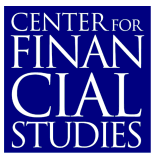

No. 2003/32

Evaluating VaR Forecasts under Stress – The German Experience

Stefan Jaschke, Gerhard Stahl, Richard Stehle

Center for Financial Studies
an der Johann Wolfgang Goethe-Universität ▪ Taunusanlage 6 ▪ D-60329 Frankfurt am Main
Tel: (+49)069/242941-0 ▪ Fax: (+49)069/242941-77 ▪ E-Mail: ifk@ifk-cfs.de ▪ Internet: http://www.ifk-cfs.de

CFS Working Paper No. 2003/32

Evaluating VaR Forecasts under Stress – The German Experience*

Stefan Jaschke[1], Gerhard Stahl[1], Richard Stehle[2]

October 2003

Abstract:

We present an analysis of VaR forecasts and P&L-series of all 13 German banks that used internal models for regulatory purposes in the year 2001. To this end, we introduce the notion of well-behaved forecast systems. Furthermore, we provide a series of statistical tools to perform our analyses. The results shed light on the forecast quality of VaR models of the individual banks, the regulator's portfolio as a whole, and the main ingredients of the computation of the regulatory capital required by the Basel rules.

JEL Classification: K23, G28

Keywords: banking supervision, VaR, exploratory data analysis, backtesting

*The first two authors point out that the views expressed herein should not be construed as being endorsed by the BaFin.

[1] Bundesanstalt für Finanzdienstleistungsaufsicht, Graurheindorfer Str. 108, 53117 Bonn

[2] Lehrstuhl für Bank- und Börsenwesen, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin

1. Introduction

In an important regulatory innovation the Basel Committee on Banking Supervision has allowed for banks to use their own internal models – so-called Value-at-Risk (VaR) models – for calculating the regulatory capital cushion needed to cover the market risk of open positions in an institute's trading book. Compared with the standardized methods the internal models approach offers a whole bunch of advantages within the process of risk management, i.e., in measuring, monitoring and managing market risk for trading portfolios. These advantages include the convergence of economic and regulatory capital (Matten; 2000), the avoidance of duplicated efforts for internal and regulatory risk measurement and, among others, the signaling of competence to the market, especially to rating agencies, by regulatory approval of an internal model.

Looking back on the year 2001 we recognize that a number of severe events took place in the financial markets. Not only the terrorist attack from September 11, but also a significant number of interest rate decisions of central banks, the Enron case, the continuing demise of the "new economy", and others, took place. Obviously, most of the events mentioned were unpredictable but dramatic in some sense. That the year 2001 was special is also highlighted by the fact that 13 German banks produced 33 VaR exceptions in the year 2001 while 14 German banks produced 17 VaR exceptions in 2002. Given the turbulences of that year, it is natural to ask whether the forecast quality of VaR models over the year 2001 turned out to be satisfactory. A second question is whether the portfolios of the supervised banks behave similarly and whether the similarity increases in stress periods. Put in another way, how well is the supervisor's portfolio diversified and is this affected by the special events of the year 2001? This is the first article to provide a detailed empirical analysis of (1) the performance of the actual VaR forecasts of all German banks that used internal models for regulatory purposes in 2001 and (2) the interrelations among the profits and losses (P&L) of these banks, especially during the stress periods. Insofar, it is comparable to the empirical analysis of similar data from six US banks by Berkowitz and O'Brien (2002). The methodology employed, however, differs significantly.

The Basel paper on backtesting describes in-depth the regulatory requirements on the forecast quality of the trading book as a whole in order to ensure an adequate calculation of regulatory capital (Basel Committee on Banking Supervision; 1996b). Numerous publications reflect VaR forecast evaluation, starting with Kupiec (1995), who points out the lack of statistical power of a backtesting that is based on a binomial test statistic. Crnkovic and Drachman (1996) propose the use of the Kuiper statistic, which is a goodness-of-fit type statistic based on the whole forecast distribution. Christoffersen (1998) draws the attention to backtesting based on specification tests, focussing on the independence property to evaluate forecast quality. Inspired by the ideas of Rosenblatt (1952), Berkowitz (2000) proposed an approach that relies on conditioned forecast distributions. This is very much in the spirit of the papers of Dawid, see (Dawid; 1982a,b, 1984, 1986; Seillier-Moiseiwitsch and Dawid; 1993). This again is influenced by the literature on wheather forecasting (Murphy and Winkler; 1987, 1992). A detailed overview of backtesting issues is given by Overbeck and Stahl (2000). The empirical analysis of the forecast quality in this paper is in the spirit of Dawid's forecast evaluation. The fact that the whole forecast distribution is not available is bridged by the empirical observation

that the ratio of P&L over VaR is surprisingly close to normal, in a certain sense. The rigorous justification of our approach is given in the appendix. The methodology used to analyze the joint behavior of all banks includes a measure of co-movement and several stress variables used to define stress periods. Finally, we consider how distributional parameters change conditional on stress.

Before describing the data set, let us introduce some notation. We denote by $\mathrm{VaR}(\pi_t, \alpha)$ the *Value-at-Risk* of a portfolio π_t, given the level of significance α. Within the framework of the Basel Committee, α is set to 99%. The VaR number $\mathrm{VaR}(\pi_t, \alpha)$ at time t denotes the $\alpha-$quantile of the distribution of the hypothetical, or clean, P&L

$$C_t = v_{t+1}(\pi_t) - v_t(\pi_t), \tag{1}$$

where $v_t(\pi)$ is the value of a given portfolio π at time t. The random variable $v_{t+1}(\pi_t)$ denotes the at time t frozen portfolio, evaluated at prices of time $t+1$. The VaR is interpreted as an upper bound of losses that might be surpassed only with probability $1-\alpha$. In the sequel we will use the shorthand notation V_t to denote $\mathrm{VaR}(\pi_t, 99\%)$.

Within an observation period of 250 trading days 2.5 violations of the forecasts V_t are to be expected on the average. Hence, if too many violations occur there is good reason to doubt that the internal model's level of significance is correctly covered. In order to ensure a sufficient forecast quality, the Basel Committee tied the amount of the capital requirements to the number of VaR exceptions of the model (Basel Committee on Banking Supervision; 1996a).

It is of great practical importance to note that freezing the portfolio – as stated in (1) – is not imperative. The Basel Committee also acknowledges backtesting VaR that is based on actual trading outcomes. In that case changes of the portfolio composition during the holding period, fees etc. are superimposed on the hypothetical P&L. From a statistician's point of view, the judgment of forecast quality should be based on the hypothetical P&L, (1), because the VaR forecast assumes a static portfolio by construction. From a risk managers point of view, however, the actual or economic P&L is actively managed and reported. Obviously both ways to tackle the backtesting problem have their intrinsic merits. German legislation prescribes a backtesting based on (1), whereas the US legislation admits to base the backtesting on the economic P&L, see (Berkowitz and O'Brien; 2002).

2. Description of the Data Set

The data set considered here contains data from all thirteen German banks that used internal models for regulatory purposes in the year 2001. The data set for each bank consists of first, VaR forecasts and second, the so-called clean, or hypothetical P&L for all 253 trading days of the year 2001. Most of the following figures and tables are based on the normalized P&L and VaR time series. I.e., they are divided by the banks' full sample standard deviation of P&L to insure confidentiality.

Table 1 shows summary statistics of each bank's data. The coefficient of variation (i.e., the ratio of the standard deviation and the mean) of VaR in column 5 shows that for the majority of the banks the variability of the VaR is relatively small compared to its mean. Only banks A, E, and I have a coefficient of variation that is two times larger than the average coefficient of variation of all banks. The last column reports the average loss

bank name	kurtosis	skewness	99%-quantile of losses	average VaR	coefficient of variation of VaR	average of the loss exceeding VaR
A	22.51	3.15	1.83	2.68	0.84	0.00
B	6.48	−0.00	2.62	4.51	0.29	1.34
C	5.98	0.61	2.18	4.05	0.14	0.00
D	5.58	0.19	2.55	3.29	0.26	0.00
E	7.79	0.43	2.56	1.69	0.87	0.56
F	16.99	−2.10	3.52	2.11	0.52	1.63
G	3.71	−0.20	2.42	2.15	0.23	0.24
H	5.72	0.01	2.12	2.25	0.23	1.49
I	15.03	−1.58	4.55	1.07	0.87	0.73
J	4.04	−0.67	2.98	3.03	0.39	0.37
K	5.28	0.00	2.24	2.20	0.11	1.37
L	4.57	−0.22	2.75	3.85	0.20	0.00
M	3.35	−0.12	2.45	2.58	0.18	0.00

Table 1: **Summary statistics of standardized data.** The columns give kurtosis, skewness, minus the 1%-quantile of the P&L, the average VaR, the coefficient of variation of the VaR, and the average size of the loss exceeding VaR. Both P&L and VaR are re-normalized in such a way that the standard deviation of the P&L is 1 for all banks.

exceeding VaR, i.e., the estimate of the expected shortfall $E[-C_t - V_t | - C_t > V_t]$. Note, that for the standard normal distribution, the expected shortfall is approximately 0.34 for the 99% confidence level. The comparatively large average losses exceeding VaR indicate the presence of outliers. Three out of the thirteen banks had more than four violations and only four had no violations at all. For reasons of confidentiality, the individual numbers of violations are not reported here.

The six histograms in figure 1 are representative for the dataset. The histograms are fairly symmetric, which is in line with the skewness in table 1. Taking into account that the data are standardized, the large range of the x-axis for bank I shows the presence of extreme outliers. This suggests the use of robust estimates of location and scale. To be specific, we use the median $F^{-1}_{0.5}$ as an estimate for the location parameter and the interquartile range $F^{-1}_{0.75} - F^{-1}_{0.25}$ as an estimate of the scale parameter. In figure 1 the fitted normal densities use the median as a robust estimator of the mean and normalized interquartile range $(F^{-1}_{0.75} - F^{-1}_{0.25})/(\Phi^{-1}_{0.75} - \Phi^{-1}_{0.25})$ is a robust estimator of the standard deviation. ($\Phi^{-1}_{0.75} - \Phi^{-1}_{0.25} \approx 1.349$ is the interquartile range of the standard normal distribution.) This improves the representation of the bulk of the data by a Gaussian distribution. Furthermore, it is evident that the large excess kurtosis and skewness of bank I is caused by a few extreme outliers relative to the normal distribution.

Figure 2 shows the time series of P&L and −VaR of the selected institutes. As can be seen from banks B and F, for example, it is not reasonable to assume stationarity of the P&L and VaR time series. Hence, the summary statistics given in table 1 as well as the histograms in figure 1 are to be interpreted with care. Banks B, D, and F give a picture of the conservativeness of the VaR forecasts as mentioned in the interpretation of table 1.

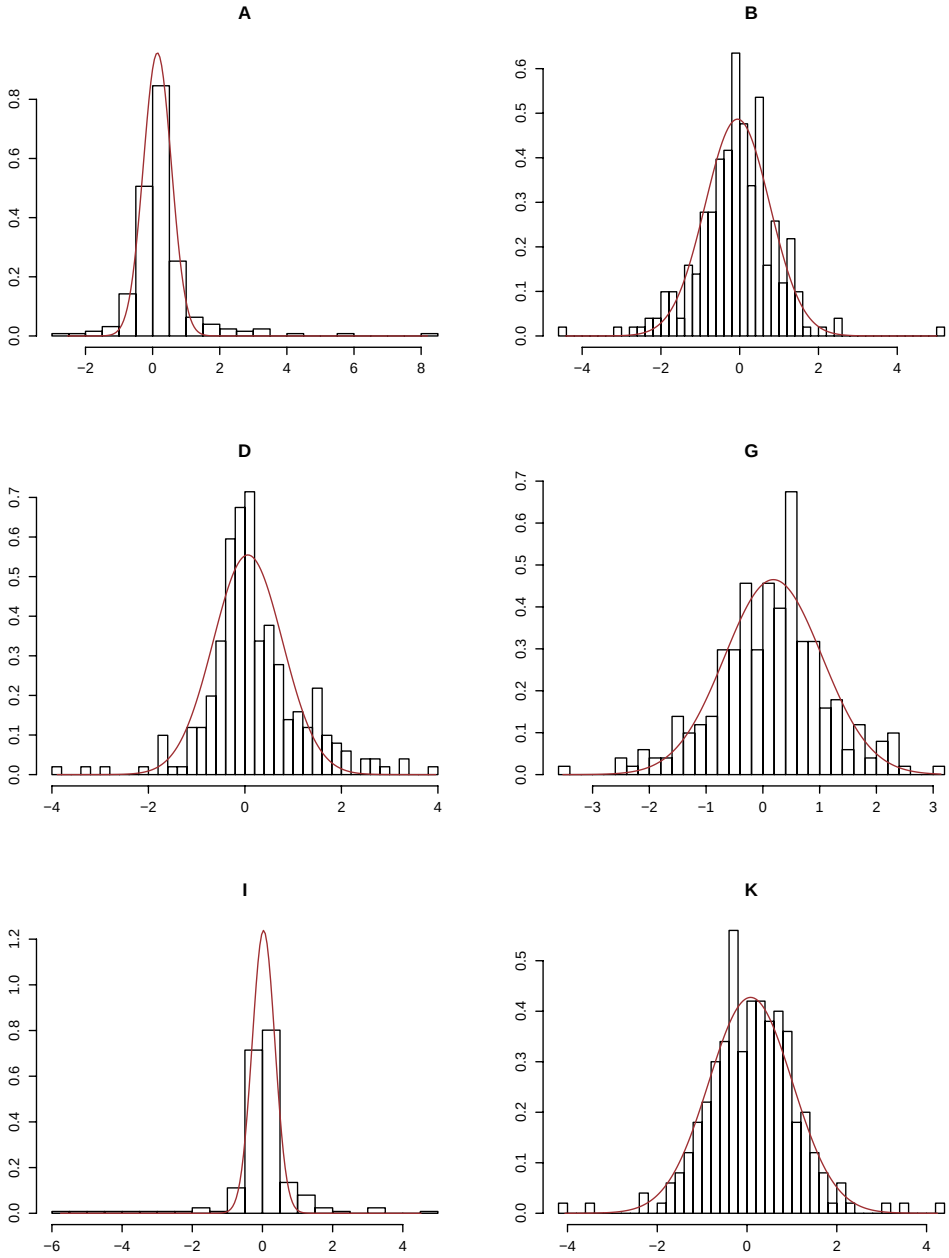

Figure 1: **Histograms of the P&L and fitted normal densities.** The mean and standard deviation of the fitted normal density are estimated robustly by the median and the (scaled) interquartile range.

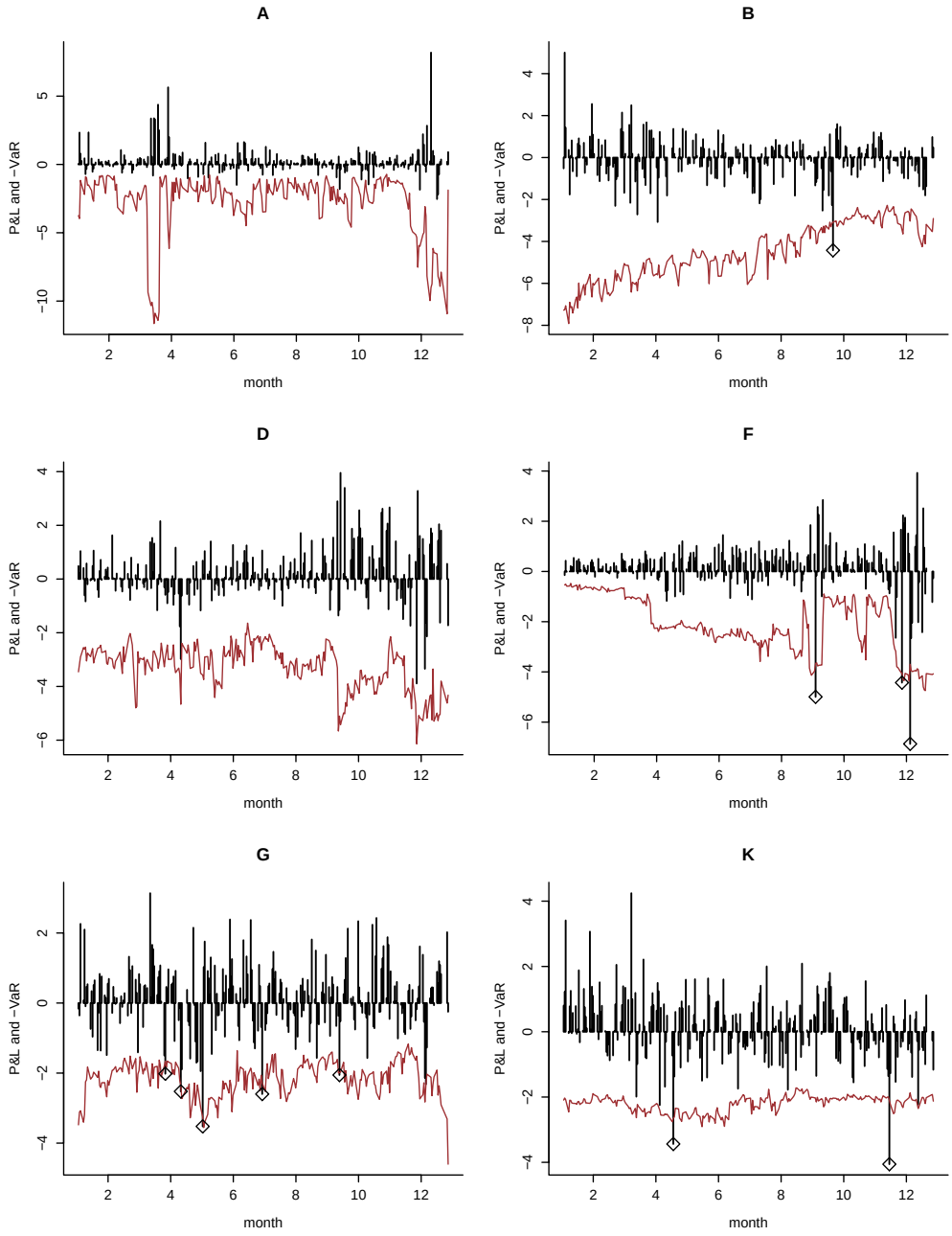

Figure 2: **Time series of P&L and -VaR.** The vertical dotted lines denote the date 2001-09-11. Diamonds are added to mark the violations.

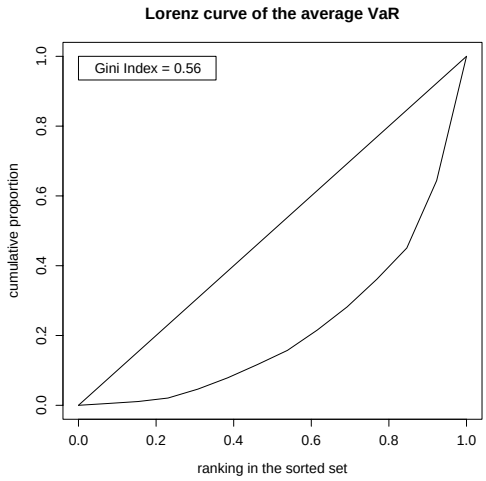

Figure 3: **Lorenz curve of the average VaR of all banks.** The largest three banks in terms of average VaR capture 61% of the aggregated average VaRs.

As the descriptive statistics show, banks tend to be conservative in the sense that they overestimate their VaR. The traffic light approach is related to a one-sided loss function, as reported in table 1. Hence, this backtest does not recover the information in the data about forecast quality of a VaR-model as a whole. In the next section we will have a closer look at forecast quality using more powerful tools.

We conclude this section with the Lorenz curve in figure 3, which depicts the concentration of risks among the thirteen banks and gives some complementary information about the relative "size" of the banks.

3. Analyzing VaR-Forecasts for Each Bank Individually

In their paper on backtesting, the Basel Committee on Banking Supervision (1996b) encourages institutes to apply backtesting procedures beyond the so-called traffic light approach. In this section, we make proposals for possible refinements. The proposed tools have been used by the *Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin)*, the German single regulator for integrated financial services supervision, for a couple of years.

In the following, we use the term *prediction-realization pair* for the class of objects $(\theta(F_t), C_t)$, where F_t denotes the whole forecast distribution and $\theta(F_t)$ a parameter thereof, e.g., a quantile, or the distribution as a whole ($\theta = \mathrm{id}$). The summary statistics from section 2 give a description of the prediction-realization pairs (V_t, C_t). In order to evaluate the forecast quality of VaR models we introduce a statistical model that allows to incorporate additional information from the prediction-realization pairs (F_t, C_t).

The time plots displayed in figure 2 show that neither of the time series C_t nor V_t are stationary. For the evaluation of a forecast model it is natural to look for transformations T such that the time series $T(\theta(F_t), V_t)$ is nearly stationary. In the literature on forecast

evaluation the transformation

$$T(F_t, C_t) = F_t(C_t) \tag{2}$$

is used as a starting point. A forecast system is considered "good" (Dawid; 1984, p.281), if the values $F_t(C_t)$ are independent and uniformly distributed, i.e., $F_t(C_t)$ iid $U[0,1]$.

This corresponds to the concepts "well-calibration" and "refinement" (= "resolution") as used in the literature on weather forecasting (Murphy and Winkler; 1987, 1992) insofar as the condition $F_t(C_t) \sim U[0,1]$ essentially means "well-calibration" and the temporal independence of $F_t(C_t)$ is related to "refinement" (Dawid; 1986; Seillier-Moiseiwitsch; 1993).

While it has been proposed (Berkowitz; 2000) that banks should report the realized probabilities $F_t(C_t)$ to the supervisory authorities, the current rules only require the reporting of C_t and $V_t = -F^{-1}(0.01)$. The forecast evaluation (backtesting) as laid down in the Basel Amendment is defined in terms of the VaR-exceptions

$$T(V_t, C_t) = \mathbf{1}_{\{V_t \geq C_t\}}. \tag{3}$$

The drawbacks of this approach are discussed by Kupiec (1995). See also Lopez (1999) and (Jorion; 2001, chapter 6).

In this section, we suggest an approach that bridges the gap between the information presented by the prediction-realization pairs (2) and (3). In the delta-normal RiskMetrics framework, the VaR-forecast V_t is related to the forecast of the standard deviation of the P&L, σ_t, as

$$V_t = z\,\sigma_t, \tag{4}$$

where z is the standard normal 99%-quantile, $\Phi(z) = 0.99$. There, the *standardized returns* are defined as

$$R_t := T(C_t, V_t) = z\,C_t/V_t, \tag{5}$$

(Longerstaey; 1996, chapter 11) and are iid standard normal under the specific model assumptions of the delta-normal approach. This definition (5) of standardized returns is useful even if (4) is invalid. A rigorous justification of this statement and the following approach is provided in the appendix. The standardized returns provide the basis of the forecast evaluation in this paper.

We call a VaR forecast system *well-behaved*[3] if

$$R_t \text{ iid } F \in \mathcal{U}_\varepsilon^G,$$

where $\mathcal{U}_\varepsilon$ is the ε-neighborhood of the set of Gaussian distributions with respect to the Kolmogorov distance in the set of probability distributions:

$$\mathcal{U}_\varepsilon^G := \{F \mid \sup_x |F(x) - \Phi(\mu, \sigma^2)(x)| \leq \varepsilon, F \text{ is pdf}, \mu \in \mathbb{R}, \sigma^2 \in (0, \infty)\} \tag{6}$$

[3]Note that "well-behaved" is not to be interpreted as "favorable from a supervisory point of view". It rather stands for "tractable from a statistical point of view".

for some fixed, small probability ε, say 5%. Note that the semiparametric model, known as Huber's gross error model in the literature of robust statistics,

$$\mathcal{F}_\varepsilon^G := \{F \mid F = (1-\varepsilon)\Phi(\mu,\sigma^2) + \varepsilon H, \mu \in \mathbb{R}, \sigma^2 \in (0,\infty)\}, \tag{7}$$

where H is an arbitrary probability distribution, is an important subset of $\mathcal{U}_\varepsilon^G$. In fact, if $F \in \mathcal{F}_\varepsilon^G$, then exist μ, σ, and H such that $F = (1-\varepsilon)\Phi(\mu,\sigma^2) + \varepsilon H$. Then the Kolmogorov distance between the distributions F and $\Phi(\mu,\sigma^2)$ is

$$\begin{aligned} d(F,\Phi(\mu,\sigma^2)) &= \sup_{x\in\mathbb{R}} |F(x) - \Phi(\mu,\sigma^2)(x)|, \\ &= \varepsilon \sup_{x\in\mathbb{R}} |H(x) - \Phi(\mu,\sigma^2)(x)| \\ &= \varepsilon\, d(H,\Phi(\mu,\sigma^2)) \le \varepsilon. \end{aligned}$$

In terms of the P-P plot of F against normal, which plots $F(x)$ against $\Phi(\mu,\sigma^2)(x)$ for $x \in \mathbb{R}$, this means that the graph is (vertically) within an ε-band of the diagonal. On the other hand, any point within that band can be reached by the graph of a distribution $F \in \mathcal{U}_\varepsilon^G$.

The Gaussian distribution with robustly estimated parameters fits (by eye) remarkably well to the center of standardized returns R_t, see figure 4. The P-P plot against normal (figure 5) also shows that the distributions of the standardized returns are relatively close to normal, in the sense that they are in the class $\mathcal{U}_\varepsilon^G$ with $\varepsilon = 0.05$. (In fact, the empirical distributions of all banks are in this class.)

The P-P plot (and the Kolmogorov distance) does not detect *large* deviations from normality as long as they happen with *small* probabilities. A better picture of the fit in the tails is provided by the Q-Q plot of F against normal, which plots $F^{-1}(t)$ against $\Phi^{-1}(\mu,\sigma^2)(\alpha)$ for $\alpha \in (0,1)$, see figure 6. Another display with emphasis on the tails is the boxplot, see figure 7.

Diagnostics that illustrate serial (in-) dependence of R_t are provided by the cumulative periodogram applied to both standardized returns and their absolute values. As expected, figure 8 shows that there is no significant autocorrelation in the standardized return series. Unlike the absolute returns of many financial securities, the absolute standardized returns of the banking books show no marked heteroscedasticity, see figure 9. Possibly reasons are that the banks' VaR-forecasts successfully predict some of the heteroscedasticity of their P&L-series. Another possible reason could be that the risk control limit system filters out some of the heteroscedasticity of the original returns.

The set of well-behaved VaR-forecasts contains forecasts that grossly over- or underestimate the respective quantile. Under the assumption that standardized returns are i.i.d. standard normal, the inverse of the standard deviation of R_t can be interpreted as the *recalibration factor*. The larger it is, the more the bank overestimates its VaR. We will call a forecast well-calibrated, if the recalibration factor equals one[4]. Instead of looking only at the empirical standard deviation of the standardized return R_t, one can alternatively

[4]The relation between this notion of "well-calibration" and its definition in the forecast evaluation literature can be made rigorous in more general settings than the delta-normal method, which is explained in appendix A.

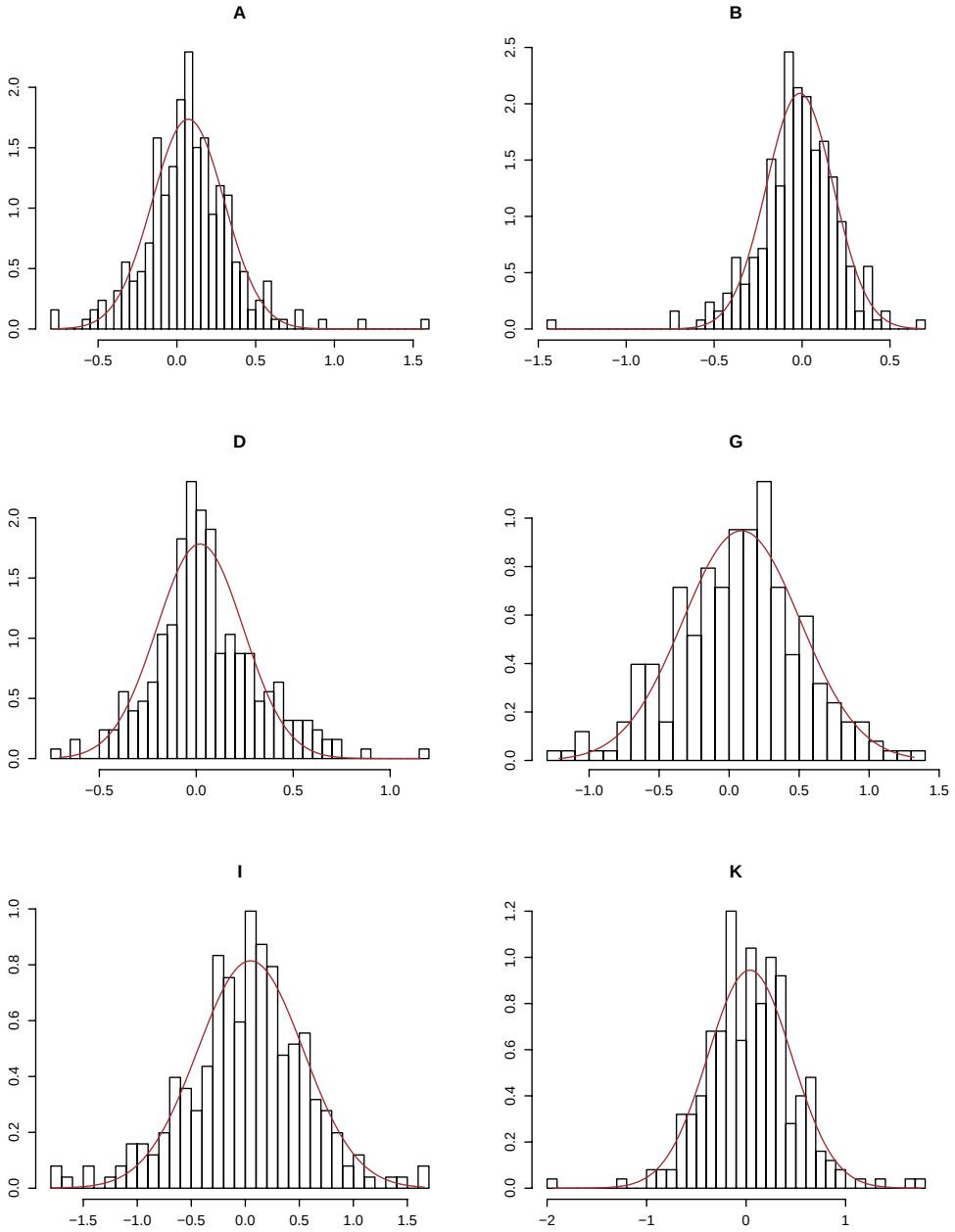

Figure 4: **Histograms of the standardized return and fitted normal density.** The mean and standard deviation of the fitted normal density are estimated robustly by the median and the (scaled) interquartile range.

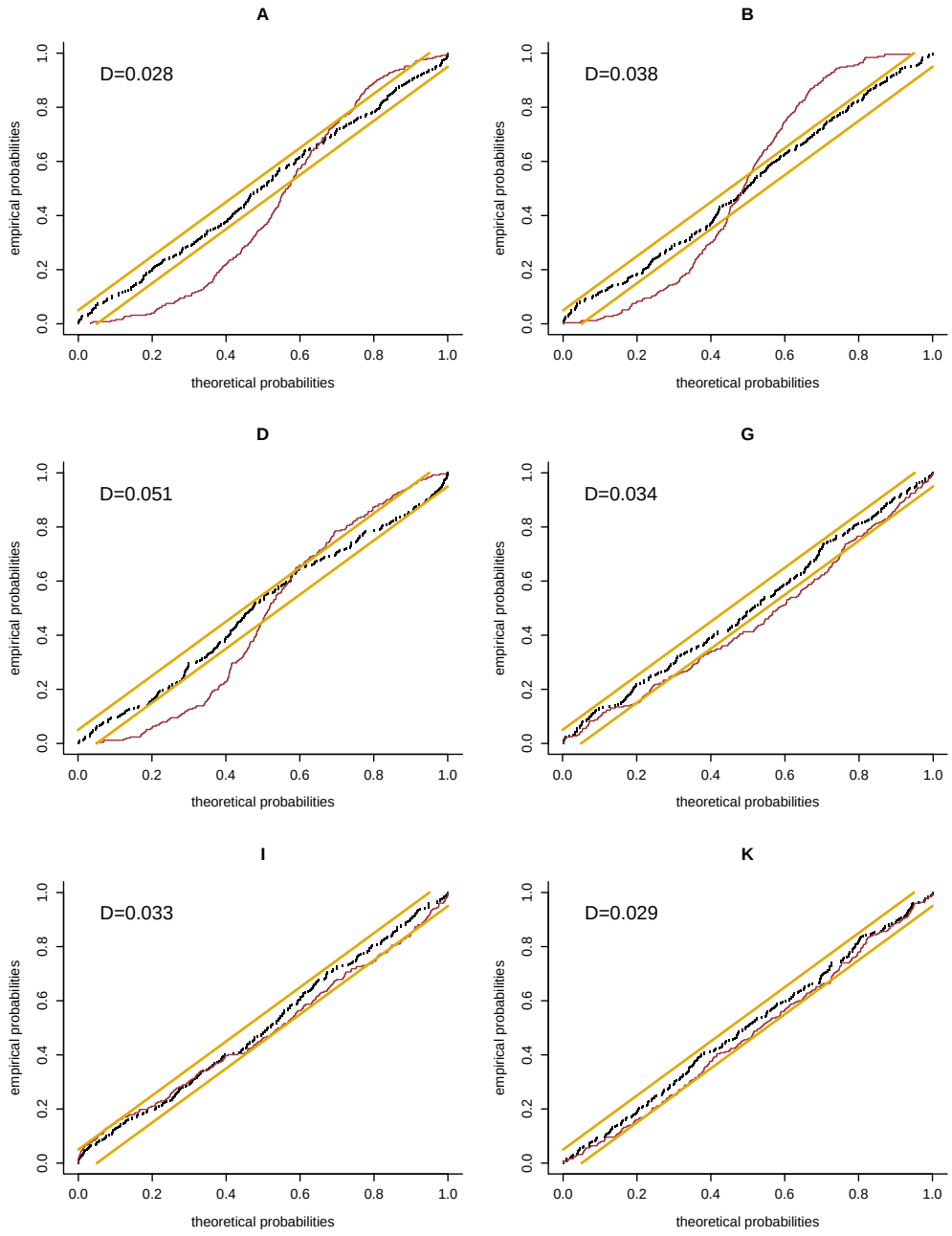

Figure 5: **P-P plot of the distribution of the standardized return against a standard normal (dotted line) and a fitted normal distribution (fat dotted line).** The mean and standard deviation of the fitted normal distribution are chosen to minimize the Kolmogorov distance to the empirical distribution function. D is that minimal distance. The dashed lines are the envelope of all distributions in the 5%-neighborhood with respect to the Kolmogorov distance.

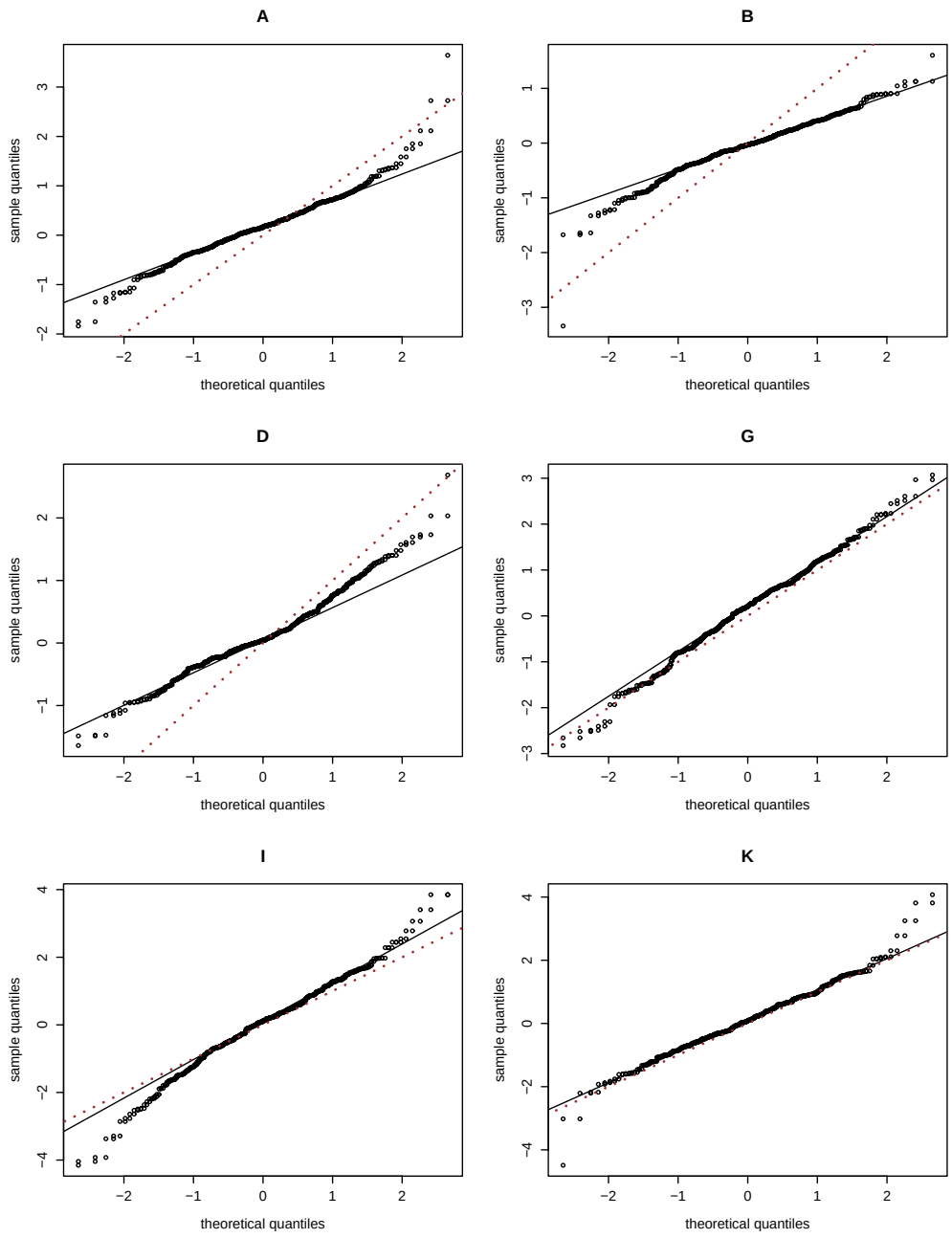

Figure 6: **Q-Q plot of the standardized return.** The dotted line denotes the standard normal distribution, the solid line denotes the normal distribution with mean and variance estimated from the sample, and the small circles represent the empirical distribution.

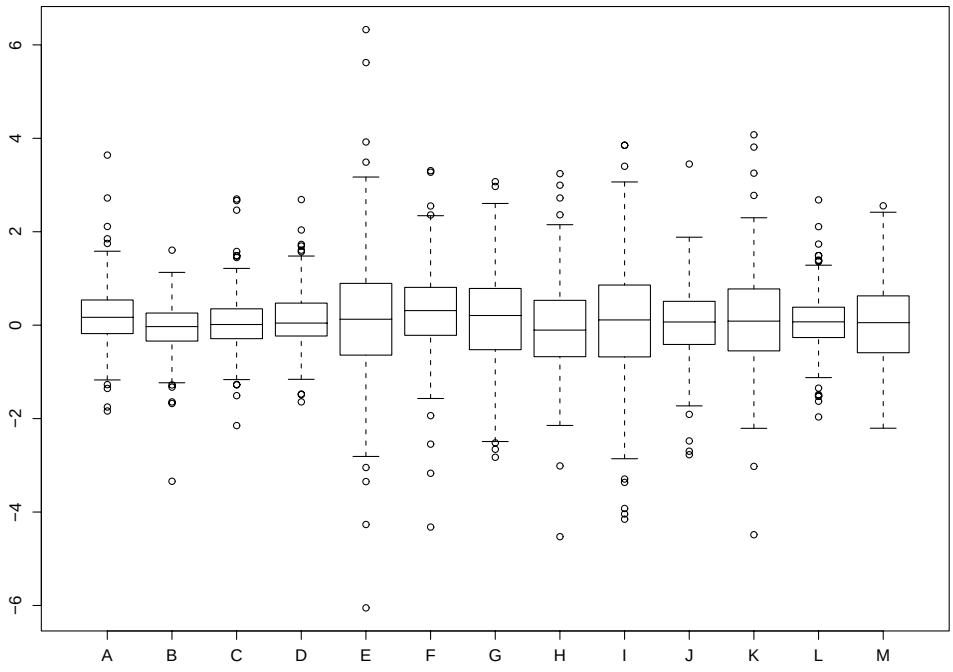

Figure 7: **Boxplot of the standardized return.** The box shows the interquartile range and the middle line the median. The "whiskers" are drawn 1.5 times the interquartile range away from the box. Any point further away is shown individually.

look at different estimates

$$\hat{\sigma}_p(R) := \left(\frac{1}{T} \sum_{t=1}^{T} |R_t|^p \right)^{1/p} / c_p \tag{8}$$

of the standard deviation of R_t for powers $p > 0$. (c_p is the constant $c_p = (E|X|^p)^{1/p}$ with X being standard normal.) Using smaller powers $p < 2$ gives more robust estimates than the usual estimate of the standard deviation. Using larger powers $p > 2$ gives estimates of the standard deviation that are more sensitive, i.e., more dependent on the extreme values of R_t, than the usual estimate of the standard deviation ($p = 2$). By looking at $\hat{\sigma}_p(r)$ for different powers p, one can observe whether banks are well-calibrated at the center or at the tails, respectively.[5]

Table 2 shows estimates of the recalibration factor, as well as p-values from testing the hypothesis that this factor equals 1. These results are further illustrated by figure 10,

[5]Note that the various different $\hat{\sigma}_p(r)$ are estimates of the standard deviation of R_t only if the R_t are (approximately) i.i.d. standard normal. Otherwise, (8) provides estimates of scaled versions of *different* moments of $|R_t|$.

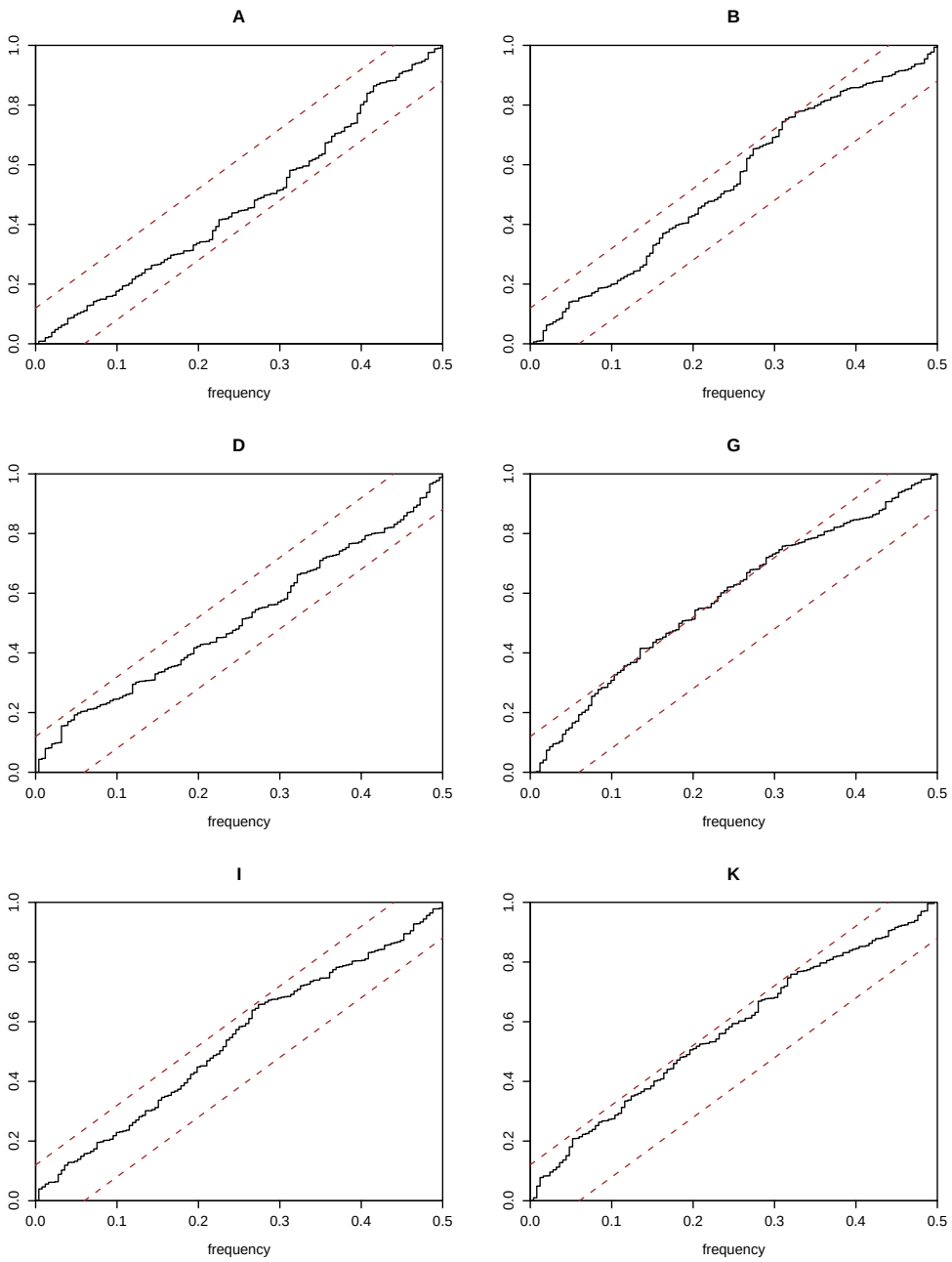

Figure 8: **Cumulative periodogram of standardized returns.** White noise is characterized by equal probabilities for all frequencies, which shows as a straight line in the cumulative periodogram. 95%-confidence bands are given for the null hypothesis of Gaussian white noise.

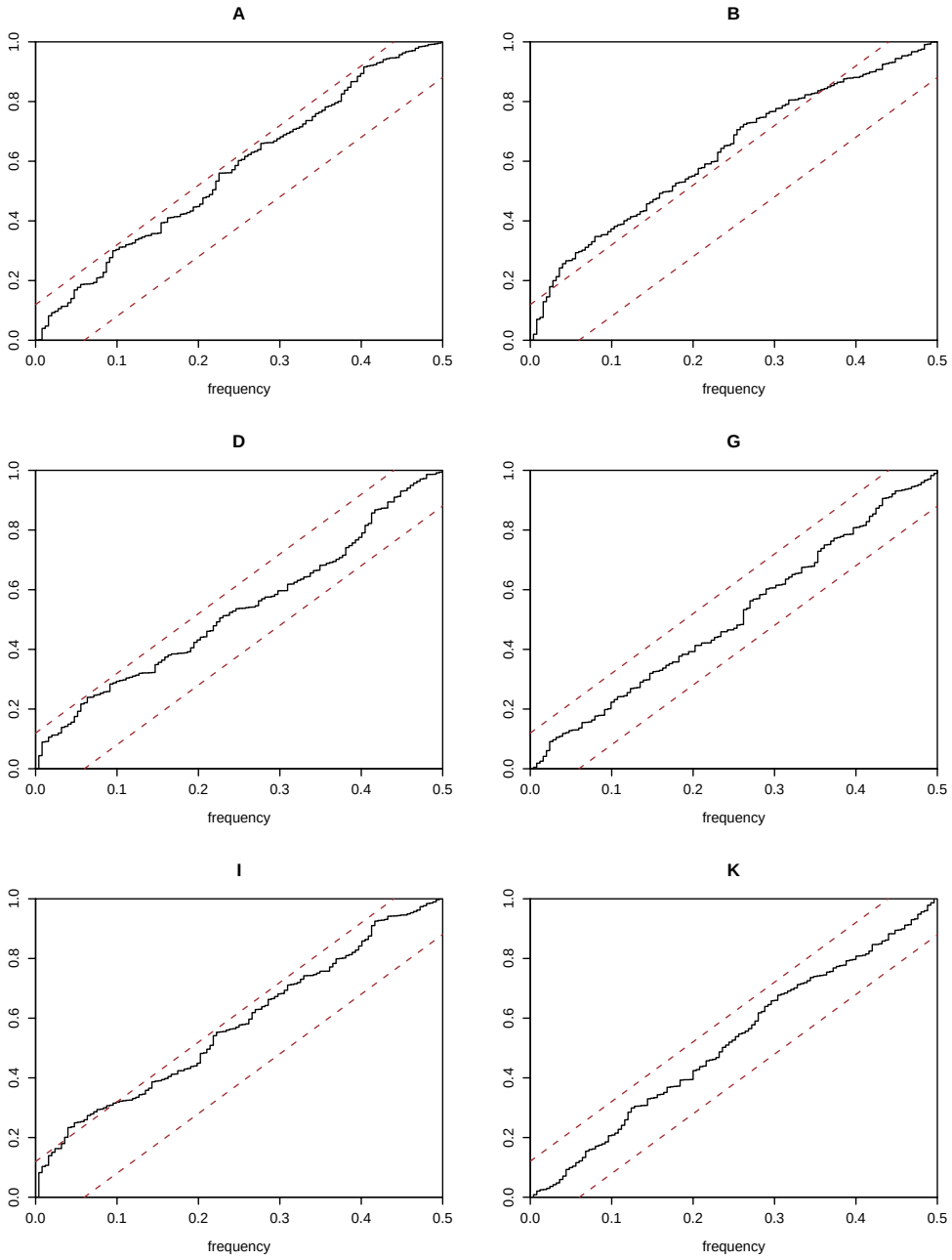

Figure 9: **Cumulative periodogram of absolute values of standardized returns.**

	p=0.5	p=1	p=2
A	1.670^{**}	1.614^{**}	1.480^{**}
B	2.162^{**}	2.053^{**}	1.849^{**}
C	1.899^{**}	1.832^{**}	1.659^{**}
D	1.805^{**}	1.690^{**}	1.552^{**}
E	0.825^{**}	0.786^{**}	0.710^{**}
F	1.147^{*}	1.129^{*}	1.055
G	0.938	0.946	0.940
H	1.074	1.037	0.980
I	0.831^{**}	0.808^{**}	0.769^{**}
J	1.466^{**}	1.403^{**}	1.296^{**}
K	1.004	0.992	0.944
L	1.838^{**}	1.765^{**}	1.625^{**}
M	1.143^{*}	1.110^{*}	1.090

Table 2: **Robust and sensitive estimates of the recalibration factor.** The estimates $1/\hat{\sigma}_p(r)$ (see equation (8)) of the recalibration factor are computed for powers $p \in \{0.5, 1, 2\}$. p-values for the corresponding null hypothesis that the recalibration factor is one (i.e., that the R_t are iid standard normal) are computed by Monte-Carlo simulation. Significance at the 5%-level is shown by one star and significance at the 1%-level with two stars.

showing the recalibration factor estimated with the empirical standard deviation ($p = 2$) and the interquartile range, respectively. Note that banks E and I are consistently underestimating its VaR. Also note that all banks are close to the line with slope 0.84 and intercept 0. This means first, that the distributions of their standardized returns have slightly heavier tails than the normal distribution. Second, the estimate of the recalibration factor depends only moderately on how robustly it is estimated.

The local averages of the absolute values of the standardized returns (= estimates of the inverse of the recalibration factor using $p = 1$) in figure 11 show only moderate variation in time, except for some unpredicted absolute returns of banks B and I.

In summary, the standardized return series of all trading books are remarkably close to being "well-behaved" as defined in this section. I.e., they are close to normal in the sense of the Kolmogorov distance and there are no obvious temporal dependencies. This is in stark contrast to the conclusions drawn by Berkowitz and O'Brien (2002). They show that the VaR-forecasts of the six US banks considered are inferior to that of a simple "reduced-form" model, i.e., a univariate time series model applied directly to the individual P&L series. Their conclusion is that the banks' forecasts "did not adequately reflect changes in P&L volatility". The conclusion from the empirical evidence presented in this section, however, is that the VaR-forecasts of the considered German banks are "essentially OK": the conservative VaR-forecasts of most banks can be corrected by simply re-calibrating them. Much of the non-normality of the standardized return distributions in the view of the tail-sensitive graphs (figure 6) can be traced back to the essentially unpredictable events in September 2001 (figure 11).

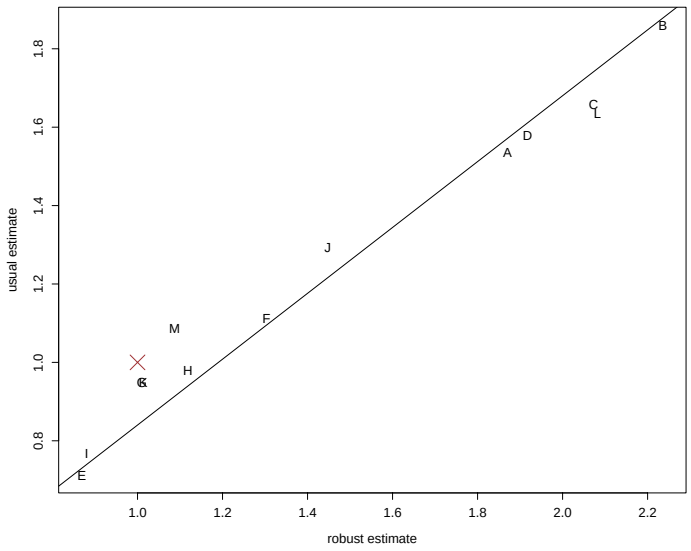

Figure 10: **Sensitive and robust estimates of the recalibration factor, estimated with the usual estimate of the standard deviation ($p = 2$) and the interquartile range.** Banks towards north east overestimate their VaR. Banks towards south east have outliers or heavy tails. The black line represents the best linear fit with intercept 0. Its slope is 0.84.

4. Analyzing P&L Series For All Banks Simultaneously

The analysis of the model quality for banks individually gives the supervisor important information concerning the supervision of each bank. On the other hand, the analysis of cross-sectional interrelations of the banks, e.g., co-movements, is at least as important because it sheds light on the aspects of systemic risk. From the regulator's point of view, two adverse scenarios are important. The first case is when the regulator's portfolio is not well-diversified in the sense that all portfolios behave similarly. A more specific case is, when portfolios behave similarly in periods of stress.

The economic relevance of the P&L series dominates that of the VaR numbers. Hence we perform our cross-sectional analysis from a P&L point of view. A natural point to start with is the calculation of correlations, say. Yet tables for cross correlations quickly get unmanageable for larger numbers of banks. Instead of looking at $m(m-1)/2 = 78$ cross correlations, it may be more useful to look at the $m = 13$ correlations between the individual P&L series and an aggregated P&L-series. The identity

$$\sum_{j=1}^{m} \operatorname{cov}(C^i, C^j) = \operatorname{cov}(C^i, \sum_{j=1}^{m} C^j) = \operatorname{cov}(C^i, \sum_{j \neq i} C^j) + \operatorname{var}(C^i)$$

motivates to consider the covariances $\operatorname{cov}(C^i, \sum_{j \neq i} C^j)$, as the covariance with the sum of all P&L series is dominated by the last (variance) term for large banks.

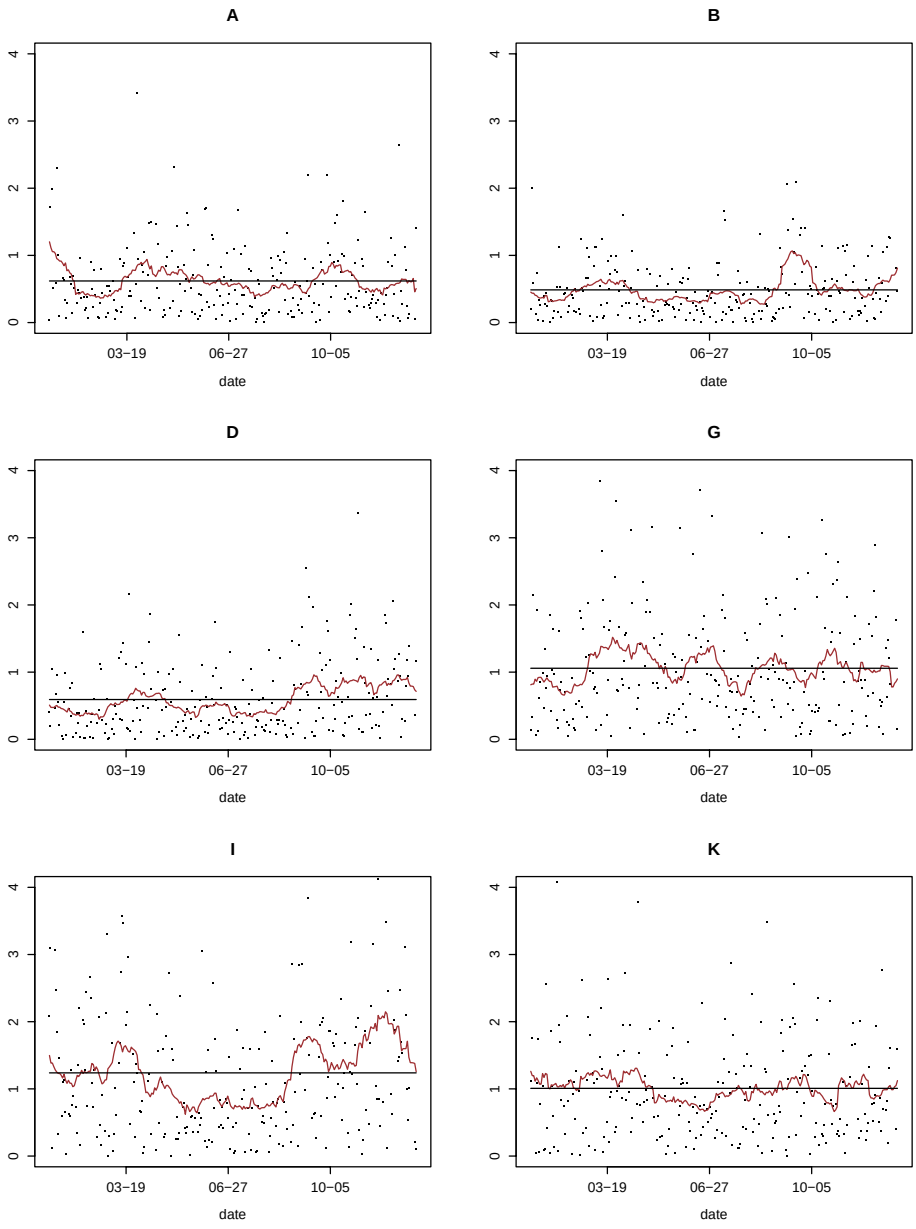

Figure 11: **Local estimates of the inverse recalibration factor.** The dots show the absolute values of the standardized returns, $|r_t|/c_1$, rescaled such, that they can be viewed as estimates of the inverse recalibration factor under the normality assumption. The horizontal line shows the overall average and the other line a sliding 60-day average of the values $|r_t|/c_1$. I.e., the horizontal line shows the inverse of the estimated recalibration factor given in the column "$p = 1$" of table 2.

	Pearson corr.	Spearman corr.
A	0.114	0.070
B	0.206**	0.154*
C	-0.102	-0.096
D	0.527**	0.431**
E	0.081	0.103
F	0.488**	0.389**
G	0.242**	0.213**
H	-0.118	-0.076
I	0.475**	0.333**
J	0.241**	0.186**
K	0.244**	0.163*
L	0.392**	0.270**
M	0.289**	0.264**

Table 3: **Linear and rank correlations between individual banks' P&L with the sum of the P&L of all *other* banks.** The table shows Pearson's linear and Spearman's rank correlations. The significance in terms of the null hypothesis of zero correlation is marked with one or two stars as before.

From table 3 we conclude that only bank H is noticably negatively correlated with the rest, but this is not significant at the 5%-level. It further tells that banks A, C, and E are not significantly correlated to the rest. All other banks have moderate, but significantly positive correlations with the rest.

In order to understand the common movement of the time series of P&Ls in real, monetary terms, we performed a principal component analysis, i.e., an eigen value decomposition of the covariance matrix of P&Ls. The results show, that (1) although the three biggest banks dominate the principal component decomposition of P&Ls, the most important factor only explains 46% of the variance, the second 27%, and the third 13%, and (2) the dominant factor has the meaning "the big banks move in the same direction". A different question is what possible explanatory variables for explaining the co-movement of standardized returns are. For this purpose, we performed an eigen value decomposition of the correlation matrix of standardized returns. It shows that the common movement of returns is remarkably "diverse" in the sense that the three most important factors together explain only 47% of the variance (22%, 14% and 11% each) and there is no dominant factor.

A natural way to measure the degree of diversification is through the ratio of the variance of the combined portfolio and the sum of the variances of its components:

$$I_t := \frac{(\sum_{i=1}^m C_t^i)^2}{\sum_{i=1}^m (C_t^i)^2}, \tag{9}$$

If we replace in

$$I_t = \frac{(\sum_{i=1}^m R_t^i V_t^i)^2}{\sum_{i=1}^m (R_t^i V_t^i)^2}$$

the time-dependent V_t^i by its average $\bar{V}^i$ and define $w_i := \bar{V}^i / \sum_{i=1}^m \bar{V}^i$, then we get

$$\bar{I}_t := \frac{(\sum_{i=1}^m R_t^i \bar{V}^i)^2}{\sum_{i=1}^m (R_t^i \bar{V}^i)^2} = \frac{(\sum_{i=1}^m R_t^i w_i)^2}{\sum_{i=1}^m (R_t^i w_i)^2},$$

which in this context can be interpreted as an *index of co-movement*.

It is bounded below by zero and above as follows. If the banks' P&L-series would be collinear, i.e., $C_t^i = w_i X_t$ (with $X_t = \sum_{i=1}^m C_t^i), w_i > 0$), then $I_t = \frac{1}{\sum_{i=1}^i w_i^2}$. This upper bound on I depends on the size distribution among banks and lies in the interval $[1, m]$. For the 13 banks under consideration, this upper bound is about 5.23, when w_i is taken to be proportional to the average VaR of bank i. If the banks' P&L-series would be stochastically independent, then $I_t = 1$. The other extreme occurs when the banks' trading would be a zero sum game, i.e., $\sum_{i=1}^m C^i = 0$ and consequently $I_t = 0$. In summary, $I_t > 1$ stands for positively correlated P&Ls, $I_t = 1$ for "normal" diversification, and $I_t < 1$ for the partial offsetting of risks beyond "normal" diversification.

Figure 12 shows that local estimates of I_t are above 1 most of the time, but well below its upper bound 5.23. The dramatic stock market movements in September 2001 seem to have had some impact, but the increases in medium-term interest rates beginning in October a lot more. In our view, the index of co-movement shows that the portfolio of the German regulator is quite well-diversified in general. In periods of stress the index increases, but only by a factor of about two, which lends support to the Basel safety factor three.

Figure 12 suggests that in the latter part of the year 2001 there was a common "stress" factor. In order to define "stress" we have to distinguish between stress *variables* that measure stress and stress *events* that denote periods of stress.

Consider the *aggregate stress defined in terms of excess losses*

$$L_t(c) := \sum_{i=1}^m (-\frac{C_t^i}{V_t^i} - c)^+ V_t^i = \sum_{i=1}^m (-C_t^i - cV_t^i)^+ \tag{10}$$

where c is a constant treshold and $x^+ = \max(0, x)$ denotes the positive part. Note, that the aggregate stress in terms of excess losses has a monetary unit. $L_t(1)$ is the sum of excess losses at the VaR-level and $L_t(0)$ is the sum of losses.

Given the notorious unpredictability of financial market returns, extremely high profits for some banks may also point to "stress". Hence, the *aggregate stress defined in terms of excess profits*

$$G_t(c) := \sum_{i=1}^m (\frac{C_t^i}{V_t^i} - c)^+ V_t^i = \sum_{i=1}^m (C_t^i - cV_t^i)^+ \tag{11}$$

may also be interesting to look at.

In the absence of data on the counter party exposures of each bank and an analysis of contagion risk based on such data, both $L_t(c)$ and $G_t(c)$ are stress variables that a supervisory authority might want to monitor.

The lines in the two upper graphs of figure 13 show 2-week averages of aggregate stress for the tresholds $c = 0$, $c = 0.5$, and $c = 1$. The peaks show the excess losses at the

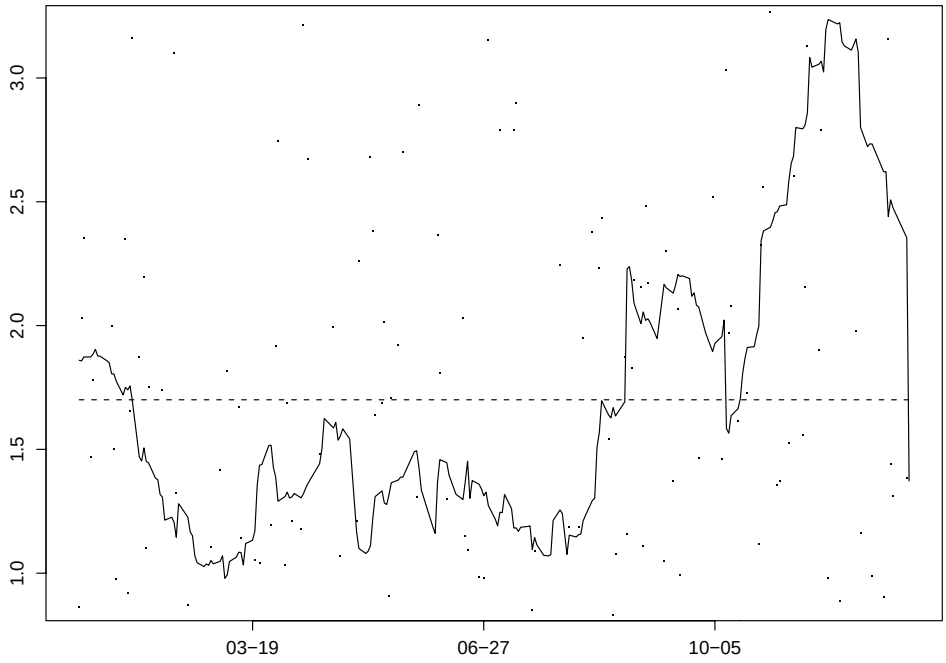

Figure 12: **Measure of co-movement.** The graph shows the local (solid line) and global (dotted line) estimate of the index of co-movement. The points are the squared aggregated P&L divided by the sum of the squared individual P&Ls of a given day. The local estimate is a 60-day sliding average of the points.

VaR-level ($c = 1$). The top graph is based on profits while the graph in the middle is based on losses. In order to see the relation to key risk factors, the bottom graph shows a stock index and an interest rate. Several interesting aspects of the data appear:

- Periods of large excess profits do not in general coincide with periods of large excess losses, which is most pronounced in the first four months.
- There were three periods of "stress" in terms of losses: in April, in September, and in November.
- The November stress was the largest.
- While the April stress was larger than the September stress in terms of aggregate losses ($c = 0$), the September stress was larger than the April stress in terms of more extreme losses ($c = 0.5$ and $c = 1$).
- Visual comparison between the two risk factors and the aggregate stress in terms of losses suggests that losses may be related to increases in short to medium term interest rates.

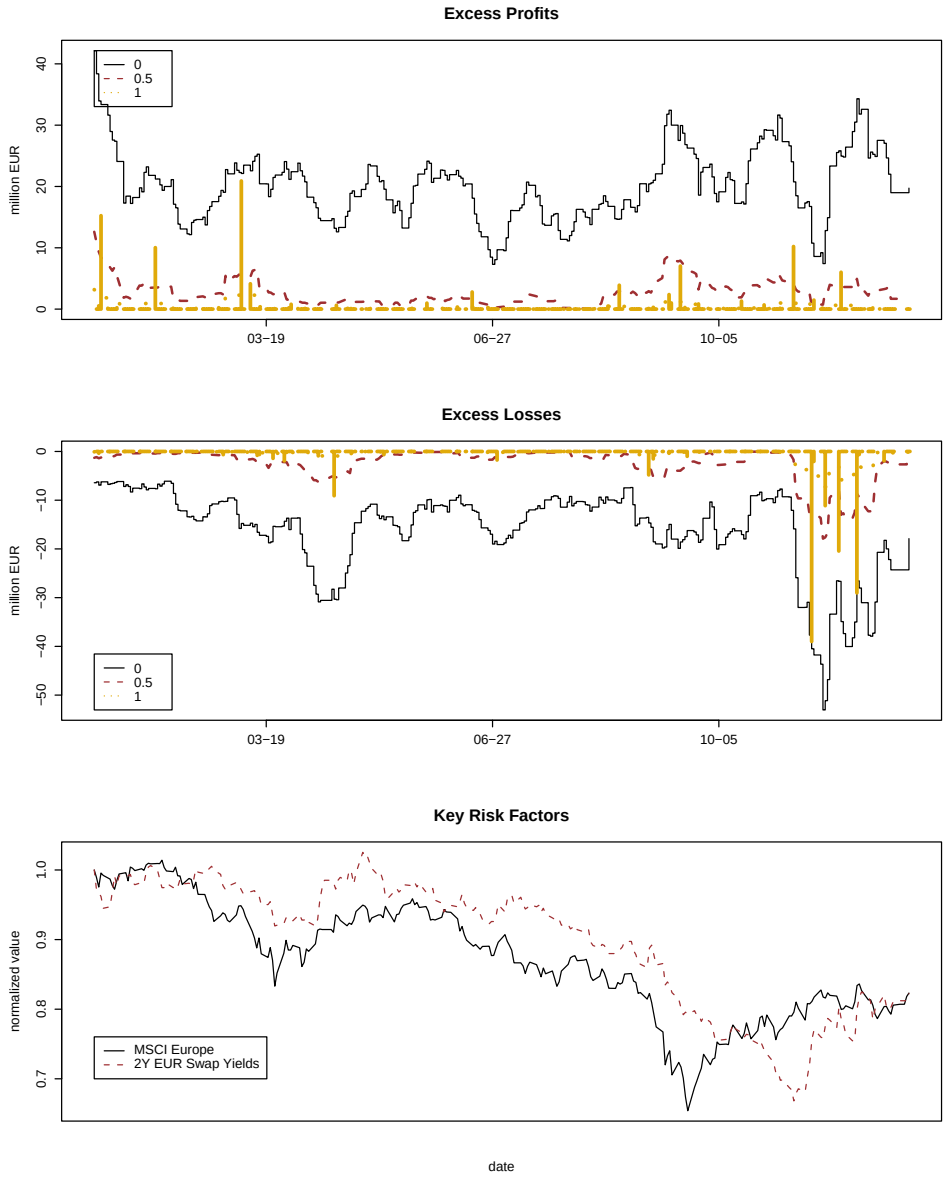

Figure 13: **Two-week averages of aggregate stress for the excess loss tresholds 0, 0.5, and 1.** The lines in the first two graphs show 2-week averages of $G_t(c)$ (top) and $-L_t(c)$ (middle) for the tresholds $c \in \{0, 0.5, 1\}$. The peaks show the sum of excess losses at the VaR-level ($c = 1$). The lower graph shows the MSCI Europe and the euro 2-year swap rate.

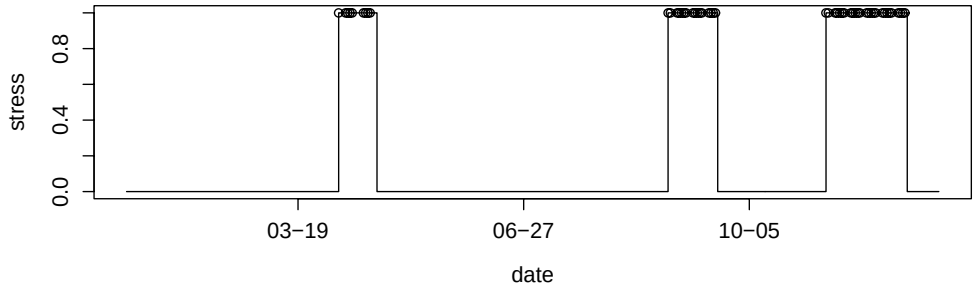

Figure 14: **Stress periods.** The stress periods defined by $\{t \mid L_t(c) > q\}$ with $c = 0.5$ und q being the 80%-quantile of $L_t(c)$.

Any stress variable like $L_t(c)$ can be used to define a set of "stress days" $\{t \mid L_t(c) > q\}$, where q is some quantile of $L_t(c)$. Figure 14 shows the thus defined stress period, using the treshold $c = 0.5$ and q the 80%-quantile of $L_t(c)$.

Having identified periods of stress, the next question is "How do properties of V, C, and R change under stress?". In other words, we compare empirical distributions from the whole time interval with empirical distributions from the stress period as defined in the previous section and depicted in figure 14.

Figure 15 compares the conditional and unconditional densities of the banks' P&L. It shows that the variance increases drastically for banks D and I under stress. Figure 16 is an alternative look at the conditional means and standard deviations. It shows that while the standard deviations increase under stress, they do not exceed the Basel safety factor 3.

Conclusion

The notion of a well-behaved forecast system is closely related to the established concepts of well-calibration and refinement. It generalizes the calibration criterion $F_t(C_t) \sim U[0, 1]$ to a neighborhood of the uniform distribution. We introduced exploratory statistical tools (local estimates of the recalibration factor, the measure of co-movement, conditional distributions under stress) in order to study this property for the VaR forecast systems of German banks. In the light of the analyses based on these tools we can draw three kinds of conclusions.

First, the VaR models of German banks that use these models for regulatory purposes work surprisingly well, even for the special year under consideration. I.e., the forecast systems are well-behaved and the forecast quality is good. The fact that many banks estimate their VaR with a bias can simply be rectified by re-calibrating the forecasts, see also (Bühler et al.; 2002).

Second, we introduced empirical measures related to systemic risk. Their estimates show that the regulator's portfolio is quite well diversified.

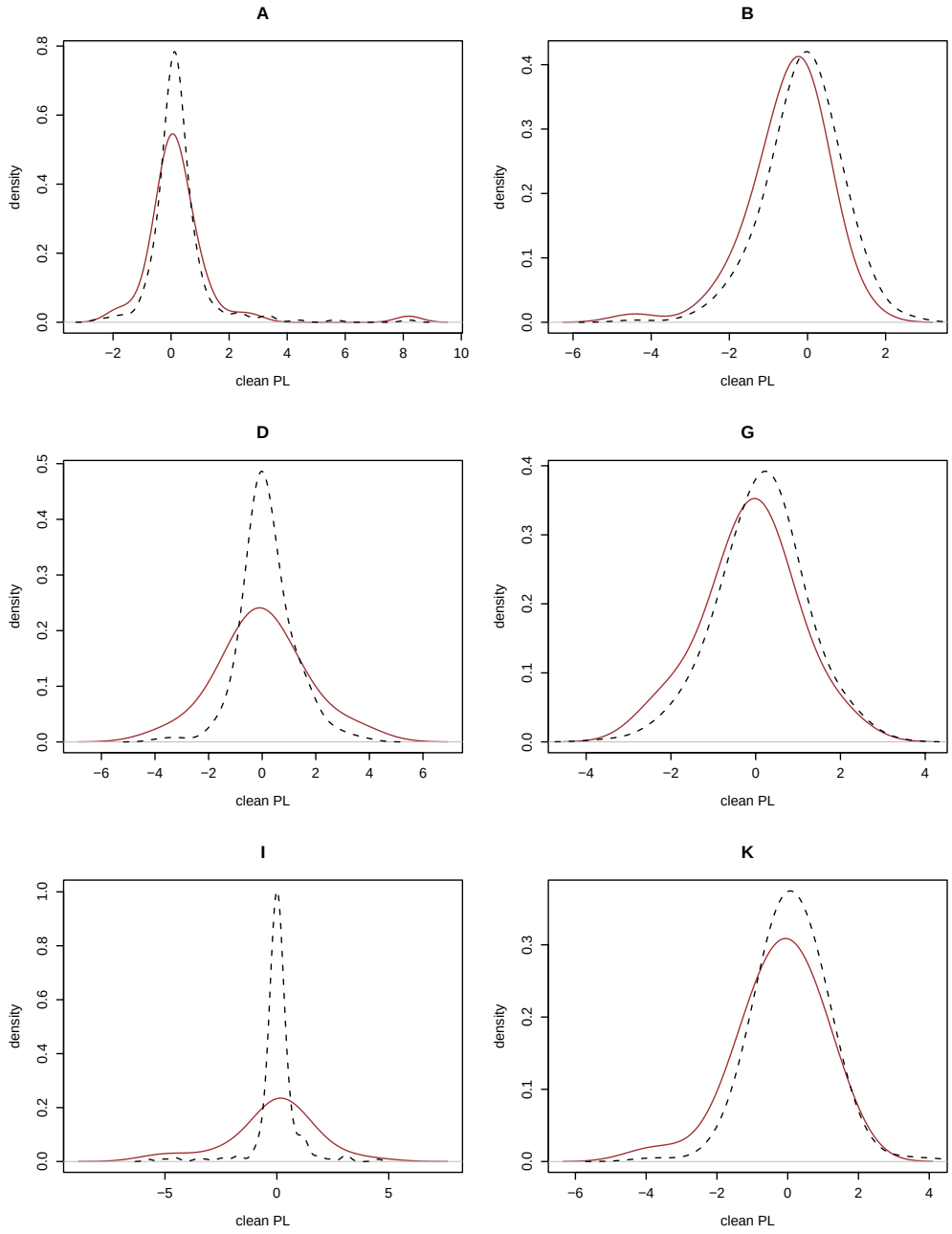

Figure 15: **Conditional densities.** The solid line shows a kernel estimate of the conditional density of the banks' P&L under stress, while the dotted line shows the unconditional density.

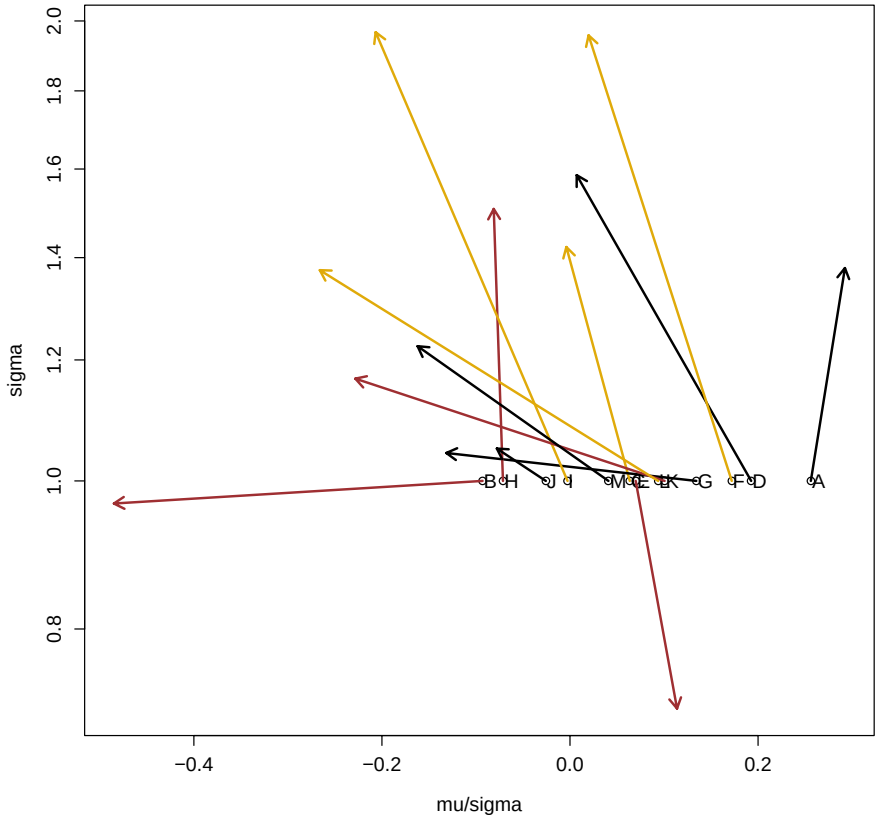

Figure 16: **Conditional means and standard deviations of clean P&L.** The bank's symbol represents the unconditional parameters, while the arrow points to the parameters conditioned on stress. The x-axis shows the conditional and unconditional mean divided by the unconditional standard deviation.

Third, our empirical analyses confirm the architecture of the regulatory framework for Value-at-Risk models. In particular, the main ingredients, i.e., the multiplication factor three and the backtesting penalty function are justified.

A. Well-Calibrated Forecasts

The literature on weather forecasting has developed an elaborate set of concepts and diagnostics for the evaluation of probability forecasts (Murphy and Winkler; 1987, 1992). The two main concepts are *calibration* and *refinement*. This section shows how our definition of "well-calibrated" is related to the established one.

Given the joint distribution of an event a and a probability forecast p for this event, the forecast p is called *well-calibrated*, if the probability of a conditional on the fact that the forecast p has been made, is p: $E[a|p] = p$. The constant forecast $\tilde{p} := E[a]$ (the

"climatological probability") is well-calibrated, but not "refined". The aspect how much information the forecast p contains about the event a is called *refinement* or *resolution* and is formalized and measured in different ways, see Dawid (1986). Partial orderings among forecasts are based on the notion that p_A is more refined than p_B if the forecast p_B can be derived from p_A in a certain way. Complete orderings among forecasts can be defined by the expected loss $ES(a,p)$ for a loss function S, called *scoring rule* in this context. Another way to define refinement for a sequence of forecasts and events (a_i, p_i) is to say that the subsequence of events corresponding to a specific forecast p^*, $(a_i)_{\{i|p_i=p^*\}}$, should be stochastically independent. (Otherwise, it would be possible to improve the forecast.)

In the context where we have a forecast $\hat{F}$ for the probability distribution F of a continuous random variable C, the concept of well-calibration becomes

$$P\{C \leq x|\hat{F}\} = \hat{F}(x), \tag{12}$$

i.e., the forecast of each event based on C is well-calibrated (Dawid; 1984, p.281). If $\hat{F}$ is continuous, (12) implies that $\hat{F}(C)$ is uniformly distributed on $[0,1]$. The requirement that a sequence of "realized probabilities" $\hat{F}_t(C_t)$ is stochastically independent is a kind of refinement requirement.

Given that banks do not report the whole forecast distribution $\hat{F}$ but only a quantile $\hat{q}(\alpha) := \hat{F}^{-1}(\alpha)$, the important question is in which sense and under which conditions the realized probabilities $F(R_t)$ of the standardized returns

$$R_t := C_t \frac{q(\alpha)}{\hat{q}_t(\alpha)} \qquad (q(\alpha) = F^{-1}(\alpha)) \tag{13}$$

for some fixed p.d.f. F can be taken as a substitute for the "true" realized probabilities $\hat{F}_t(C_t)$.

If the forecast $\hat{F}$ comes from a location-scale family of distributions, i.e.,

$$\hat{F}(x) = F((x - \hat{\mu})/\hat{\sigma}), \tag{14}$$

and the location $\hat{\mu}$ is 0, then the realized probabilities of the standardized returns are obviously a perfect substitute for the "true" realized probabilities:

$$\hat{F}_t(C_t) = F(C_t/\hat{\sigma}_t) = F(C_t \frac{q(\alpha)}{\hat{q}_t(\alpha)}) = F(R_t).$$

Interestingly, the converse also holds.

Proposition 1 *Let C be a random variable and F_0 its distribution under the true probability. Let $\hat{F}$ be a forecast for F_0 and F a fixed "benchmark" distribution. Assume F, F_0, and $\hat{F}$ are continuous and have support $(-\infty, \infty)$. Let q and $\hat{q}$ denote the inverses of F and $\hat{F}$, respectively. The value $U = F(C\frac{q(\alpha)}{\hat{q}(\alpha)})$ has the same distribution as the realized probability $\hat{F}(C)$ under the true probability measure if and only if $\hat{F}$ comes from the scale-family $\hat{F}(x) = F(x\frac{q(\alpha)}{\hat{q}(\alpha)})$.*

Proof. We only have to show the non-trivial direction. Assume that

$$P_0\{U \leq p\} = P_0\{\hat{F}(C) \leq p\} \qquad \forall p \in [0,1].$$

This implies

$$F_0\left(F^{-1}(p)\frac{\hat{q}(\alpha)}{q(\alpha)}\right) = F_0\left(\hat{F}^{-1}(p)\right).$$

Since F_0 is invertible, this equation also holds for the arguments of $F_0(.)$. Setting $x := F^{-1}(p)\frac{\hat{q}(\alpha)}{q(\alpha)}$, this leads to

$$\hat{F}(x) = p$$

which equals

$$= F\left(x\frac{q(\alpha)}{\hat{q}(\alpha)}\right)$$

by definition of x. □

The assumption that the banks' forecasts $\hat{F}$ come from a location-scale family with location 0 is a rather unrealistic one. We actually know from those banks that use historical simulation and various delta-gamma-normal methods that the forecasts $\hat{F}$ do *not* in general come from such a family. In the empirical analysis, however, we observed that the empirical distribution of the standardized returns is remarkably close to normal. An interesting question is now how the two concepts of well-calibration are related under the additional assumption

$$R = C\frac{q(\alpha)}{\hat{q}(\alpha)} \sim F(./\sigma) \tag{15}$$

for some fixed benchmark distribution F and scale $\sigma > 0$.

Proposition 2 *Let $\hat{F}$ be a forecast for the distribution F_0 of a random variable C and F a fixed "benchmark" distribution. Assume F, F_0, and $\hat{F}$ are continuous and have support $(-\infty, \infty)$. Let q and $\hat{q}$ denote the inverses of F and $\hat{F}$, respectively. Assume that the standardized returns $R = C\frac{q(\alpha)}{\hat{q}(\alpha)}$ are known to come from the scale family (15) and the forecast $\hat{F}$ is well-calibrated as in (12). Then $\sigma = 1$ and the forecast $\hat{F}$ "comes on average from a scale-family" in the sense of*

$$F(x) = E\left[\hat{F}\left(x\frac{\hat{q}(\alpha)}{q(\alpha)}\right)\right]. \tag{16}$$

In general it cannot be concluded, however, that each single forecast comes from a scale family $\hat{F}(x) = F(x/\hat{\sigma})$.

Proof. We first prove $\sigma = 1$. Since $\hat{F}$ is well-calibrated,

$$\begin{aligned}\alpha &= P\{C \leq \hat{q}(\alpha)|\hat{F}\} \\ &= P\{R \leq q(\alpha)|\hat{F}\}.\end{aligned}$$

Taking expectation, this also holds unconditionally:

$$\alpha = P\{R \leq q(\alpha)\},$$

which equals $F(q(\alpha)/\sigma)$ because of (15). But this implies $\sigma = 1$.

Furthermore,

$$\begin{aligned} F(x) &= P\left\{C\frac{q(\alpha)}{\hat{q}(\alpha)} \leq x\right\} \\ &= E\left[P\{C\frac{q(\alpha)}{\hat{q}(\alpha)} \leq x|\hat{F}\}\right] \\ &= E\left[\hat{F}\left(x\frac{\hat{q}(\alpha)}{q(\alpha)}\right)\right]. \end{aligned}$$

Example. Assume $C = m + X$; $m \sim G$ and $X \sim F_0(./\sigma_0)$ are independent random variables. Assume further that the bank has advance knowledge of m (but no knowledge of X), then the forecast

$$\hat{F}(x) := F_0((x - m)/\sigma_0)$$

is well-calibrated in the sense (12), it comes from a location-scale family, but the location is not necessarily 0. The standardized return becomes

$$R = C\frac{q(\alpha)}{\hat{q}(\alpha)} = (m + X)\frac{q(\alpha)}{m + \sigma_0 q_0}$$

where $q_0 := F_0^{-1}(\alpha)$. This equation shows that the conditional distribution of R given m comes from the location-scale family defined by F_0.

Being more specific, assume F_0 is the standard normal distribution. Then the distribution F observed by the supervisor is a certain mixture of normal distributions and will usually have a different shape as and fatter tails than F_0. □

The proposition shows that the notion of well-calibration based on the standard deviation of R_t differs from the notion of well-calibration in the sense of Dawid (1986), unless one makes additional assumptions. For equivalence, one needs the rather unrealistic assumption that all individual forecasts come from a scale-family. Under the weaker condition (15), well-calibration implies $\sigma = 1$. I.e., if we "observe" that the standardized returns R_t are approximately normal, then the variance of the standardized returns is 1 for well-calibrated forecasts. The condition $\sigma(R_t) = 1$ as a criterion of well-calibration is to be understood in this sense in the paper.

References

Basel Committee on Banking Supervision (1996a). Amendment to the capital accord to incorporate market risks, `http://www.bis.org/publ/bcbs24.pdf`.

Basel Committee on Banking Supervision (1996b). Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements, `http://www.bis.org/publ/bcbs22.pdf`.

Berkowitz, J. (2000). Testing density forecasts, with applications to risk management, `http://www.uh.edu/~jberkowi/back.pdf`. appeared in the *Journal of Business and Economic Statistics* (October 2001).

Berkowitz, J. and O'Brien, J. (2002). How accurate are value-at-risk models at commercial banks?, *Journal of Finance* **57**: 1093–1112.

Bühler, W., Engel, C., Korn, O. and Stahl, G. (2002). Backtesting von Kreditrisikomodellen, *in* A. Oehler (ed.), *Kreditrisikomanagement - Portfoliomodelle, Derivate, Backtesting und Aufsicht*, 2nd edn, Schäfer-Poeschel, Stuttgart, pp. 181–217.

Christoffersen, P. (1998). Evaluating interval forecasts, *International Economic Review* **39**: 841–862.

Crnkovic, C. and Drachman, J. (1996). Quality control, *RISK* **9**(9): 138–143.

Dawid, A. P. (1982a). Intersubjective statistical models, *Exchangeability in probability and statistics (Rome, 1981)*, North-Holland, Amsterdam, pp. 217–232.

Dawid, A. P. (1982b). The well-calibrated Bayesian, *J. Amer. Statist. Assoc.* **77**(379): 605–613.

Dawid, A. P. (1984). Statistical theory. The prequential approach, *J. Roy. Statist. Soc. Ser. A* **147**(2): 278–292.

Dawid, A. P. (1986). Probability forecasting, *Encyclopedia of Statistical Sciences*, John Wiley & Sons, New York, pp. 210–218.

Jorion, P. (2001). *Value at Risk: The New Benchmark for Managing Financial Risk*, 2 edn, McGraw-Hill, New York.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models, *Journal of Derivatives* **2**: 73–84.

Longerstaey, J. (1996). RiskMetrics technical document, *Technical Report fourth edition*, J.P.Morgan. originally from `http://www.jpmorgan.com/RiskManagement/RiskMetrics/`, now `http://www.riskmetrics.com`.

Lopez, J. (1999). Regulatory evaluation of Value-at-Risk models, *Journal of Risk* **1**: 37–63.

Matten, C. (2000). *Managing Bank Capital*, 2nd edn, John Wiley & Sons, Chichester.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification, *Monthly Weather Review* **116**: 1330–1338.

Murphy, A. H. and Winkler, R. L. (1992). Diagnostic verification of probability forecasts, *International Journal of Forecasting* **7**: 435–455.

Overbeck, L. and Stahl, G. (2000). Backtesting: Allgemeine Theorie, Praxis und Perspektiven, *in* L. Johanning and B. Rudolph (eds), *Handbuch Risikomanagement*, Uhlenbruch Verlag, Bad Soden, pp. 289–320.

Riedel, A. (2002). *Qualität der Value-at-Risk-Schätzung deutscher und US-amerikanischer Banken*, Diplomarbeit, Humboldt-Universität, Berlin, Germany.

Rosenblatt, M. (1952). Remarks on a multivariate transformation, *Ann. Math. Statistics* **23**: 470–472.

Seillier-Moiseiwitsch, F. (1993). Sequential probability forecasts and the probability integral transform, *International Statistical Review* **61**(3): 395–408.

Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993). On testing the validity of sequential probability forecasts, *J. Amer. Statist. Assoc.* **88**(421): 355–359.

CFS Working Paper Series:

No.	Author(s)	Title
2003/22	**Roman Kräussl**	Do Changes in Sovereign Credit Ratings Contribute to Financial Contagion in Emerging Market Crises?
2003/23	**Roman Kräussl**	A Critique on the Proposed Use of External Sovereign Credit Ratings in Basel II
2003/24	**Tereza Tykvová**	Is the Behavior of German Venture Capitalists Different? Evidence from the Neuer Markt
2003/25	**Tereza Tykvová**	The Role of the Value Added by the Venture Capitalists in Timing and Extent of IPOs
2003/26	**Stefanie Franzke** **Stefanie Grohs** **Christian Laux**	Initial Public Offerings and Venture Capital in Germany
2003/27	**Andreas Hackethal**	German banks – a declining industry?
2003/28	**Frank A. Schmid** **Mark Wahrenburg**	Mergers and Acquisitions in Germany-Social Setting and Regulatory Framework
2003/29	**Christina E. Bannier**	Privacy or Publicity – Who Drives the Wheel?
2003/30	**Markus Mentz** **Steffen P. Sebastian**	Inflation convergence after the introduction of the Euro
2003/31	**Seppo Honkapohja** **Kaushik Mitra**	Performance of Inflation Targeting Based On Constant Interest Rate Projections
2003/32	**Stefan Jaschke** **Gerhard Stahl** **Richard Stehle**	Evaluating VaR Forecasts under Stress – The German Experience

Copies of working papers are available at the Center for Financial Studies or can be downloaded (http://www.ifk-cfs.de).