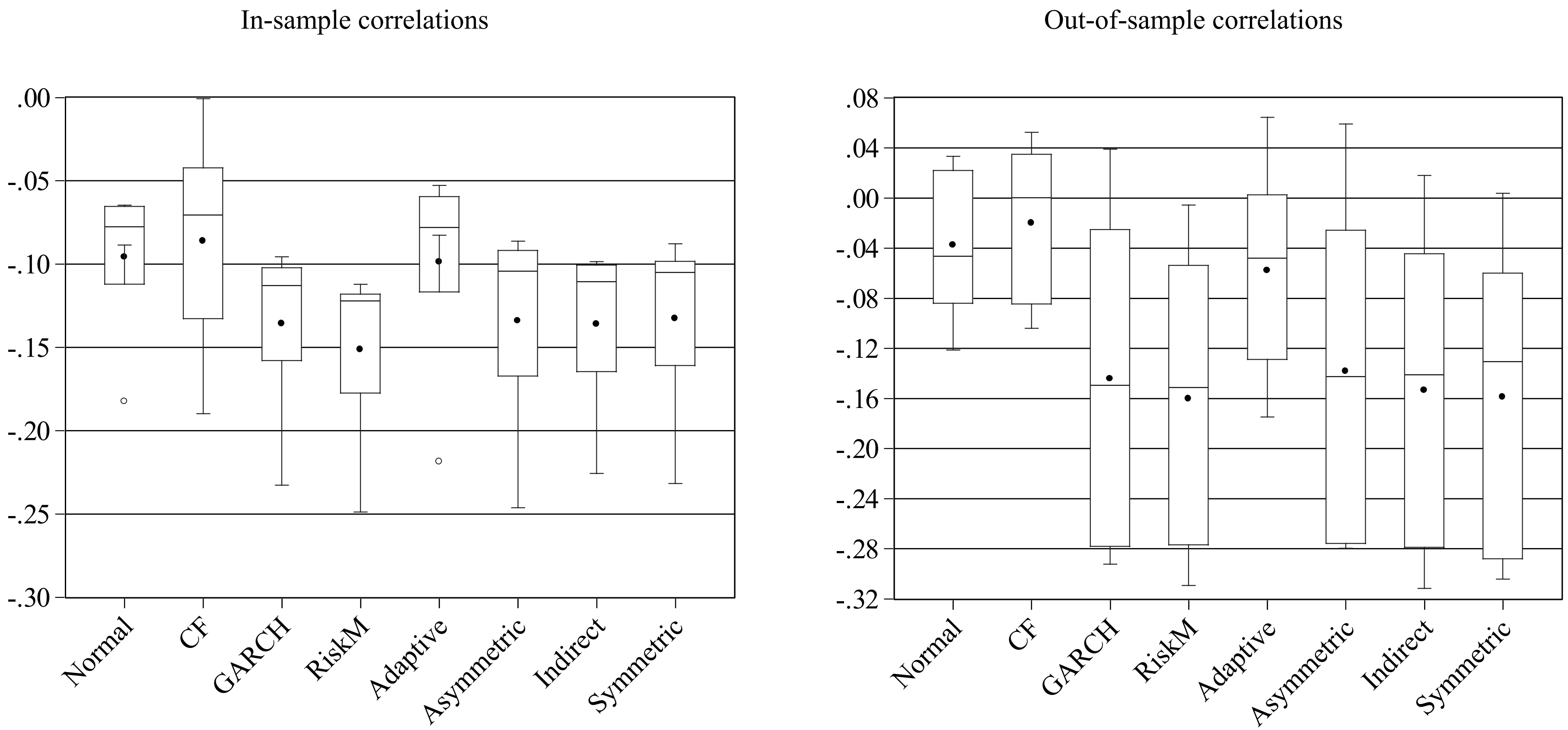


Figure 4 In-sample and out-of-sample correlations between squared returns and value-at-risk models for GS commodity indices

These two graphs show box-plots of in- and out-of-sample correlations between squared returns of different VaR models for the five commodity futures indices. The in-sample period ranges from 01/01/1991 to 12/28/2004 (3,652 observations). The out-of-sample period ranges from 12/28/2004 to 11/28/2006 (500 observations).

The hit ratio is perhaps the most common performance measure for value-at-risk models. The hit ratio reflects the number of times the return was more negative than the VaR. Thus, in our 5% VaR models, we expect hits to occur 5% of the time.

The upper panel of Figure 5 shows the in- and out-of-sample hit ratios for the industrial metals index. During the in-sample period, the three CAViaR models are close to 5%, while the GARCH and the CF-VaR overestimate and underestimate the risk, respectively. During the out-of-sample period, however, the GARCH model is now closer to the expected value and the CAViaR models underestimate the risk, resulting in hit ratios larger than 5%. This again shows the importance of measuring risk with several VaR methods, rather than relying on just one.

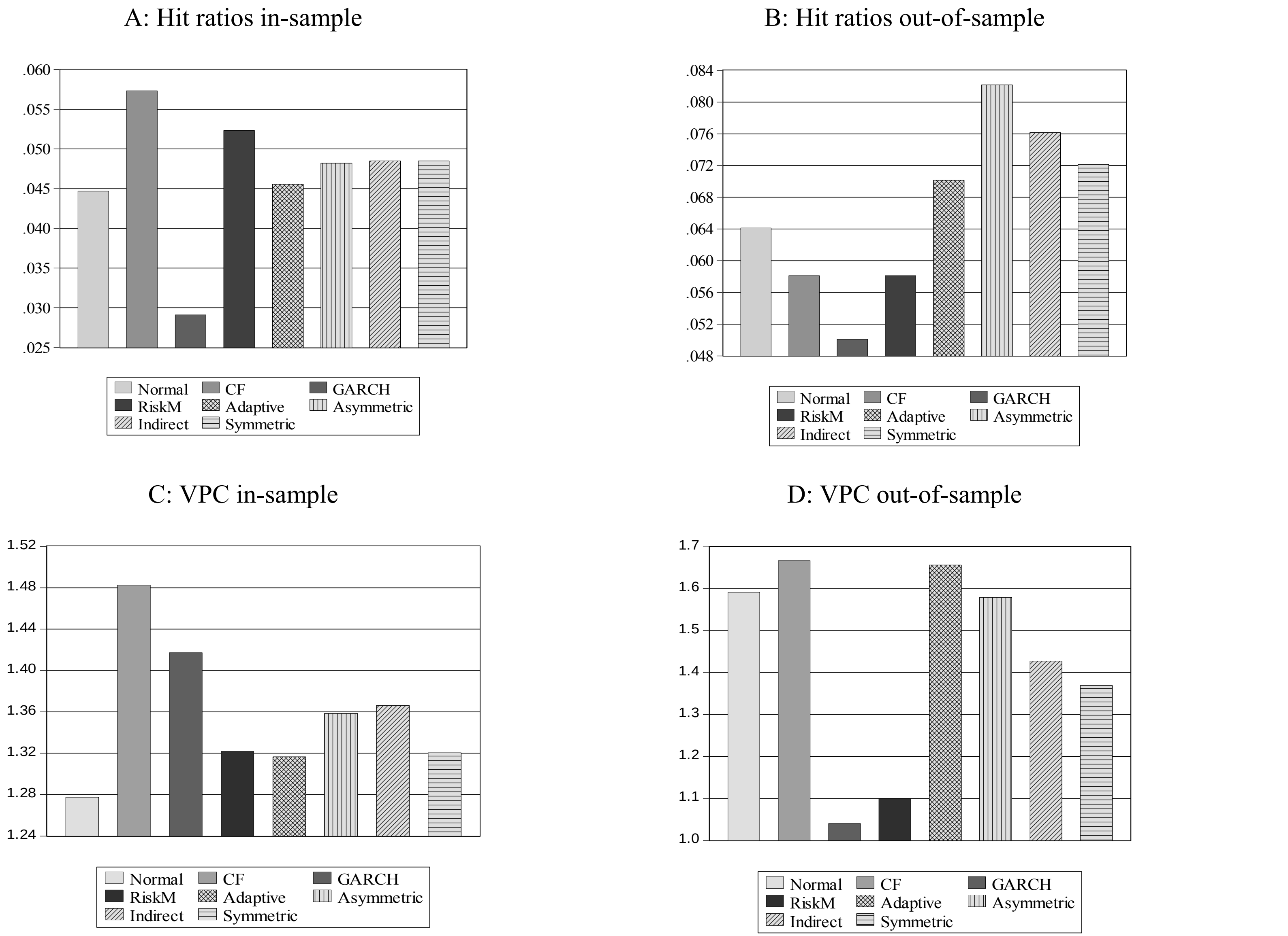


Figure 5 Hit ratios and VaR performance criterion for industrial metals index

The graphs in panel A and B show the hit ratios for the in-sample and out-of sample period, respectively. The graphs in panel C and D present the values of the VaR performance criterion (VPC) for the in-sample and out-of sample period. The in-sample period ranges from 01/01/1991 to 12/28/2004 (3,652 observations), while the out-of-sample period ranges from 12/28/2004 to 11/28/2006 (500 observations).

An overview of the hit ratios for all indices in Table 4 reveals that the hit ratios are fairly close together during the in-sample period, but can vary substantially during the short out-of-sample period. The GARCH model produces the best results. However, there is a general tendency to over- or underestimate the risk depending on the underlying index.

Table 4 Hit ratios for GS commodity futures indices

Sectors	Normal	CF	GARCH	Risk Metrics	Adaptive	Asym-metric	Indirect	Sym-metric
				In-sample				
Agriculture	0.0473	0.0523	0.0306	0.0526	0.0479	0.0538	0.0517	0.0514
Energy	0.0535	0.0547	0.0335	0.0556	0.0523	0.0526	0.0538	0.0529
Industrial M.	0.0500	0.0617	0.0323	0.0556	0.0509	0.0532	0.0535	0.0529
Livestock	0.0570	0.0567	0.0353	0.0600	0.0494	0.0526	0.0520	0.0509
Precious M.	0.0485	0.0747	0.0282	0.0497	0.0491	0.0511	0.0514	0.0509
				Out-of-sample				
Agriculture	0.0401	0.0441	0.0381	0.0381	0.0521	0.0421	0.0481	0.0401
Energy	0.0341	0.0421	0.0321	0.0481	0.0441	0.0361	0.0481	0.0401
Industrial M.	0.0641	0.0581	0.0501	0.0581	0.0701	0.0822	0.0762	0.0721
Livestock	0.0561	0.0561	0.0481	0.0661	0.0461	0.0541	0.0641	0.0601
Precious M.	0.0701	0.0641	0.0601	0.0641	0.0701	0.0641	0.0822	0.0541

This table show the hit-ratios, i.e. the percentage of times when the return is more negative than the VaR for the in-sample and out-of-sample period. The in-sample period ranges from 01/01/1991 to 12/28/2004 (3,652 observations), while the out-of-sample period ranges from 12/28/2004 to 11/28/2006 (500 observations).

Comparing hit ratios, however, can be dangerous. In fact, it is not hard to find a horizontal line where 5% of the returns ex-post lie below the line, and 95% of the returns are above the line. A good VaR measure must recede during tranquil periods to allow for the investment of unneeded reserves, and increase enough in absolute terms during volatile periods so that the corresponding hit value is close to the VaR. After all, risk managers are most concerned about the unexpected hits that constitute a multiple of the VaR.

In order to consider the trade-off between efficient capital allocation and sufficient reserves, we compare VaR models using what we call the VaR Performance Criterion (VPC). Similarly to Bao *et al.* (2006), this performance measure is based partly on a loss function that measures the distance from the returns to VaR. It also incorporates the correlation measure, and imposes penalties if the hit ratio deviates from its theoretical value.

$$VPC = w_1 \cdot \frac{1}{n}\sum_{h=1}^{H}\left(hitvalue_h - VaR_h\right)^2 + w_2 \cdot \left|\frac{1}{n}\sum_{j=1}^{J}\sqrt{VaR_j - R_j}\right| \cdot I\left(R_j < 0\right) + w_3\rho + w_4\left|\theta - ratio\right| \tag{18}$$

where n is the number of observations, H is the number of hits, J is the number of negative returns that do not constitute hits, and θ is the theoretical hit ratio (here: 5%). The first two terms in Equation (18) show the trade-off.

Because hits are dangerous when they get large, the first term is squared. In contrast, the opportunity costs of reserves are less important as long as risk management is concerned;

therefore, the second term comes with a square root. We calculate the correlation coefficient ρ as the correlation between the VaR and squared returns.

The last term in (18) measures the deviation of the hit ratio from its theoretical value. The weights w_1, w_2, w_3, and w_4 sum to unity, and are set so that each term contributes a fraction to the VPC measure that reflects the risk manager preferences. To compute the weights, the VPC is first calculated for a specific index over each of the eight VaR models using equal weights, i.e., $w_i = 0.25$. The weights are then calculated for each VaR model according to the risk manager preferences. In a second step, we take the median over all weights and reestimate the VPC using the new weights to obtain the final VPC measure.[19] This procedure is also repeated for the other indices. In consideration of risk protection as the most important term, we set $w_1 = 0.55$, $w_2 = 0.10$, $w_3 = 0.30$, and $w_4 = 0.05$.

The VPC for the industrial metals index is shown in the lower panel of Figure 5, where lower values indicate better VaR models. The GARCH-type models and the three CAViaR models (i.e., except the adaptive model) are the best performers, with the GARCH-type models featuring particularly low VPC values in the out-of-sample period. The VPC values for all indices shown in Table 5 confirm the results of the industrial metals index for the in-sample period.

[19] Taking the mean instead of the median does not result in large changes of the VPC. However, some robustness checks indicated outliers in the weights, so we prefer the median. Furthermore, an earlier version of the VPC involved using the weights of only one model (the benchmark VaR) instead of the median over all models. This also did not result in very large changes, but raised the problem of selecting the appropriate benchmark VaR.

Table 5 VaR performance criterion for S&P GSCI commodity futures indices

Sector	VaR Measures							
	Normal	CF	GARCH	Risk Metrics	Adap-tive	Asym-metric	Indirect	Sym-metric
	In-Sample							
Agriculture	1.4243	1.4410	1.2636	1.1492	1.4453	**1.1948**	1.2196	1.2233
Energy	1.5599	1.5260	1.6057	1.4811	1.6467	1.5542	**1.4740**	1.5562
Industrial M.	1.3433	1.4344	1.3109	**1.2498**	1.3191	1.2520	1.2788	1.2556
Livestock	1.3813	1.3155	1.0798	1.2332	1.1145	**1.0306**	1.1028	1.0890
Precious M.	1.3290	2.1658	1.2667	**1.2215**	1.4038	1.2335	1.2703	1.2368
	Out-of-Sample							
Agriculture	1.5929	1.6947	1.3060	**1.2099**	1.5991	1.3100	1.2861	1.2860
Energy	**1.3145**	1.3644	1.7154	1.8882	1.6428	1.4565	1.7865	1.4860
Industrial M.	1.5913	1.6669	**1.0402**	1.0988	1.6562	1.5793	1.4274	1.3689
Livestock	1.4102	1.3532	1.3542	1.8909	**1.2664**	1.4783	1.7147	1.5946
Precious M.	1.7011	1.6088	1.0637	1.1402	1.5271	1.1940	1.3776	**0.9789**

This table shows the values for the VaR performance criterion (VPC) for all five commodity indices and all eight VaR models for the in-sample and out-of sample period. The in-sample period ranges from 01/01/1991 to 12/28/2004 (3,652 observations), while the out-of-sample period ranges from 12/28/2004 to 11/28/2006 (500 observations); bold numbers indicate the best performance.

In most cases, the normal, CF, and adaptive VaR have much higher VPC values than the other VaRs. At first glance, the evidence for the out-of-sample period seems mixed. For the energy and the livestock index, these VaR models are not the worst performers, but they have relatively low values. This is because the out-of-sample period for those indices was relatively tranquil. The GARCH-type and CAViaR models show their power only during waves of high volatility. To illustrate, Figure 6 shows one extreme example.

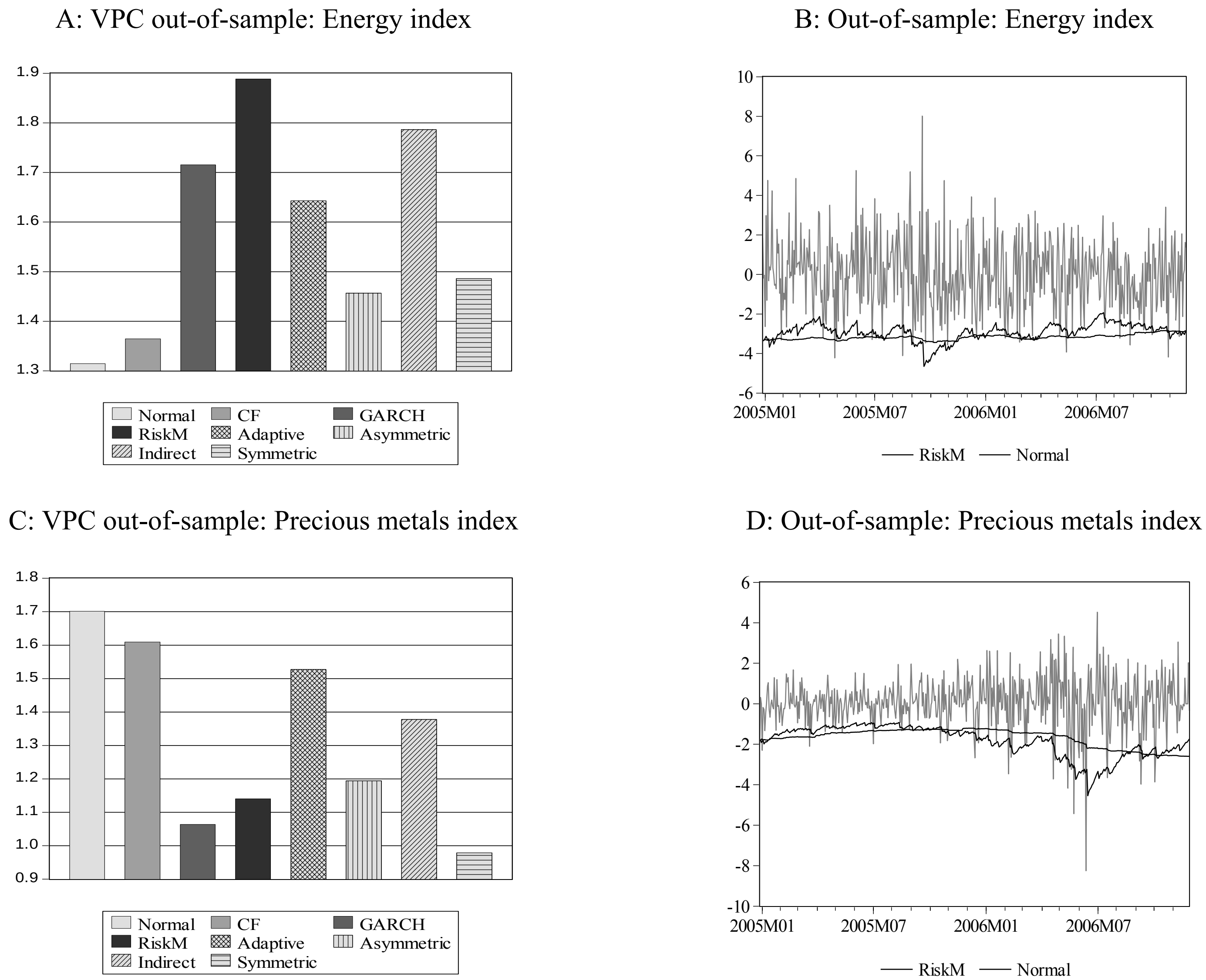


Fig. 6. VaR performance criterion for energy and precious metals index.

Panel A shows the VPC for the energy index. The unexpected good performance of the normal VaR and the CF VaR is due to the specific time period of the energy index shown in panel B and does not generalize to the whole sample range or all indices. This is shown in panel C and D where the normal VaR and CF VaR turn out to be the worst performers since they are unable to adjust to the change in volatility of the precious metals index adequately. The out-of-sample period ranges from 12/28/2004 to 11/28/2006 (500 observations).

The upper panel of Figure 6 shows the out-of-sample period for the energy index and the corresponding VPC values. Here, the normal VaR has the lowest value, while the RiskMetrics-VaR has the highest. If we compare this situation to a more volatile period like the out-of-sample period of the precious metals index in the lower panel, we see quite an opposite picture. The normal VaR now becomes the worst performer, and the RiskMetrics-VaR is among the best performers. This shows that short out-of-sample periods should be read together with longer in-sample periods that cover a time span where accidentally tranquil periods can be ruled out.

We next compare the VPC to Christoffersen's (1998) conditional coverage test. This popular performance measure applies a likelihood ratio test for the (unconditional) hit ratios, and also tests for independence of the series of indicator variables, $\{I\}_{t=1}^{T}$, denoted as 0 if a hit occurs and 1 otherwise. We can express the conditional coverage test as the sum of the two components: $LR_{cc} = LR_{uc} + LR_{ind}$, where

$$LR_{uc} = -2\ln\left[L\left(\theta; I_1, I_2, ..., I_T\right) \big/ L\left(\hat{\theta}; I_1, I_2, ..., I_T\right)\right] \overset{asy}{\sim} \chi^2(1) \tag{19}$$

and

$$LR_{ind} = -2\ln\left[L\left(\hat{\Pi}_2; I_1, I_2, ..., I_T\right) \big/ L\left(\hat{\Pi}_1; I_1, I_2, ..., I_T\right)\right] \overset{asy}{\sim} \chi^2(1). \tag{20}$$

For the test on unconditional coverage, we compare the theoretical hit probability (here 5%) to its maximum likelihood estimate, $\hat{\theta}$. The test for independence of the indicator variables compares a matrix of independent probabilities $\hat{\Pi}_2$ that measure the probability of switching from a hit state to a non-hit state (and vice versa) to a first-order Markov chain transition probability matrix $\hat{\Pi}_1$ that measures the switching probabilities depend on yesterday's state and are thus not independent. Rejection of the null hypothesis signifies that the indicator variables are not independent but tend to cluster around certain times. The test statistic LR_{cc} is then asymptotically Chi-square distributed with two degrees of freedom.

Like the Christoffersen test, the VPC is completely distribution-free: No distributional assumptions about the returns are needed, and VaR models of all kinds can be compared. The VPC does not explicitly test for conditional coverage, but it does incorporate clustering through the loss function $\sum\left(hitvalue_h - VaR_h\right)^2$. We believe this is sufficient.

One drawback of the Christoffersen test is that it relies on the number of consecutive hits. Even during crisis periods, however, returns can fluctuate strongly. Thus, negative returns may be followed by positive returns, before turning negative again. So this clustering of hits may not be consecutive, and therefore would not be detected by the conditional coverage test.

Table 6 Christoffersen test of unconditional coverage, independence and conditional coverage for S&P GSCI commodity futures indices

Sectors	Test	VaR Measures							
		Normal	CF	GARCH	Risk Metrics	Adap-tive	Asym-metric	Indirect	Sym-metric
In-sample									
Agriculture	UC	0.52	0.38	**31.21**	0.48	0.32	1.01	0.21	0.15
	IND	**107.00**	**94.90**	**117.60**	**71.10**	**91.10**	**73.10**	**70.90**	**76.00**
	CC	**107.50**	**95.20**	**148.80**	**71.50**	**91.50**	**74.10**	**71.20**	**76.10**
Energy	UC	0.86	1.52	**21.93**	2.14	0.38	0.48	1.01	0.60
	IND	**58.30**	**60.30**	**83.40**	**54.80**	**58.50**	**52.60**	**50.60**	**54.50**
	CC	**59.20**	**61.80**	**105.30**	**56.90**	**58.90**	**53.10**	**51.60**	**55.10**
Industrial Metals	UC	0.00	**9.20**	**25.41**	2.14	0.05	0.72	0.86	0.60
	IND	**163.20**	**163.00**	**197.50**	**137.70**	**167.10**	**143.90**	**146.30**	**157.60**
	CC	**163.20**	**172.00**	**222.90**	**139.80**	**167.20**	**144.60**	**147.20**	**158.20**
Livestock	UC	3.39	3.12	**17.24**	**6.70**	0.03	0.48	0.29	0.05
	IND	**98.80**	**96.60**	**121.30**	**69.10**	**86.10**	**70.90**	**82.90**	**86.50**
	CC	**102.20**	**99.70**	**138.50**	**75.80**	**86.10**	**71.40**	**83.20**	**86.60**
Precious Metals	UC	0.16	**38.10**	**40.05**	0.01	0.06	0.09	0.15	0.05
	IND	**79.70**	**82.40**	**114.70**	**68.10**	**76.20**	**60.70**	**66.30**	**61.40**
	CC	**79.80**	**120.00**	**154.70**	**68.10**	**76.30**	**60.80**	**66.40**	**61.50**
Out-of-sample									
Agriculture	UC	1.11	0.38	1.62	1.62	0.05	0.69	0.04	1.11
	IND	1.41	0.93	0.10	0.10	0.30	0.02	0.02	0.05
	CC	2.52	1.32	1.72	1.72	0.35	0.71	0.06	1.16
Energy	UC	2.99	0.69	3.85	0.04	0.38	2.25	0.04	1.11
	IND	1.20	1.85	1.06	0.57	2.03	1.35	0.57	0.05
	CC	4.19	2.54	4.91	0.61	2.41	3.59	0.61	1.16
Industrial Metals	UC	1.93	0.66	0.00	0.66	3.81	**9.18**	**6.24**	4.56
	IND	0.75	0.37	2.64	3.59	1.25	0.75	0.36	1.44
	CC	2.68	1.02	2.64	4.25	5.05	**9.93**	**6.59**	**6.00**
Livestock	UC	0.38	0.38	0.04	2.49	0.16	0.17	1.93	1.01
	IND	0.12	0.12	0.57	0.32	0.00	0.18	0.00	0.02
	CC	0.50	0.50	0.61	2.81	0.17	0.35	1.93	1.04
Precious Metals	UC	3.81	1.93	1.01	1.93	3.81	1.93	9.18	0.17
	IND	**6.94**	**6.09**	0.02	0.44	**12.66**	**5.56**	**5.22**	**6.16**
	CC	**10.75**	**8.02**	1.04	2.37	**16.47**	**7.49**	**14.40**	**6.33**

This table shows the results of the Christoffersen test for the in-sample and out-of-sample period. The Christoffersen test rejects independence for most models in the in-sample period while the opposite is true during the out-of-sample period. The test thus fails to distinguish between models that perform well during high volatility periods and inflexible models that do not perform adequately. The in-sample period ranges from 01/01/1991 to 12/28/2004 (3,652 observations), while the out-of-sample period ranges from 12/28/2004 to 11/28/2006 (500 observations). Critical values for rejection of the null hypothesis of unconditional coverage (UC), independence in the series of hits (IND), and conditional coverage (CC) are 5.02, 3.84, and 5.99, respectively (bold figures).

Table 6 gives the results of the Christoffersen test. During the out-of-sample period, few hits occurred on two consecutive days, so the results are not reliable. This might account for the finding that the null hypothesis of conditional coverage cannot be rejected for most VaR models during this period. In contrast, during the in-sample period, none of the VaR models passes the test for conditional coverage, mainly because of the clustering of hits. During this period, the conditional coverage test is therefore not useful for cross-model comparison. The VPC, however, correctly accounts for the number *and* size of (consecutive) hits via the loss function, thus providing more information than the conditional coverage test.

We conclude that the GARCH-type models (GARCH and RiskMetrics) and the three CAViaR models (Asymmetric Slope, Indirect GARCH, and Symmetric Absolute Value) are most qualified for value-at-risk modelling in commodity futures markets. However, the choice of the best model also depends on the underlying return series, so various models should be analyzed for every dataset.

Conclusion

The challenge of risk modelling is to incorporate time-varying volatility and the distribution of returns adequately, because under- or overestimating risk can lead to high losses or opportunity costs. This paper examines the in- and out-of-sample performance of various parametric and semi-parametric value-at-risk (VaR) models for commodity futures investments. The existence of significant skewness and excess kurtosis in daily commodity excess returns results in a systematic underestimation of risk when using conventional VaR or Cornish-Fisher VaR. Moreover, empirical evidence shows that GARCH-VaR and RiskMetrics, which model the evolution of conditional volatility, lead to an overall improvement, because they are more sensitive to changes in the return process.

One possible weakness of the parametric VaR models is the assumption of a specific analytical distribution of independently distributed returns. The semi-parametric CAViaR models do not depend on any distributional assumptions and may be the preferred choice when distributional assumptions of other models are likely to be violated, e.g. if the return series does not follow a normal distribution and the standardized residuals in a GARCH model are not of common distributional form. We propose an extensive performance measure in order to evaluate VaR models on a broader basis. This performance measure reveals that the GARCH-type VaR and three of the CAViaR models are the best performers, because they can 1) react sufficiently to changes in volatility and limit the risk of large negative shocks, and 2) find the best trade-off between risk and the efficient allocation of reserves. Various performance measures indicate different models as the best choice depending on the underlying return series and its length. Using only short time periods of returns for model comparison may lead to wrong conclusions concerning the best VaR model since the whole return space including rare but extreme events may not be covered. Hence, risk managers should compare VaR measures for every return process over short and long time horizons to find the most adequate value-at-risk model for security portfolios that include commodity futures investments.

References

Anson, M. (2004) *Managing Downside Risk in Return Distributions using Hedge Funds, Managed Futures and Commodity Futures*, The CTA Reader.

Bao, Y., Lee, T.-H. and Saltoğlu, B (2006) 'Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check', *Journal of Forecasting*, 25, 101–128.

Bessembinder, H., Coughenour, J. F., Seguin, P. J. and Smoller, M. M. (1996) 'Is there a Term Structure of Futures Volatilities? Reevaluating the Samuelson Hypothesis', *Journal of Derivatives*, 4, 45–58.

Bloomfield, P. and Steiger, W. L. (1983) *Least Absolute Deviations: Theory, Applications and Algorithms*, Birkhauser.

Bollerslev, T., 'Generalized Autoregressive Conditional Heteroscedasticity', *Journal of Econometrics*, Vol. 31, 1986, pp. 307–327.

Brooks, C., Clare, A. D., Dalle Molle, J. W. and Persand, G. (2005) 'A Comparison of Extreme Value Theory Approaches for Determining Value at Risk', *Journal of Empirical Finance*, 12, 339–352.

Christoffersen, P. (1998) 'Evaluating Interval Forecasts', *International Economic Review*, 39, 841–862.

Cotter, J. (2005) 'Extreme Risk in Futures Contracts', *Applied Economics Letters*, 12, 489–492.

Danielson, J., and De Vries, C. G. (2000) 'Value-at-Risk and Extreme Returns', *Annales D'Économie et de Statistique*, 60, 239-270.

Engle, R. F. (1982) 'Autoregressive Conditional Heteroscedasticity 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', *Journal of Business & Economic Statistics*, 22, 367–381.

Erb, C. B. and Harvey, C. R. (2006) 'The Tactical and Strategic Value of Commodity Futures', *Financial Analysts Journal*, 62, 69–97.

Favre, L. and Galeano, J. (2002) 'Mean-Modified Value-at-Risk Optimization with Hedge Funds', *Journal of Alternative Investments*, 5, 21–25.

Galloway, T. M. and Kolb, R. W. (1996) 'Futures Prices and the Maturity Effect', *Journal of Futures Markets*, 16(7), 809–828.

Giot, P. and Laurent, S. (2003) 'Market Risk in Commodity Markets: A VaR Approach', *Energy Economics*, 25, 435–457.

Jensen, G. R., Johnson, R. R. and Mercer, J. M. (2000) 'Efficient Use of Commodity Futures in Diversified Portfolios', *Journal of Futures Markets*, 20, 489–506.

Jensen, G. R., Johnson, R. R. and Mercer, J. M. (2002) 'Tactical Asset Allocation and Commodity Futures', *Journal of Portfolio Management*, 28, 100–111.

Jorion, P. (2007) *Value at Risk. The New Benchmark for Managing Financial Risk*, Graw-Hill.

Kat, H. M. and Oomen, R. C. A. (2006) 'What Every Investor should know about Commodities, Part I: Univariate Return Analysis', *Journal of Investment Management*, 5, 1–25.

Koenker, R. and Bassett, G. (1978) 'Regression Quantiles', *Econometrica*, 46, 33–50.

Koenker, R. and Bassett, G. (1982) 'Robust Tests for Heteroskedasticity based on Regression Quantiles', *Econometrica*, 50, 43–61.

Kuester, K., Mittnik. S. and Paollella, M. S. (2006) 'Vale-at-risk Prediction: A Comparison of Alternative Strategies', *Journal of Financial Econometrics*, 4, 53–89.

Ljung, G. M. and Box, G. E. P. (1978) 'On a Measure of Lack of Fit in Time Series Models', *Biometrika*, 65, 297–303.

Manganelli, S. and Engle, R. F. (2001) 'Value at Risk Models in Finance', *Working Paper*, No. 75, European Central Bank.

RiskMetrics Group (1996) *RiskMetrics – Technical Document*,,J.P. Morgan.

Samuelson, P. A. (1965) 'Proof that Properly Anticipated Prices Fluctuate Randomly', *Industrial Management Review*, 6, 41–49.

Till, H. (2006) 'Portfolio Risk Measurement in Commodity Futures Investments', in T. P. Ryan, ed., *Portfolio Analysis: Advanced Topics in Performance Measurement, Risk and Attribution*, Risk Books, 243-251.

Urzua, C. (1996) ‘On the Correct Use of Omnibus Tests for Normality’, *Economics Letters*, 53, 247–251.
White, H. (1994) *Estimation, Inference and Specification Analysis*, Cambridge University Press.
White, H. (2000) ‘A Reality Check for Data Snooping’, *Econometrica*, 68, 1097–1126.