Jon Danielsson, Kevin R. James, Marcela Valenzuela, Ilknur Zer

Model risk of risk models

Article (Accepted version) (Refereed)

Original citation:
Danielsson, Jon, James, Kevin R., Valenzuela, Marcela and Zer, Ilknur (2016) *Model risk of risk models.* Journal of Financial Stability, 23. pp. 79-91. ISSN 1572-3089

DOI: 10.1016/j.jfs.2016.02.002

Model Risk of Risk Models*

Jon Danielsson
Systemic Risk Centre
London School of Economics

Kevin James
Systemic Risk Centre
London School of Economics

Marcela Valenzuela
University of Chile, DII

Ilknur Zer
Federal Reserve Board

February 2016

Forthcoming in Journal of Financial Stability

Abstract

This paper evaluates the model risk of models used for forecasting systemic and market risk. Model risk, which is the potential for different models to provide inconsistent outcomes, is shown to be increasing with market uncertainty. During calm periods, the underlying risk forecast models produce similar risk readings; hence, model risk is typically negligible. However, the disagreement between the various candidate models increases significantly during market distress, further frustrating the reliability of risk readings. Finally, particular conclusions on the underlying reasons for the high model risk and the implications for practitioners and policy makers are discussed.

Keywords: Market risk, systemic risk, Value-at-Risk, expected shortfall, MES, CoVaR, financial stability, risk management, Basel III
JEL classification: G01, G10, G18, G20, G28, G38

*Corresponding author Ilknur Zer, Federal Reserve Board, 20th Street and Constitution Avenue N.W. Washington, D.C. 20551, USA, ilknur.zerboudet@frb.gov, +1-202-384-4868. The views in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. The early version of this paper is circulated under the title "Model Risk of Systemic Risk Models". We thank the Economic and Social Research Council (UK) [grant number: ES/K002309/1], and the AXA Research Fund for its financial support provided via the LSE Financial Market Group's research programme on risk management and regulation of financial institutions. Valenzuela acknowledges the support of Fondecyt Project No. 11140541 and Instituto Milenio ICM IS130002. We also thank Kezhou (Spencer) Xiao for excellent research assistance. Finally we thank Seth Pruitt, Kyle Moore, John W. Schindler, and participants at various seminars and conferences where earlier versions of this paper were presented. All errors are ours. Updated versions of this paper can be found on www.RiskResearch.org and the Webappendix for the paper is at www.ModelsandRisk.org/modelrisk.

1 Introduction

Following the 2008 crisis, risk forecasting has emerged as a key public concern. Statistical risk measures are set to play a much more fundamental role in policy and decision making within financial institutions than before the crisis. Hence, an understanding of the model risk of risk forecast models—that is, the potential for different underlying risk forecast models to provide inconsistent outcomes—is of considerable interest to both policymakers and practitioners. The empirical study of such risk for macroprudential and internal management purposes constitutes the main motivation of this paper.

Why does model risk matter? Risk models play a fundamental role in the regulatory process and are directly embedded within the Basel regulations and are therefore used to determine bank capital. While their use for macroprudential purposes is not as clear, there are a number of proposals from the academic and public sectors for using these models for setting bank capital and surcharges to meet systemic risk. Hence, the output of these models has a real economic impact. For these reasons, it is important to understand to what extent decision makers can rely on risk models and when their use is not advisable.

We start by proposing a general framework for quantifying model risk. To this end, we focus on the level of disagreement amongst the candidate models and propose a new method we term *risk ratio.* This entails applying a range of common risk forecast methodologies to the problem of forecasting risk, and calculating the ratio of the maximum to the minimum risk forecasts. This provides a succinct way of capturing model risk because if the underlying models have passed some model evaluation criteria used by the authorities and financial institutions, they can be considered reputable risk forecasting candidates. If risk is forecasted by a number of equally good models, the risk ratio should be close to 1. If the risk ratio is very different from 1, then it captures the degree to which different models disagree, providing a measure of model risk.

We first focus our attention on the five most commonly used risk forecast models: historical simulation, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory. In addition, we include six hybrid models identified in the literature as high quality: both extreme value theory and historical simulation applied to GARCH filtered data under the assumptions of normal, student-t, and skewed-t error term distributions. While it would be straightforward to expand the universe of models if another prominent candidate emerges, it will not materially affect the results since any additional model can only increase model risk.

We first apply the risk ratio methodology on market risk measures. Value-at-Risk (VaR) has been the main building block of market risk regulations since its first incorporation into the Basel Accords in 1996; hence, the model risk of VaR is our starting point. In addition, we consider the model risk of expected shortfall (ES), since the Basel committee (2013, 2014) has proposed replacing VaR with ES in market risk regulations.

We then propose a general setup for the classification of systemic risk models (SRMs), providing a lens through which to analyze the most common market data based systemic risk models (SRMs). The prominent marginal expected shortfall (MES) (Acharya et al., 2010), conditional value at risk (CoVaR) (Adrian and Brunnermeier, 2011), SRISK (Brownlees and Engle, 2015; Acharya et al., 2012), Co-Risk (IMF, 2009), and BIS's Shapley value method (Tarashev et al., 2010) all fall under our classification setup. While intended for different purposes, these measures and market risk regulation techniques are closely related; both elementally depend on VaR, suggesting that the model risk of VaR is likely to pass through to market data based SRMs. One could apply the risk ratio approach to the various market data based SRMs, but given their common ancestry, we expect the results to be fundamentally the same, and in the interest of brevity we focus on two SRMs: MES and CoVaR.

The data set consists of large financial institutions traded on the NYSE, AMEX, and NASDAQ from the banking, insurance, real estate, and trading sectors over a sample period spanning 1970 to 2012. We find that on average, model risk is quite low, indicating that in typical situations decision makers do not have to be too concerned about model choice or model risk. However, the situation changes when looking at individual stocks and periods of stress in financial markets. Model risk is significantly higher when an individual stock is subject to idiosyncratic shocks or when financial markets are stressed. The average maximum 99% VaR risk ratio across the whole sample is 9.23, and in the most extreme case it reaches 55.32, during the 1987 crash. None of the models *systematically* gives the lowest or highest forecasts, and the large risk ratios are not driven by the inclusion of a particular model.

The empirical results are a cause for concern, as the degree of model risk documented here frustrates internal risk management as well as macro-prudential and micro-prudential policy. For this reason, our results should be of considerable value to policymakers and risk managers alike, who will get a better understanding of the reliability of risk models and how to understand the problem of conflicting measurements of the same underlying risk. Ultimately, a better understanding of model risk should lead to more robust policymaking and asset allocation.

We suspect the problem of model risk arises for two reasons. The first is the low frequency of actual financial crises. Developing a model to capture risk during crises is quite challenging,

since the actual events of interest have almost never happened during the observation period. Such modeling requires strong assumptions about the stochastic processes governing market prices that are likely to fail when the economy enters a crisis.

Second, common statistical models assume risk is exogenous—extreme events arrive to the markets from outside, like an asteroid would, and the behavior of market participants has nothing to do with the crisis. However, as argued by Danielsson and Shin (2003); Brunnermeir and Sannikov (2014), risk is really endogenous, created by the interaction between market participants and by their desire to bypass risk control systems. As both risk takers and regulators learn over time, we can also expect price dynamics to change, further frustrating statistical modeling.

It is important to recognize that the output of risk forecast models is used as an input into expensive decisions, be they portfolio allocations or the amount of capital held. Hence, the minimum acceptable criterion for a risk model should not be to weakly beat noise, but the quality of the risk forecasts should be sufficiently high so the cost of type I and type II errors are minimized, as argued by Danielsson et al. (2015a).

Furthermore, most successful market risk methodologies, including all of those discussed here, were originally designed for the day-to-day management of market risk in financial institutions. In our view, one should be careful when using the same statistical toolkit for the more demanding job of systemic and tail risk identification.

The outline of the rest of the paper is as follows. Section 2 gives the details of the model risk analysis conducted. Section 3 presents the empirical findings for market regulatory models. Section 4 provides a classification system for systemic risk methodologies and examines the model risk of market data based systemic risk models. Section 5 features a discussion of our main findings. Section 6 concludes.

2 Model risk analysis

Broadly speaking, model risk relates to the uncertainty created by not knowing the data generating process. That high level definition does not provide guidance on how to assess model risk, and any test for model risk will be context dependent.

Within the finance literature, Green and Figlewski (1999); Cont (2006); Hull and Suo (2002) underline three different sources of model risk. First, there is uncertainty on the choice of the model itself. Second, the underlying theoretical model could be misspecified. Third, some of the input parameters in the underlying model could be unobservable and hence

may require assumptions for empirical implementation. For Gibson (2000), model risk is defined as uncertainty over the risk factor distribution, whereas Alexander and Sarabia (2012) distinguish two sources of model risk: inappropriate assumptions about the form of the statistical model, and parameter uncertainty (i.e., estimation error in the parameters of the chosen model). Finally, Hendricks (1996); Glasserman and Xu (2013); Boucher et al. (2014) define model risk as inaccuracy in risk forecasting that arises from estimation error and the use of an incorrect model.

Our interest here is in a particular practical aspect of model risk—how the use of different candidate models, all feasible ex-ante, may lead to widely different risk forecasts. It has been known since the very first days of financial risk forecasting that different models can produce vastly different outcomes, where it can be difficult or impossible to identify the best model, (e.g., Hendricks, 1996; Berkowitz and O'Brien, 2002; Danielsson, 2002; O'Brien and Szerszen, 2014). This problem arises because financial risk is latent, it cannot be directly measured and instead has to be forecasted by a statistical model. Hence, the definition of model risk we focus on most closely resembles that of Green and Figlewski (1999); uncertainty on the model choice itself creates model risk as there are many standard VaR forecast models used.

2.1 Model and measure choices

Our objective is to capture the resulting model disagreement into one statistical measure, the risk ratio. We focus on the two most commonly used market risk measures, VaR and ES, explicitly addressing the specific case of Basel III. We also consider market data based systemic risk measures, and provide the first empirical evidence documenting how model risk in VaR and ES passes through to the SRMs.

At this stage, we are left with the choice of whether to identify the best model amongst the candidates. In that, we are guided by a large literature on model choice (see for instance Bao et al., 2006; Kuester et al., 2006; Brownlees and Gallo, 2009; O'Brien and Szerszen, 2014, among others).

A typical way to identify the best risk forecast model is by backtesting, usually through the analysis of violation ratios. While any systematic occurrence of violations quickly shows up in backtest results, violations are just one of many criteria for evaluating the performance of risk models such as volatility of risk forecasts, clustering of violations, extreme tail risk, overestimation or underestimation only, and more. This means that passing one criteria might be relatively unimportant for any particular user.

Furthermore, in the specific case of SRMs, what matters is extreme outcomes, which by definition are very infrequent. The paucity of data during such time periods makes it difficult, if not impossible, to formally test for violations and to obtain robust backtest results.

Thus, instead of identifying the best model, we opt to focus on a different relevant question: examining the consistency/discrepancy of a set of candidate models. That leaves us with the final question of how to quantify the degree of model disagreement. We opted for a simple approach: the ratio of the highest to the lowest forecasts, or risk ratio. We tried other summary measures of disagreement, such as standard deviations and absolute deviations, but the results do not change qualitatively.

2.2 The risk ratio approach to model risk

Consider the problem of forecasting risk for day $t+1$ using information available on day t. Suppose we have N candidate models to forecast risk on day $t+1$, each providing different forecasts:

$$\left\{\text{Risk}_{t+1}^{n}\right\}_{n=1}^{N}.$$

We then define model risk as the ratio of the highest to the lowest risk forecasts

$$\text{Risk Ratio}_{t+1} = RR_{t+1} = \frac{\max\left\{\text{Risk}_{t+1}^{n}\right\}_{n=1}^{N}}{\min\left\{\text{Risk}_{t+1}^{n}\right\}_{n=1}^{N}}.$$

The risk ratios provide a clear unit-free way to compare the degree of divergence, as long as the underlying models are recognized as high quality, are in use by financial institutions, and have passed muster with the authorities.

The baseline risk ratio estimate is 1. If we forecast the risk by a number of equally good models, the risk ratio should be close to 1; a small deviance can be explained by estimation risk. Therefore a risk ratio very different from 1 captures the degree to which different models disagree. In this case, both practitioners and regulators end up with valid but inconsistent risk forecasts.

2.3 Models

A very large number of models have been proposed for forecasting market risk. Unfortunately, it is difficult to map out all of the models used in the industry. While one can get some guidance in reading the annual reports of financial institutions, the stated model choice may

be different from the model used for capital purposes, which is yet different from the models used for internal risk control.

To the best of our knowledge, there is no comprehensive survey listing the most commonly used models by industry. As the model choice depends on the specific industry and jurisdiction, it may not be possible to create such a survey. Local regulators create their own rulebooks, and even within the context of the Basel regulations, the same model may be allowed in one jurisdiction and disallowed in another. Different industries—especially those regulated more lightly, such as asset managers—might use different models while not being required to disclose the model used.

Faced with a large number of models, the decision of which models to include is not straightforward. We do not want to simply pick every model because including a model that is not in use and that persistently delivers the highest or lowest risk forecasts would artificially inflate the risk ratio. We therefore opt for a two-level approach. First, we pick the five mainstream models that are most commonly discussed in the academic, practitioner, and regulatory literatures. Second, we include six more sophisticated models that are identified in the academic literature as the best at forecasting risk.

That said, such a choice is inevitably subjective; hence, we also present the sensitivity of the results by excluding specific models in Section 3.2. In addition, in the web appendix at `www.ModelsandRisk.org/modelrisk`, we provide the individual risk forecasts so that anybody can calculate model risk for their particular subset of our universe of models.

First, being one of the simplest, historical simulation (HS) is one of the most common risk forecast models preferred in the industry. For instance, the two largest bank holding companies in our sample, Bank of America and JPMorgan, calculate trading risk via HS for their annual reports.

Second, we include three GARCH family models. Normal GARCH (G) is one of the most popular models to forecast volatility, along with exponentially weighted moving average (EWMA), as the non-stationary version of GARCH (1,1) (see for example Poon and Granger, 2003, for a literature review on volatility forecasting models). Several authors have documented that the normal GARCH is not sufficient for capturing tail events, proposing student-t GARCH (tG) (see for example Bauwens and Laurent, 2005; Bali and Theodossiou, 2007; Marimoutou et al., 2009).

Third, we consider extreme value theory (EVT) models since several authors have argued that they provide more accuracy and stability than GARCH-derived risk forecasts (see for example Danielsson and Morimoto, 2000; Bekiros and Georgoutsos, 2005).

We also include several models identified by the academic literature as being of particularly high quality. In this, we make use of Kuester et al. (2006) who compare alternative VaR forecasting models and show that hybrid models, in general, perform better compared with the simple approaches. Though it is not possible to capture all possible hybrid/mixing approaches, we include six models that have been shown to perform the best within the universe considered by Kuester et al. (2006). Specifically, we include filtered historical simulation model assuming the error terms are normal (FHS), student-t (tFHS), and skewed-t distributed (stFHS). In addition, another hybrid model combining a GARCH filter with an EVT approach is included under the assumption of normal (G-EVT), student-t (tG-EVT), and skewed-t (stG-EVT) distributed error terms.

Many of the models we consider are related to each other and can be expected to deliver similar and correlated results. For example, the EWMA is nested within GARCH. The HS and EVT models are both based on unconditional empirical quantiles, where EVT fits a parametric function to the tails. The normal and student-t GARCH models are related by the volatility specification but with a different conditional distribution. Finally, the hybrid models produce correlated VaR readings as they all require GARCH filtering as a first step in their calculation. However, the degree of similarity between the various models does not necessarily imply that the model risk is negligible. Even related models can deliver different results.

Hence, a relevant question is whether we can only consider, for instance, GARCH family models and then conclude that the model risk is negligible. Our answer is no for two reasons. First, different models have different merits. Some end users may prefer to use a model that reacts quickly to the news (e.g., GARCH), others may prefer easy computation (e.g. EWMA or HS), or low volatility of risk forecasts and hence, stability (e.g. HS), and still others may want to best capture the fat tails (such as EVT or student-t GARCH). In addition, although computationally trickier, hybrid models are shown to have good risk forecast properties (e.g., Kuester et al., 2006). Therefore, the choice of risk measure is tailored to the preferences of the end user.

Second, even the same risk forecast model may be subject to model risk simply due to different parameters chosen by the end-user. For example, according to the 2014 annual reports, Bank of America calculates its trading risk via 99% VaR using HS with a three-year estimation window period. Similarly, JPMorgan calculates its trading risk by employing HS, though via 95% VaR with one-year estimation windows. We calculate the market risk of the S&P 500 index using the above two sets of parameters. We find that the minimum risk ratio is 1, as expected, but only for 5% of the times. The correlation of the risk readings is only 66%, where the disagreement between the two risk readings can go up to 4.

2.4 Data

Our sample includes all NYSE, AMEX, and NASDAQ–traded financial institutions from the banking, insurance, real estate, and trading sectors with SIC codes from 6000 to 6799. We collect daily prices, holding period returns, and number of shares outstanding from CRSP 1925 US Stock Database for the period January 1970 to December 2012. We then keep a company in the sample if (1) it has more than 1,000 return observations, (2) it has less than 30 days of consecutively missing return data, and (3) it is one of the largest 100 institutions in terms of market capitalization at the beginning of each year. This yields a sample of 439 institutions.

We estimate daily 99% VaR values for each model and each company, where the portfolio value is set to be \$100 and the estimation window is 1,000 days. Some of the stocks in the sample have periods of infrequent trading. For those illiquid stocks, it was not possible to obtain VaR forecasts for every estimation method and each day.[1] In particular, the models were unable to simultaneously find parameter combinations that work for market outcomes when a company is not traded for consecutive days. We could either exclude those methods, or exclude the data with such illiquidity. We opt for the latter and remove the part of a stock's sample that contains more than one week's worth of zero returns; that is, we truncated the sample instead of removing the stock or the zeroes. However, in practice, truncation does not have an impact on our results. For comparison, we also run the model without truncation, and found that the results did not change in any substantial way. Yet, we end up with more outliers in the risk forecasts. Similarly, increasing or decreasing the weekly threshold did not materially alter the results.

3 Model risk of market risk models

When we apply the risk ratio analysis to our range of risk forecast models and risk measures we find that model risk is always present, regardless of the asset. Overall, such risk is low in periods when economic and market conditions are benign, tending to pick up along with economic and market uncertainty. The results do not appear to be driven by any particular model, the two models identifying the risk ratios change quite rapidly with time, and every model has at one point delivered either the highest or lowest risk readings.

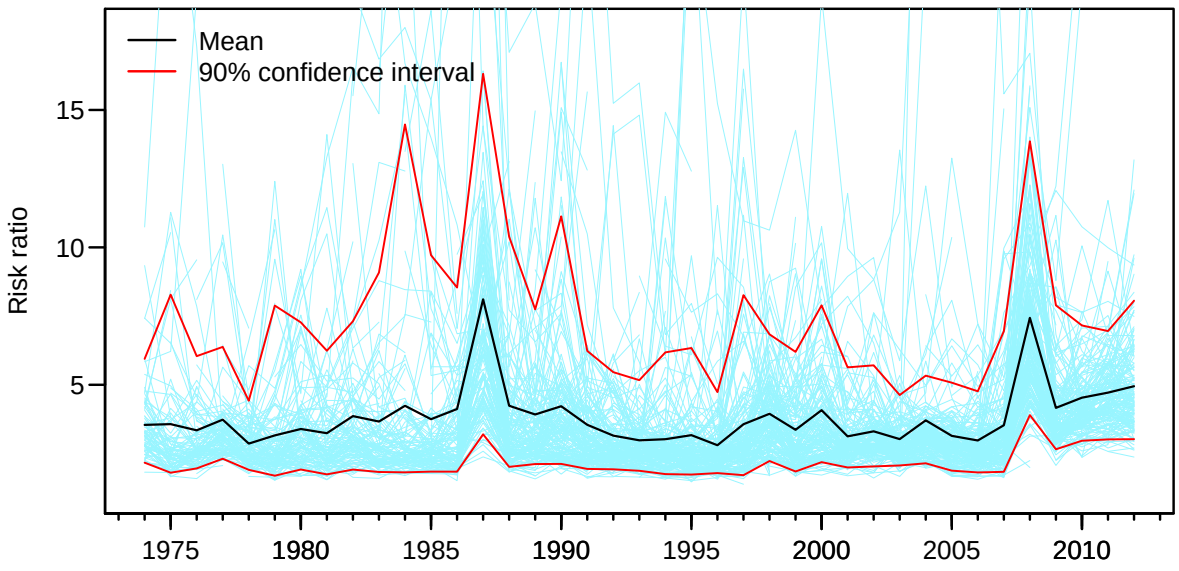
[1] In rare cases, the nonlinear optimization methods did not converge or the solution resulted in numerical instabilities pushing up the VaR numbers, usually for student-t GARCH. We omitted those outlier cases. As there are so few instances, they do not qualitatively affect our results.

Figure 1 shows the model risk of all institutions in our sample, along with the mean and 95% empirical quantiles. While a complete set of results can be obtained from the web appendix, below we present the main results, focusing on daily VaR at the 99% level.

The risk ratio across the whole sample is, on average, about 4. Thus, for a typical asset on a typical day, the highest risk forecast is about quadruple that of the lowest risk forecast. The results, however, vary quite a lot with time, and in periods of stress model risk increases significantly. During the 1987 crash, it exceeds 15 and at the most extreme case, it reaches 55.32. Another consistent observation of high model risk is in the years during and after the 2008 crisis, as we observe an increasing trend in risk ratios.

Figure 1: Maximum annual model risk: Across all institutions — 99% VaR

The plot displays the maximum daily risk ratio in a given year for each of the 439 financial institutions in our sample. The averages across all institutions along the 95% confidence intervals are presented. The risk ratio is the ratio of the highest to the lowest VaR estimate that is calculated at a 99% probability level based on 11 different models, including five mainstream models—historical simulation, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory—and six mixed models—historical simulation and extreme value theory applied to a GARCH filtered data under the assumption of normal, student-t, and skewed-t distributions.

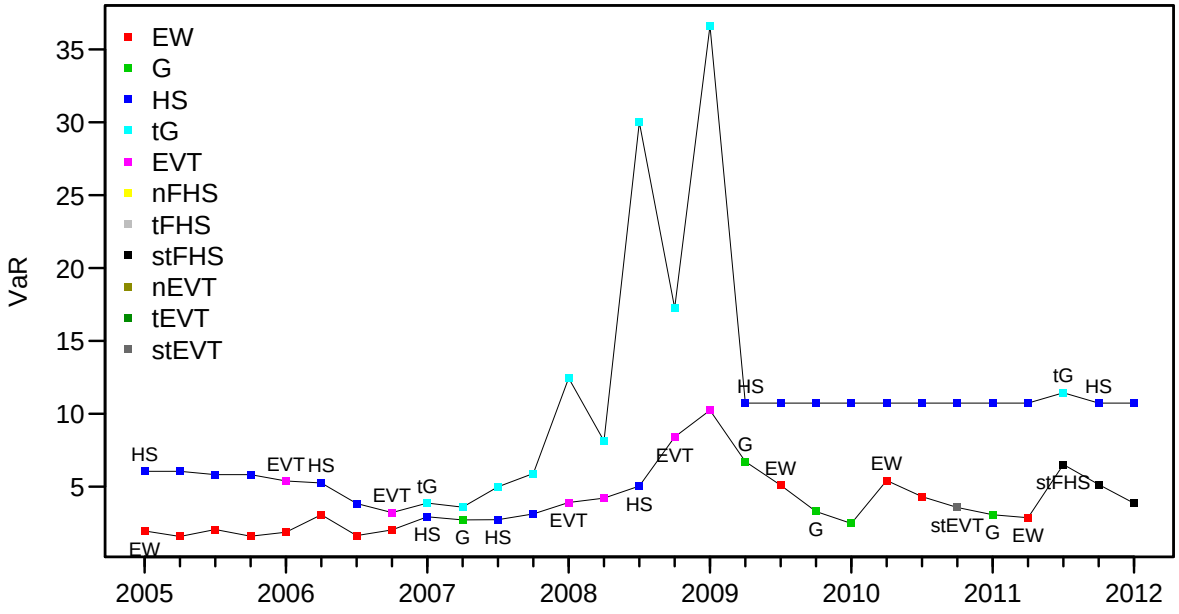


In order to paint a more detailed picture of model risk, Figure 2 shows the detailed results for the biggest stock in our sample in terms of asset size: JP Morgan. The figure shows end-of-quarter maximum and minimum VaR forecasts, along with the particular methods generating the readings.

There is no clear pattern among model outcomes. Generally, the highest observations tend to be generated by the fat-tailed methods (tG and EVT), whereas the thin-tailed methods dominate the low risk readings (EWMA and G). HS is represented in both the maxima and

Figure 2: End-of-quarter model risk for JP Morgan around the crisis

The highest and the lowest 99% daily VaR forecasts for JP Morgan based on 11 different models, including five mainstream models—historical simulation (HS), exponentially weighted moving average (EW), normal GARCH (G), student-t GARCH (tG), and extreme value theory (EVT)—and six mixed models—HS and EVT applied to a GARCH filtered data under the assumption of normal, student-t, and skewed-t distributions (nFHS, tFHS, stFHS, nEVT, tEVT, stEVT). Estimation window is 1,000 days. To minimize clutter, end-of-quarter results are plotted. Every time the VaR method changes, the label changes. Portfolio value is $100.



minima and EVT sometimes produces the lowest forecasts. The hybrid models sometimes have the lowest risk forecasts and while not visible in the end-of-quarter numbers in the plot, occasionally the highest.

3.1 Focus on Basel II and III

Table 1 presents the maximum daily risk ratios across the NBER recession dates, the stock market crashes of 1977 and 1987, and the 1998 LTCM/Russian crisis. We present the results from a small number of stocks for illustrative purposes, with full results in the web appendix. We consider the largest depository, JP Morgan (JPM), non-depository, American Express (AXP), insurance, American International Group (AIG), and broker-dealer, Goldman Sachs (GS) in the sample.[2] In addition, in order to study the model risk of the overall system, we

[2]Metlife and Prudential are the largest and second largest insurance companies in our sample in terms of asset size, respectively. However, we present the results for the American International Group (AIG), which is the third largest insurance company in the sample because both Metlife and Prudential have available observations only after 2000.

employ the daily returns of the S&P 500 index and the Fama-French value-weighted financial industry portfolio (FF). In addition, we create a financial equity portfolio, Fin100, by assuming that an investor holds the 100 biggest financial institutions in her portfolio. The portfolio is rebalanced annually and the random weights are based on the market capitalization of each stock at the beginning of the year.

Panel 1(a) shows the results when risk is calculated via daily 99% VaR, in line with the Basel II market risk regulations. We find that risk ratios across the entire time period, range from 1.71 to 1.82 for the portfolios and from 1.85 to 2.16 for the individual stocks we consider, suggesting that model risk is generally quite moderate throughout the sample period. A clearer picture emerges by examining the maximum risk ratios across the various subsamples. Model risk remains quite temperate during economic recessions, but increases substantially during periods of financial turmoil, exceeding 10 during the 1987 crash and 5 during the 2008 global crisis for the market portfolio.

The Basel committee has proposed a number of changes to the existing market risk capital accords, most importantly changing the core measure from 99% VaR to 97.5% ES. In the first round of proposals in 2013, the Committee also proposed using 10-day overlapping estimation windows. In practice, this means that one would use the returns from days 1 to 10 as the first observation, days 2 to 11 for the second, and so forth. However, in the subsequent revision, the committee withdrew the overlapping proposal.

We consider both cases in Panels 1(b) and 1(c) from the point of view of model risk and find that switching to ES from VaR does not overcome the model disagreement. In particular, with 97.5% ES 10-day overlapping estimation windows, the model risk increases significantly, with the risk ratios during turmoil periods double for S&P 500, triple for FF, and quadruple for Fin100, on average.

We suspect the reason for the impact of the overlapping estimation windows on model risk is because of how observations are repeated. Not only will it introduce dependence in the underlying time series, but anomalous events will gain artificial prominence, as they are repeated 10 times in the sample. Both sources, in turn, may bias the estimation. Since different estimation methods react differently to these introduced artifacts, it is not surprising that model risk increases so sharply.

3.2 Sensitivity analysis

These results give rise to the question of whether any particular model is responsible for the highest or lowest risk forecasts systematically, and therefore driving the results. In order to

Table 1: Daily risk ratios: non-overlapping 99% VaR and overlapping 97.5% ES

This table reports the maximum of the ratio of the highest to the lowest daily VaR and ES forecasts (risk ratios) for the crises periods spanning from January 1974 to December 2012 for the S&P-500, Fama-French financial sector portfolio (FF), the value-weighted portfolio of the biggest 100 stocks in our sample (Fin100), JP Morgan (JPM), American Express (AXP), American International Group (AIG), and Goldman Sachs (GS). Panel 1(a) presents the risk ratio estimates where the risk is calculated via daily 99% VaR. In Panel 1(b) we calculate the risk ratios via 97.5% daily ES with 10–day overlapping estimation windows. Five mainstream models: historical simulation, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory, and six hybrid models: historical simulation and extreme value theory applied on a GARCH filtered data with normal, student-t, and skewed-t distributed error terms are employed to calculate the VaR and ES estimates. Estimation window size is 1,000 days. Finally, the last row of each panel reports the average risk ratio for the whole sample period.

(a) Basel II requirements: VaR, $p = 99\%$, non-overlapping

Event	Peak	Trough	SP-500	FF	Fin100	JPM	AXP	AIG	GS
1977 crash	1977-05	1977-10	2.65	3.16	3.20	3.39	4.30	13.02	
1980 recession	1980-01	1980-07	1.92	2.23	2.14	2.37	1.92	2.63	
1981 recession	1981-07	1982-11	2.14	2.17	2.35	2.96	2.88	2.99	
1987 crash	1987-10	1988-01	10.39	10.10	10.10	11.01	6.37	3.71	
1990 recession	1990-07	1991-03	2.06	2.26	2.30	3.77	2.17	1.82	
LTCM crisis	1998-08	1998-11	4.34	3.34	2.98	2.97	5.13	3.00	
2001 recession	2001-03	2001-11	1.96	2.48	2.45	2.31	2.12	2.80	
2008 recession	2007-12	2009-06	5.22	5.26	6.39	6.90	5.44	13.89	5.79
Full sample (ave.)	1974-01	2012-12	1.71	1.78	1.82	1.85	1.85	2.09	2.16

(b) Basel III 2013 proposal: ES, $p = 97.5\%$, 10–day overlapping

Event	Peak	Trough	SP-500	FF	Fin100	JPM	AXP	AIG	GS
1977 crash	1977-05	1977-10	4.79	12.38	5.42	5.93	6.69	3.71	
1980 recession	1980-01	1980-07	5.72	18.55	12.32	15.61	5.94	8.45	
1981 recession	1981-07	1982-11	7.40	10.90	16.51	21.64	5.59	10.46	
1987 crash	1987-10	1988-01	13.35	16.40	53.84	7.91	8.29	6.17	
1990 recession	1990-07	1991-03	10.45	13.24	19.12	7.05	4.99	17.48	
LTCM crisis	1998-08	1998-11	5.04	5.98	5.55	8.96	6.26	7.34	
2001 recession	2001-03	2001-11	4.88	3.82	3.99	4.32	5.24	3.75	
2008 recession	2007-12	2009-06	6.73	6.43	6.36	24.12	9.93	22.66	6.17
Full sample (ave.)	1974-01	2012-12	2.78	3.00	2.93	2.52	2.55	2.85	2.79

(c) Basel III 2014 proposal: ES, $p = 97.5\%$, non-overlapping

Event	Peak	Trough	SP-500	FF	Fin100	JPM	AXP	AIG	GS
1977 crash	1977-05	1977-10	2.56	3.21	3.30	3.38	4.26	16.23	
1980 recession	1980-01	1980-07	1.89	2.18	2.09	2.36	1.91	2.90	
1981 recession	1981-07	1982-11	2.13	2.19	2.35	3.04	2.95	3.25	
1987 crash	1987-10	1988-01	9.03	8.99	9.13	10.21	5.66	3.60	
1990 recession	1990-07	1991-03	2.65	2.48	2.37	3.50	2.24	2.16	
LTCM crisis	1998-08	1998-11	3.79	3.15	2.87	2.90	4.69	2.98	
2001 recession	2001-03	2001-11	1.82	2.59	2.57	2.28	2.07	2.90	
2008 recession	2007-12	2009-06	4.74	4.42	5.67	6.07	5.17	9.59	5.28
Full sample (ave.)	1974-01	2012-12	1.76	1.83	1.87	1.91	1.88	2.19	2.14

examine this eventuality, we study the sensitivity of our findings to any particular model by excluding them from the risk ratio analysis, one by one.[3] We focus on the S&P 500 index, but the results are similar for the other assets.

The results reported in Table 2 suggest that the divergence across models is not dependent on any particular model. The estimated risk ratios are almost invariant compared with the one that considers all 11 models. In other words, each of the models can, at different times, deliver the maximum or minimum risk forecasts.

Table 2: Sensitivity of daily risk ratios: non-overlapping 99% VaR

This table reports the maximum of the ratio of the highest to the lowest daily 99% VaR forecasts (risk ratios) when a particular model is excluded from the risk ratio analysis indicated by the heading column. Column two repeats the risk ratio estimates when all of the eleven models; historical simulation (HS), exponentially weighted moving average (EWMA), normal GARCH (G), student-t GARCH (tG), extreme value theory (EVT), HS and EVT applied to a GARCH filtered data with normal, student-t, and skewed-t distributed error terms (nFHS, tFHS, stFHS, nEVT, tEVT, stEVT) are employed. The risk forecasts are calculated for the period from January 1974 to December 2012 for the S&P-500 index. Estimation window size is 1,000 days. Finally, the last row of each panel reports the average risk ratio for the whole sample period.

Excluded Model	None	HS	EWMA	G	tG	nFHS	tFHS	stFHS	EVT	nEVT	tEVT	stEVT
Event												
1977 crash	2.65	2.49	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.64	2.65
1980 recession	1.92	1.92	1.92	1.92	1.92	1.90	1.92	1.92	1.92	1.92	1.92	1.92
1981 recession	2.14	2.13	2.14	2.14	1.86	2.14	2.14	2.14	2.14	2.14	2.14	2.14
1987 crash	10.39	9.62	10.39	10.39	10.39	10.39	10.39	10.39	10.39	10.08	10.39	10.39
1990 recession	2.06	2.05	1.99	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06
LTCM crisis	4.34	4.05	4.34	4.34	4.00	4.34	4.34	4.34	4.34	4.34	4.34	4.34
2001 recession	1.96	1.84	1.96	1.96	1.91	1.96	1.96	1.96	1.96	1.96	1.96	1.96
2008 recession	5.22	5.10	5.22	5.22	4.55	5.22	5.22	5.22	5.22	5.22	5.22	5.22
Full sample	1.71	1.69	1.61	1.70	1.64	1.70	1.71	1.71	1.71	1.71	1.71	1.71

We also examine the sensitivity of the results on the definition of model risk since one might employ different measures to quantify the disagreement among the alternative models other than risk ratio, such as standard deviation or absolute deviation of the VaR forecasts. Both measures take into account how much a model deviates from the average risk reading, and hence, can be considered as a plausible way to measure model risk. We find that all three measures of model risk are correlated over 80% and the main results hold irrespective of the definition.

[3] We exclude one model at a time, rather than exclude a group of models, because it would be unfeasible to report all combinations here. The total number of possible combinations is $\sum_{k=2}^{10} \binom{11}{k} = 2035$, where k is the number of models considered each time and 11 is the total number of models. The entire set of results is available on the web appendix so any combination of interest can be calculated.

3.3 Model risk and market conditions

Table 1 and Figure 1 reveal that while model risk is typically quite moderate, it sharply increases during crisis periods. Given that some of the risk measures we use are based on conditional volatilities, a part of the explanation is mechanical—whenever the volatility increases, a conditional historical volatility method, such as GARCH, will produce higher risk readings. More fundamentally, however, the results indicate that not the VaR readings, but the disagreement between those readings increases. All of the risk forecast models employed can be considered as valid candidates. Given that they have entered the canon based on their performance during non-crisis times, it is not surprising that they broadly agree at such periods; otherwise, any model that sharply disagreed, might have been dismissed. However, the models all treat history and shocks quite differently and therefore can be expected to differ when faced with a change in statistical regimes. Given that none of the methods systematically produces the highest or the lowest VaR estimates throughout the sample period, we surmise that this is what we pick up in our analysis.

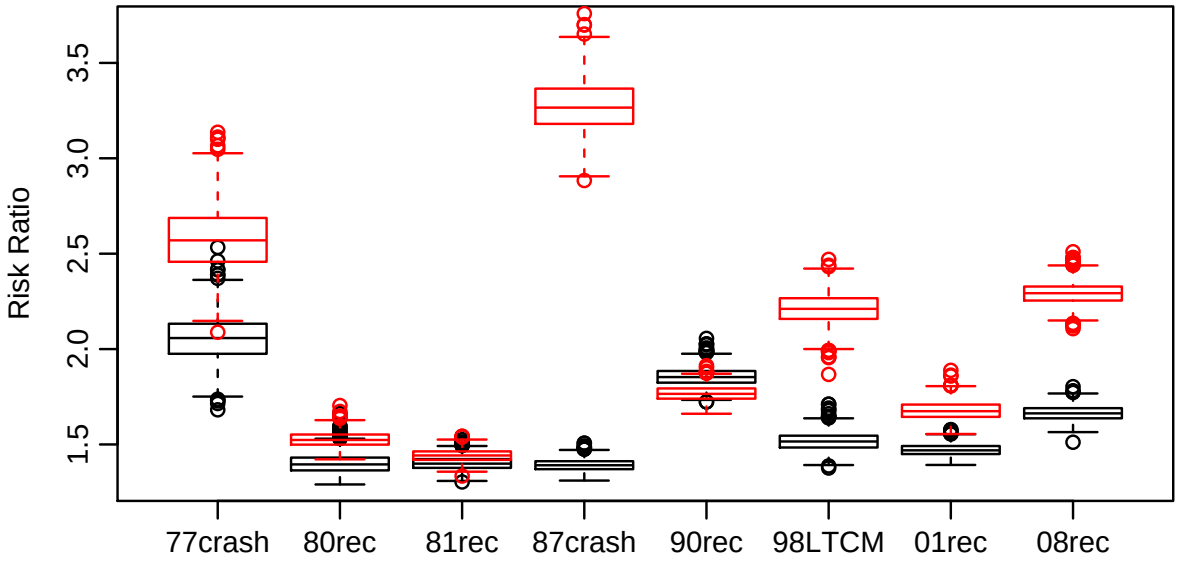
We investigate this by testing whether the difference between the model risk during crisis and immediate pre-crisis periods are statistically different from each other. To this end, we further adopt a variation of the portfolio bootstrap procedure of Hendricks (1996) to evaluate the statistical significances of risk ratios during different market conditions.[4] We assume that an investor holds the 100 biggest financial institutions in her portfolio. We also assume that the stocks in the portfolio are allowed to change at the beginning of each year and portfolio weights are random. We calculate the ratio of the highest to the lowest VaR estimates for the random portfolios employing 11 VaR approaches. The following algorithm illustrates the main steps:

1. Select the biggest 100 institutions in terms of market capitalization at the beginning of each year and obtain the daily holding period return for each stock.
2. For a given year, select a random portfolio of positions for the stocks selected in step (1) by drawing the portfolio weights from a unit-simplex. Hence, get the daily return of the random portfolio for the sample period.

[4]We also considered a stationary/block bootstrap procedure, finding similar results. However, the block bootstrap approach implicitly assumes no dependence across blocks, which can be avoided by employing the portfolio bootstrap approach. With the random weights, the value and the risk of a portfolio is random, producing different and incomparable VaR forecasts. However, our aim is not to compare the VaR of two portfolios. We rather compare the risk ratio of a given portfolio, which should be close to 1 if the model risk is negligible. As a consequence, we find that portfolio bootstrapping is a more suitable procedure for our purposes.

Figure 3: Model risk: Confidence intervals

The plot displays the first and third quartiles of risk ratios for the crisis and non-crisis periods separately between January 1974 and December 2012. The intervals for the crisis periods are plotted in red, whereas the pre-crisis periods are identified as black. The risk ratio is the ratio of the highest to the lowest VaR estimates of the simulated portfolio outlined in Section 3.3. Estimation window size is 1,000 and VaR estimates are calculated at a 99% probability level based on six different models: historical simulation, moving average, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory. Data are obtained from the CRSP 1925 US Stock Database.



3. Calculate the daily 99% VaR by employing each of the six candidate risk models for the random portfolio chosen in step (2) with an estimation window size of 1,000.
4. For a given day calculate the ratio of the highest to the lowest VaR readings (VaR risk ratios) across all models.
5. Repeat steps (2) through (4) 1,000 times. This gives a matrix of risk ratios with a dimension of number of days $\times$ number of trials.
6. Identify the crisis and pre-crisis periods. For a given episode, we consider the previous 12 months as a pre-crisis period. For instance, for the 2008 global financial crisis, with a peak on December 2007 and a trough on June 2009, the pre-crisis period covers from December 2006 to November 2007.
7. For each trial, obtain the time-series averages of risk ratios over the crisis and pre-crisis periods and calculate the confidence intervals.

Figure 3 plots the first and third quartiles of risk ratios for each of the episodes separately. The intervals for the crisis periods are plotted in red, whereas the pre-crisis periods are in black. For all of the periods, except the 1990 recession, we find that the risk ratios are higher

during crises compared to non-crises. Moreover, the difference is statistically significant for the 1987 crash, 1998 LTCM crisis, and 2008 global financial crisis.

4 Model risk of systemic risk models

Perhaps the most common way to construct a systemic risk model (SRM) is to adopt existing market risk regulation methodologies to the systemic risk problem, an approach we term *market data based methods*.[5] In this section, we examine the model risk of such models. To this end, we start by proposing a general setup for the classification of SRMs, and then we apply our risk ratio methodology to the most popular SRMs.

Our objective is not to document the degree of disagreement *across* systemic risk measures. Systemic risk measurements are very much in the early stages and use inconsistent definitions of systemic risk. Hence, it would not be surprising to see that they do not agree when measuring such risk. However, our paper focuses on a different question that has not been studied before—the disagreement of a given systemic risk measure if it is calculated based on different models, such as HS, GARCH, or EVT.

4.1 Classification of systemic risk measures

The various market data based SRMs that have been proposed, generally fall into one of three categories: the risk of an institution given the system, the risk of the system given the institution or the risk of the system or institution by itself. In order to facilitate the comparison of the various SRMs, it is beneficial to develop a formal classification scheme.

Let R_i be the risky outcome of a financial institution i on which the risk is calculated. This could be, for example, daily return risk of such an institution. Similarly, we denote the risky outcome of the entire financial system by R_S. We can then define the joint density of an institution and the system by

$$f\left(R_i, R_S\right).$$

The marginal density of the institution is then $f(R_i)$, and the two conditional densities are $f\left(R_i|R_S\right)$ and $f\left(R_S|R_i\right)$. If we then consider the marginal density of the system as a

[5] Besides the market data based methods, other approaches exist to construct SRMs, such as those based on credit risk techniques, market-implied losses, connectedness and macroeconomic conditions. See, for instance, Segoviano and Goodhart (2009), Huang et al. (2009), Alessi and Detken (2009), Borio and Drehmann (2009), Tarashev et al. (2010), Drehmann and Tarashev (2013), Gray and Jobst (2011), Huang et al. (2012), Suh (2012), Billio et al. (2012), Gray and Jobst (2013), and Bluhm and Krahnen (2014). However, given the preeminence of market data based methods amongst SRMs, that is where we focus our attention.

normalizing constant, we get the risk of the institution conditional on the system by Bayes' theorem:

$$f\left(R_i|R_S\right) \propto f\left(R_S|R_i\right) f(R_i). \tag{1}$$

The risk of the system conditional on the institution is similarly defined;

$$f\left(R_S|R_i\right) \propto f\left(R_i|R_S\right) f(R_S). \tag{2}$$

Suppose we use VaR as a risk measure. Defining Q as an event such that:

$$\text{pr}[R \leq Q] = p,$$

where Q is some extreme negative quantile and p the probability. Then, VaR equals $-Q$. Expected shortfall (ES) is similarly defined:

$$\text{ES} = \text{E}[R|R \leq Q].$$

Conditional VaR (CoVaR_i) is then obtained from (1) with VaR being the risk measure:[6]

$$\text{CoVaR}_i = \text{pr}[R_S \leq Q_S|R_i \leq Q_i] = p. \tag{3}$$

and if instead we use (2) and ES as a risk measure, we get marginal expected shortfall (MES):

$$\text{MES}_i = \text{E}[R_i|R_S \leq Q_S]. \tag{4}$$

We could just as easily have defined MVaR as

$$\text{MVaR}_i = \text{pr}[R_i \leq Q_i|R_S \leq Q_S] = p \tag{5}$$

and CoES as

$$\text{CoES}_i = \text{E}[R_S|R_i \leq Q_i]. \tag{6}$$

To summarize:

[6] Adrian and Brunnermeier (2011) identify an institution being under distress if its return is *exactly* at its Value–at–Risk (VaR) level rather than *at most* at its VaR.

Marginal risk measure	Condition on system	Condition on institution
	MVaR	CoVaR
VaR	$\text{pr}[R_i \leq Q_i \vert R_S \leq Q_S] = p$	$\text{pr}[R_S \leq Q_S \vert R_i \leq Q_i] = p$
	MES	CoES
ES	$\text{E}[R_i \vert R_S \leq Q_S]$	$\text{E}[R_S \vert R_i \leq Q_i]$

The Shapley value (SV) methodology falls under this classification scheme, by adding a characteristic function, which maps any subgroup of institutions into a measure of risk. The SV of an institution i is a function of a characteristic function θ and the system S. If we choose θ as VaR, then

$$SV_i = g(S, \theta) = g(S, \text{VaR}).$$

If the characteristic function is chosen as the expected loss of a subsystem given that the entire system is in a tail event, we end up with the same definition as MES. Similarly, the Co-Risk measure of IMF (2009) and systemic expected shortfall (SRISK) of Brownlees and Engle (2015); Acharya et al. (2012) also fall under this general classification system. SRISK is a function of MES, leverage, and firm size, where MES is calculated as in (4) with a DCC and TARCH model to estimate volatility. On the other hand, Co-Risk is similar in structure to CoVaR, except that it focuses the co-dependence between two financial institutions, rather than the co-dependence of an institution and the overall financial system. In other words, it depends on the conditional density of institution i given institution j and can be estimated via quantile regressions with market prices, specifically the CDS mid-prices being the input.

Ultimately, regardless of the risk measure or conditioning, the empirical performance of the market based systemic risk measures fundamentally depends on VaR. This applies equally whether the risk measure is directly based on VaR, like CoVaR, or indirectly, like MES. Hence, we expect the model risk of VaR to pass through to the model risk of MES and CoVaR.

4.2 Model risk of MES

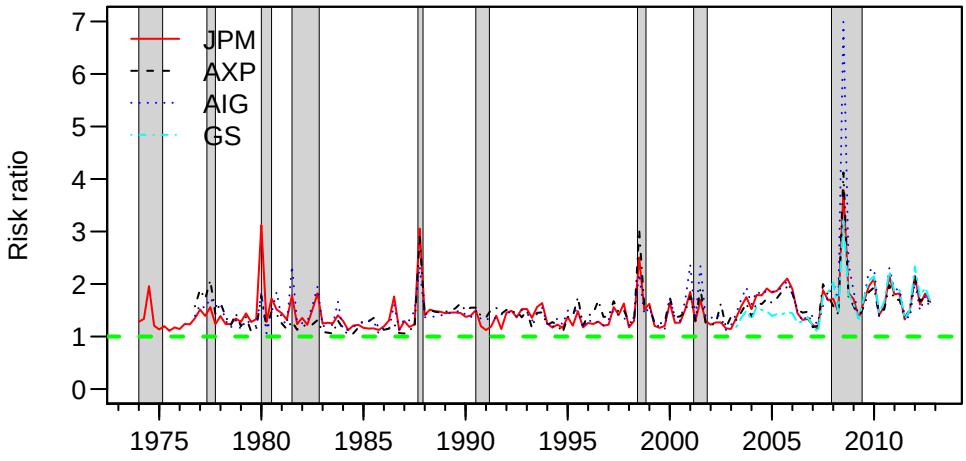
MES is defined as an institution's expected equity loss given that the system is in a tail event. Hence, it is an expected shortfall estimate modified to use a threshold from the overall system rather than the returns of the institution itself, and the first step requires calculation of VaR of the market portfolio. Following Acharya et al. (2010), we use a 95% probability level, with the S&P 500 as the market portfolio.

By using the same risk models previously employed for VaR and ES, we construct 11 MES forecasts for each day, one for each risk forecast model. We then finally calculate daily risk ratios across the risk readings.

Figure 4 illustrates the end-of-quarter risk ratios for the same four companies introduced in Section 3.1. The NBER recession dates, the stock market crashes of 1977 and 1987, and the 1998 LTCM/Russian crisis are marked with gray shades to visualize the trends in model risk during the turmoil periods. The results are in line with those for VaR, as presented in Table 1. Model risk remains low most of the time, but spikes during periods of market turmoil.

Figure 4: MES model risk

Ratio of the highest to the lowest daily 95% MES estimates for JP Morgan (JPM), American Express (AXP), American International Group (AIG), and Goldman Sachs (GS). The S&P 500 index is used as market portfolio. To calculate the system–VaR estimates, we employ five mainstream methods: historical simulation, exponentially weighted moving average, normal GARCH, student-t GARCH, and extreme value theory. In addition, we consider six hybrid models: historical simulation and extreme value theory applied to a GARCH filter under the assumptions of normal, student-t, and skewed-t distributed error terms. Estimation window size is 1,000. To minimize clutter, end-of-quarter results are plotted. Data are obtained from the CRSP 1925 US Stock Database. The NBER recession dates, the stock market crashes of 1977 and 1987, and the LTCM/Russian crisis are marked with gray shades.



Note that, in general, MES risk ratios presented in Figure 4 are closer to 1 than the VaR ratios. This is because one gets much more accurate risk forecasts in the center of the distribution compared with the tails, and therefore 95% risk forecasts are more accurate than 99% risk forecasts. The downside is that a 95% daily probability is an event that happens more than once a month. This highlights a common conclusion, it is easier to forecast risk for non-extreme events than extreme events and the less extreme the probability is, the better the forecast. That does not mean that one should therefore make use of a non-extreme probability, because the probability needs to be tailored to the ultimate objective for the risk forecast.

4.3 Model risk of CoVaR and ΔCoVaR

The other market based systemic risk measure we study in detail is CoVaR (Adrian and Brunnermeier, 2011). The CoVaR of an institution is defined as the VaR of the financial system given that the institution is under financial distress, and ΔCoVaR captures the marginal contribution of a particular institution to the systemic risk.

While Adrian and Brunnermeier (2011) estimate CoVaR by means of a quantile regression method (see Appendix B for details), one can estimate it with all of the methods considered in this study, with the exception of risk forecast models based on historical simulation. Those are quite easy to implement, but require at least $1/0.01^2 = 10,000$ observations at the 99% level. For this reason, the risk ratio results for CoVaR will inevitably be biased towards one.

If one defines an institution being under distress when its return is *at most* at its VaR level, rather than being *exactly* at its VaR, then CoVaR is defined as:[7]

$$\mathrm{pr}[R_S \leq \mathrm{CoVaR}_{S|i} \,|R_i \leq \mathrm{VaR}_i] = p.$$

It is then straightforward to show that:

$$\int_{-\infty}^{\mathrm{CoVaR}_{S|i}} \int_{-\infty}^{\mathrm{VaR}_i} f(x,y)dxdy = p^2. \tag{7}$$

Hence, CoVaR can be estimated under any distributional assumption by solving (7). Girardi and Ergun (2013) estimate CoVaR under normal GARCH and Hansen's (1994) skewed-t distribution. In addition, we extend this analysis to EWMA, student-t GARCH, EVT, and a GARCH filter with an EVT approach using the assumptions of normal, student-t, and skewed-t distributed error terms. We then compare the risk forecasts produced by these models. We model the correlation structure with Engle's (2002) DCC model and obtain CoVaR by numerically solving (7). The EVT application was based on using EVT for the tails and an extreme value copula for the dependence structure.

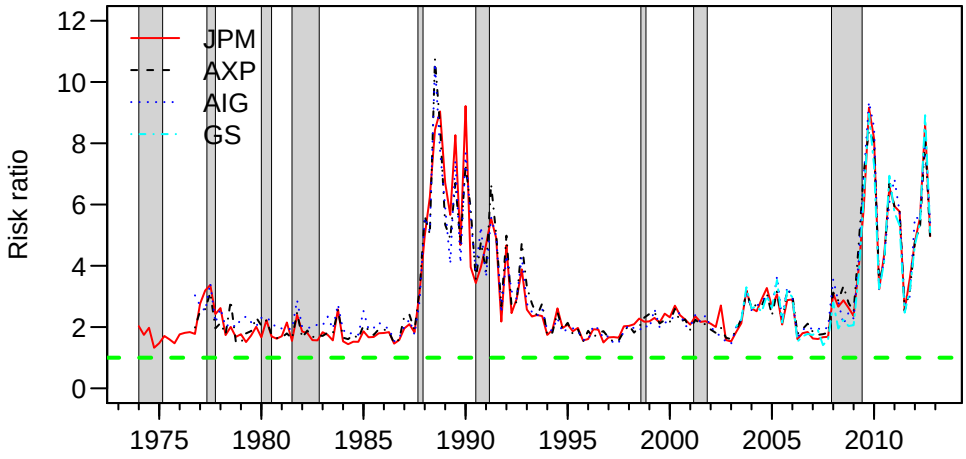
Figure 5 illustrates the end-of-quarter risk ratios for the same four companies. Similarly, recession and crisis periods are marked with gray shades to visualize the trends in model risk. We find that the model risk of CoVaR is higher on average compared with the model

[7] Mainik and Schaanning (2012) and Girardi and Ergun (2013) estimate the dynamics of CoVaR under the conditioning event $R_i \leq \mathrm{VaR}_i$. Their results show that the resulting CoVaR is not significantly different from the original CoVaR analysis proposed by Adrian and Brunnermeier (2011) conditioned on $R_i = \mathrm{VaR}_i$. This suggests that without loss of generality, one can condition the CoVaR measure on $R_i \leq \mathrm{VaR}_i$ rather than on $R_i = \mathrm{VaR}_i$, and yet it allows us to estimate the CoVaR under different distributional assumptions.

risk of VaR and MES, especially after the 2008 period. In line with other results, it increases sharply with market turmoil.

Figure 5: CoVaR model risk

Ratio of the highest to the lowest daily 99% CoVaR estimates for JP Morgan (JPM), American Express (AXP), American International Group (AIG), and Goldman Sachs (GS). The Fama-French value-weighted financial industry portfolio index is used as the market portfolio. Seven different methods—exponentially weighted moving average, normal GARCH, student-t GARCH, EVT, and mixed models that combine EVT with a GARCH filter under the assumptions of normal, student-t, skewed-t distributions—are employed to calculate the individual stock VaR estimates. CoVaR is estimated by numerically integrating (7). Estimation window size is 1,000. To minimize clutter, end-of-quarter results are plotted. Data are obtained from the CRSP 1925 US Stock Database. The NBER recession dates, the stock market crashes of 1977 and 1987, and the LTCM/Russian crisis are marked with gray shades..



We also investigate the statistical properties of the CoVaR measure, as well as the ΔCoVaR measure, estimated by the quantile regression methods of Adrian and Brunnermeier (2011). The results are reported in Appendix B. First, we find that the unconditional correlation between VaR and ΔCoVaR mostly exceeds 99%, suggesting that the scaled signal provided by ΔCoVaR is very similar to the signal provided by VaR. Second, we show that when the estimation noise in the quantile regression is carried through to the ΔCoVaR estimates, it is hard to significantly discriminate between different financial institutions based on ΔCoVaR.

5 Analysis

Our findings indicate significant levels of model risk in the most common risk forecast methods, affecting applications of both market risk regulatory models (MRRMs) and systemic risk measures (SRMs). Importantly, we find that model risk increases along with financial market stress.

This result does not necessarily imply that the performance of the models deteriorate during a financial crisis. All it says is that the disagreement among models, captured by risk ratios, increases significantly during market turmoil. From a practical point of view, anyone who relies on risk forecast models for decision-making will be faced with a less precise signal at such times than normally is the case.

That might not be all that is bothersome for many users, especially those engaged in preventative measures and hedging. For others, an important part of their job is crisis response, identifying risk in real time and formulating short-term measures to protect their particular financial institution or the system at large. Similarly, especially after the 2008 crisis, the authorities formulating post-crisis regulatory design heavily depend on risk models.

In our view, therefore, the disagreement among the models and, hence, the risk ratio analysis is a valuable resource to both private market participants and macro and micro prudential regulators.

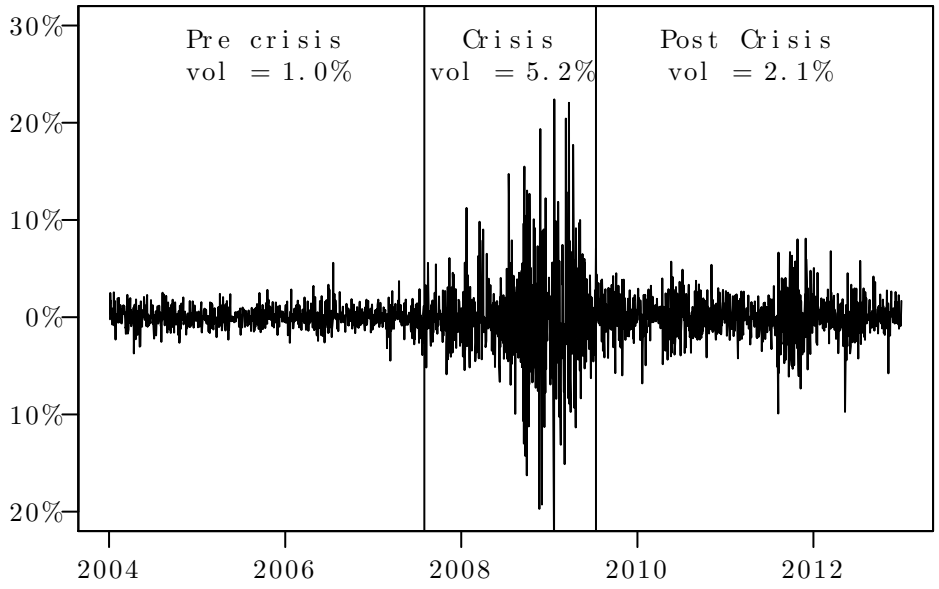
We suspect there are two main reasons for high model risk: the low frequency of financial crises and the presence of endogenous risk. Perhaps the main problem in systemic risk forecasting is the low frequency of financial crises. While fortunate from a social point of view, it causes significant difficulties for any empirical analysis. For instance, a typical OECD country suffers a banking crisis once every 35 years (Danielsson, Valenzuela and Zer, 2015b). Therefore, the empirical analyst has to make use of data from non-crisis periods to impute statistical inference on the behavior of financial markets during crises.

The challenge in building an empirical systemic risk model is therefore capturing the risk of an event that has almost never happened using market variables during times when not much is going on. In order to do so, one needs to make stronger assumptions about the stochastic process governing market prices, assumptions that may not hold as the economy transits from a calm period to a turmoil period. At the very least, this implies that a reliable method would need to consider the transition from one state of the world to another. It requires a leap of faith to believe that price dynamics during calm periods have much to say about price dynamics during crises, especially when there is no real crisis with which to compare the forecast. Ultimately this implies that from a statistical point of view, the financial system may transit between distinct stochastic processes, frustrating modeling.

As an illustration, Figure 6 plots the time series of JP Morgan returns. Visual identification shows the presence of three distinct regimes, where the volatility and extreme outcomes before the crisis do not seem to indicate the potential for future crisis events, and similarly, data during the crisis would lead to the conclusion that risk is too high after the crisis. In other words, if one were to estimate a model that does not allow for structural breaks, one is likely

to get it wrong in all states of the world; risk assessments would be too low before the crisis and too high after the crisis.

Figure 6: Daily JP Morgan returns before, during, and after the 2008 crisis, along with daily volatility.



The second reason for the rather high levels of model risk witnessed here is how risk arises in practice. Almost all risk models assume risk is exogenous, in other words that adverse events arise from the outside. However, in the language of Danielsson and Shin (2003), risk is endogenous, created by the interaction between market participants. Because market participants have an incentive to undermine any extant rules aimed at controlling risk-taking and hence, take risk in the least visible way possible, risk-taking is not observable until the actual event is realized. In the words of the former head of the BIS, Andrew Crockett (2000):

> "The received wisdom is that risk increases in recessions and falls in booms. In contrast, it may be more helpful to think of risk as increasing during upswings, as financial imbalances build up, and materializing in recessions."

6 Conclusion

Risk forecasting is a central element of financial decision-making, the control of risk-taking, and macro and micro prevention policy. Such forecasting, necessarily depends on statistical models, and our results indicate that the degree of model risk of such models is quite high. The fundamental problem of model risk in any risk model such as VaR arises because risk cannot

be measured but has to be estimated by the means of a statistical model. Many different candidate statistical models have been proposed where one cannot robustly discriminate among them. Therefore, with a range of different plausible models one obtains a range of risk readings, and their disagreement provides a succinct measure of model risk.

We propose a method, termed risk ratios, for the estimation of model risk. Our results indicate that, model risk is low during times of no financial distress. In other words, the various candidate statistical models roughly provide the same risk forecasts. However, risk forecast models are subject to significant model risk during periods of financial distress, which are, unfortunately when they are most needed. This is a cause for concern because under high model risk, risk readings do not coincide, obstructing risk inference.

High model risk casts a doubt on appositeness of market data based SRMs and MRRMs to macroprudential policy making. After all, policymakers would like to use their outputs for important purposes; perhaps to determine capital for systematically important institutions, or in the design of financial regulations. However, our analysis shows that the risk readings depend on the model employed, so it is not possible to accurately conclude which institution is (systemically) riskier than the other.

Our results suggest that risk readings should be interpreted and evaluated with caution since they may lead to costly decision mistakes. Point forecasts are not sufficient. Confidence intervals incorporating the uncertainty from a particular model should be provided along with any point forecasts, analyzing for robustness and model risk. Recently some policy authorities, such as the European Banking Authority, have moved in this direction by emphasizing the need for confidence intervals conditional on a specific model, but still not capturing the risk across models.

Finally, from a prudential policymaker's perspective, it would be of interest to assess the forecasting performance of a systemic risk measure – when estimated using different underlying models – over economic indicators. Thus in a possible extension of future work, one could employ indicators such as the Chicago Fed National Activity Index (CFNAI) or NBER recession dates and different underlying risk forecast models, to examine whether a systemic risk measure delivers better out-of-sample forecasts compared to historical averages.

A Risk forecasting models

We employ five mainstream VaR forecast models: historical simulation (HS), exponentially weighted moving average (EWMA), normal GARCH (G), student-t GARCH (tG), and extreme value theory (EVT). We also include six hybrid models that have been identified in the literature as high quality: HS and EVT methods applied to the residuals of a GARCH model, under the assumptions of normal, student-t, and skewed-t distribution.

Historical simulation is the simplest non-parametric method to forecast risk. It employs the p^{th} quantile of historical return data as the VaR estimate. The model does not require an assumption regarding the underlying distribution on asset returns. However, it relies on the assumption that returns are independent and identically distributed. Moreover, it gives the same importance to all returns, ignoring structural breaks and clustering in volatility.

For the next three models—EWMA, G, and tG—VaR is calculated as follows:

$$\text{VaR}(p)_{t+1} = -\sigma_t F_R^{-1}(\vec{\theta})\vartheta, \tag{A.1}$$

where σ_t is the time-dependent return volatility at time t, $F_R(\cdot)$ is the distribution of standardized simple returns with a set of parameters $\vec{\theta}$, and ϑ is the portfolio value. Hence, these approaches require a volatility estimate and distributional assumptions.

One of the simplest ways to estimate the time-varying volatility is the EWMA model, which modifies the standard moving average model by applying exponentially decaying weights into the past. Under the assumption that the error terms are conditionally normally distributed, $F_R(\cdot)$ represents the standard normal cumulative distribution $\Phi(\cdot)$. The volatility is calculated as:

$$\hat{\sigma}^2_{\text{EWMA},t+1} = (1-\lambda)y_t^2 + \lambda\hat{\sigma}^2_{\text{EWMA},t},$$

where λ is the decay factor set to 0.94 as suggested by J.P. Morgan for daily returns (J.P. Morgan, 1995).

In addition, we estimate the volatility by employing a standard GARCH(1,1) model under the assumption that error terms are both normally and student-t distributed. We denote the former model as normal GARCH (G) and the latter as the student-t distribution GARCH (tG):

$$\hat{\sigma}^2_{G,t+1} = \omega + \alpha y_t^2 + \beta\sigma^2_{G,t}.$$

The degrees of freedom parameter for the student-t distribution GARCH (tG) is estimated through a maximum-likelihood estimation.

Another fat-tailed model, extreme value theory (EVT), is included in our analysis under the assumption that the tails are asymptotically Pareto distributed:

$$F(x) \approx 1 - Ax^{-\iota}$$

where A is a scaling constant whose value is not needed for VaR and ι the tail index estimated by maximum likelihood (Hill, 1975):

$$\frac{1}{\hat{\iota}} = \frac{1}{q}\sum_{i=1}^{q} \log \frac{x_{(i)}}{x_{(q-1)}},$$

where q is the number of observations in the tail. The notation $x_{(i)}$ indicates sorted data. We follow the VaR derivation in Danielsson and de Vries (1997):

$$\text{VaR}(p) = x_{(q-1)} \left(\frac{q/T}{p} \right)^{1/\hat{\iota}}.$$

ES is then:

$$\text{ES}(p) = \text{VaR}\, \frac{\hat{\iota}}{\hat{\iota} - 1}.$$

Finally, we include two groups of hybrid models that combine HS and EVT with a GARCH filter, respectively. We estimate VaR as follows:

$$\text{VaR}(p)_{t+1} = \hat{\mu} + \hat{\sigma}_t Q^{\eta}$$

where $\hat{\mu}$ is the unconditional mean and $\hat{\sigma}_t$ is the one-ahead forecast conditional volatility. We estimate $\hat{\sigma}_t$ using a GARCH(1,1) model assuming that the error terms (ε_t) have a normal, student-t, and skewed-t distribution. We compute Q^{η} by applying the HS or EVT models to the standardized residuals $\eta_t = \dfrac{\varepsilon_t}{\hat{\sigma}_t}$.

B CoVaR

Following Adrian and Brunnermeier (2011) for stock i and the system S, we estimate the time-varying CoVaR via quantile regressions:

$$\begin{aligned} R_{t,i} &= \alpha_i + \gamma_i M_{t-1} + \varepsilon_{t,i} && \text{(B.1)} \\ R_{t,S} &= \alpha_{S|i} + \beta_{S|i} R_{t,i} + \gamma_{S|i} M_{t-1} + \varepsilon_{t,S|i}, && \text{(B.2)} \end{aligned}$$

where R is defined as the growth rate of marked-valued total assets. The overall financial system portfolio $R_{t,S}$ is the weighted average of individual stock $R_{t,i}$s, where the lagged market value of assets is used as a weight. Finally, M denotes the set of state variables that are listed in detail below.

By definition, VaR and CoVaR are obtained by the predicted values of the quantile regressions:

$$\begin{aligned} \text{VaR}_{t,i} &= \widehat{\alpha}_i + \widehat{\gamma}_i M_{t-1} && \text{(B.3)} \\ \text{CoVaR}_{t,i} &= \widehat{\alpha}_{S|i} + \widehat{\beta}_{S|i}\, \text{VaR}_{t,i} + \widehat{\gamma}_{S|i} M_{t-1}. \end{aligned}$$

The marginal contribution of an institution, ΔCoVaR, is defined as:

$$\Delta\text{CoVaR}_{t,i}(p) = \widehat{\beta}_{S|i}\left[\text{VaR}_{t,i}(p) - \text{VaR}_{t,i}(50\%)\right]. \tag{B.4}$$

In order to calculate CoVaR estimates, we collapse daily market value data to a weekly frequency and merge it with quarterly balance sheet data from the CRSP/Compustat Merged quarterly dataset. Following Adrian and Brunnermeier (2011), the quarterly data are filtered to remove leverage and book-to-market ratios less than zero and greater than 100, respectively.

We start our analysis by considering the time series relationship between ΔCoVaR and VaR. ΔCoVaR is defined as the difference between CoVaR conditional on the institution being in distress and CoVaR calculated in the median state of the same institution. Given that the financial returns are almost symmetrically distributed, VaR calculated at 50% is almost equal to zero. Our empirical investigation confirms this observation; we find that the unconditional correlation between VaR and ΔCoVaR mostly exceeds 99%. This suggests that the scaled signal provided by ΔCoVaR is very similar to the signal provided by VaR.

On the other hand, in a cross-sectional setting, in what is perhaps their key result, Adrian and Brunnermeier (2011) find that even if the VaR of two institutions is similar, their ΔCoVaR can be significantly different, implying that the policy maker should consider this while forming policy regarding institutions' risk.

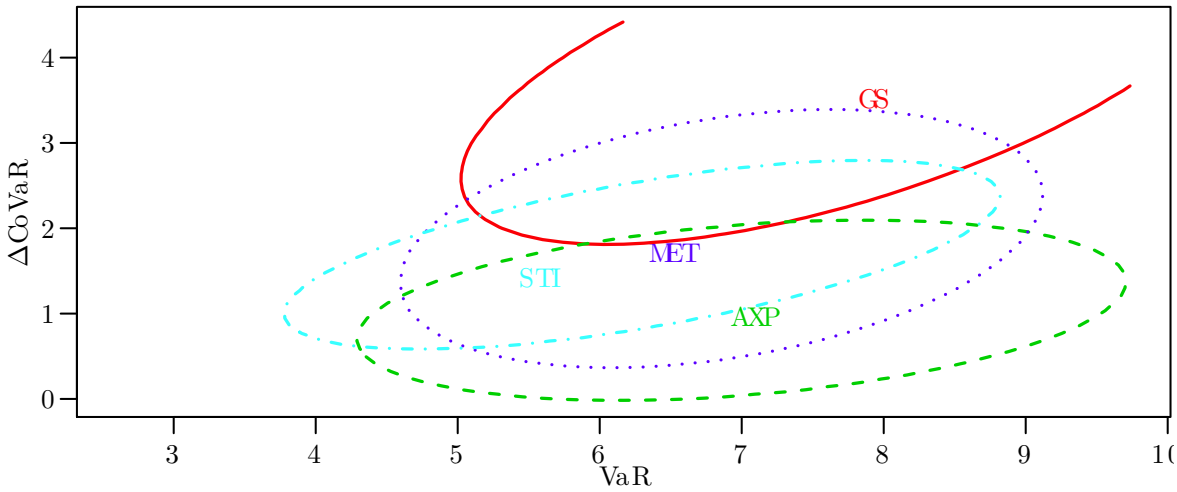
In order to get the idea of the model risk embedded in this estimation, we employ a bootstrapping exercise. For each of the stocks we re-run the quantile regressions 1,000 times by reshuffling the error terms and estimate VaR, CoVaR, and ΔCoVaR for each trial. Figure B.1 shows 99% confidence intervals of the bootstrapped estimates along with the point estimates. An institution's ΔCoVaR is plotted on the y-axis and its VaR on the x-axis, estimated as of 2006Q4 at a 1% probability level. For ease of presentation, we present the confidence intervals for Goldman Sachs (GS), American Express (AXP), Metlife (MET), and Suntrust Banks (STI). The point estimates show that there is a considerable difference between VaR and ΔCoVaR cross–sectionally, confirming the results of Figure 1 in Adrian and Brunnermeier (2011). For instance, although the VaR estimate of Goldman Sachs is comparable with its peers, its contribution to systemic risk, ΔCoVaR, is the highest. However concluding that Goldman Sachs is the systemically riskiest requires substantial caution since the confidence intervals overlap in quite wide ranges.

The following set of state variables (M) are included in the time–varying CoVaR analysis:

1. *Chicago Board Options Exchange Market Volatility Index (VIX):* Captures the implied volatility in the stock market. Index is available on the Chicago Board Options Exchange's website.
2. *Short–term liquidity spread:* Calculated as the difference between three-month U.S. repo rate and three-month US Treasury bill rate. The former is available in Bloomberg since 1991, and the latter is from the Federal Reserve Board's H.15 release.
3. The change in the three-month Treasury bill rate.

Figure B.1: 99% confidence intervals

99% confidence intervals of the 1,000 bootstrapped quantile regressions outlined in (B.3). VaR is the 1% quantile of firm asset returns, and ΔCoVaR is the marginal contribution of an institution to the systemic risk. The confidence intervals of Goldman Sachs (GS), Metlife (MET), Suntrust Banks (STI), and American Express (AXP) are presented. Portfolio value is equal to $100. Stock data are obtained from the CRSP 1925 US Stock and CRSP/Compustat Merged databases.



4. *Credit spread change:* Difference between BAA-rated corporate bonds from Moody's and 10-year Treasury rate, from the H.15 release.

5. *The change in the slope of the yield curve:* The change in difference of the yield spread between the 10-year Treasury rate and the three-month bill rate.

6. S&P 500 returns as a proxy for market return.

7. Real estate industry portfolio obtained from Kenneth French's website.

References

Acharya, V. V., R. Engle and M. Richardson, "Capital Shortfall: A New Approach to Ranking and Regulating Systemic Risk," *American Economic Review* 102 (2012), 59–64.

Acharya, V. V., L. H. Pedersen, T. Philippon and M. Richardson, "Measuring Systemic Risk," (May 2010), Working Paper.

Adrian, T. and M. K. Brunnermeier, "CoVaR," (2011), Working Paper, NBER–17454.

Alessi, L. and C. Detken, "Real Time Early Warning Indicators for Costly Asset Price Boom/Bust Cycles," ECB Working Paper Series 1039, European Central Bank, 2009.

Alexander, C. and M. Sarabia, J., "Quantile Uncertainty and Value-at-Risk Model Risk," *Risk Analysis* 32 (2012), 1293–1308.

Bali, T. G. and P. Theodossiou, "A Conditional-SGT-VaR approach with alternative GARCH models," *Annals of Operations Research* 151 (2007), 241–267.

Bao, Y., T. Lee and B. Saltoglu, "Evaluating VaR Models in Emerging Markets: A Reality Check," *Journal of Forecasting* 25 (2006), 101–128.

Basel Committee on Banking Supervision, "Overview of the amendment to the capital accord to incorporate market risk," Technical Report, Basel Committee on Banking Supervision, 1996.

———, "Fundamental Review of the Trading Book: A Revised Market Risk Framework," Technical Report, Basel Committee on Banking Supervision, 2013.

———, "Fundamental Review of the Trading Book: Outstanding Issues," Technical Report, Basel Committee on Banking Supervision, 2014.

Bauwens, L. and S. Laurent, "A New Class of Multivariate Skew Densities, With Application to Generalized Autoregressive Conditional Heteroscedasticity Models," *Journal of Business & Economic Statistics* 23 (2005), 346–354.

Bekiros, S. D. and D. A. Georgoutsos, "Estimation of Value-at-Risk by extreme value and conventional methods: a comparative evaluation of their predictive performance," *Int. Fin. Markets, Inst. and Money* 15 (2005), 209–228.

Berkowitz, J. and J. O'Brien, "How Accurate Are Value-at-Risk Models at Commercial Banks?," *Journal of Finance* 57 (2002), 977–987.

Billio, M., M. Getmansky, W. Lo, A. and L. Pelizzon, "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors," *Journal of Financial Economics* 104 (2012), 535–559.

Bluhm, M. and P. Krahnen, J., "Systemic risk in an interconnected banking system with endogenous asset markets," *Journal of Financial Stability* 13 (2014), 75–94.

Borio, C. and M. Drehmann, "Assessing the Risk of Banking Crises–Revisited," BIS Quarterly Review, pp. 29–46, Bank of International Settlements, 2009.

Boucher, M., C., J. Danielsson, S. Kouontchou, P. and B. Maillet, B., "Risk Models–at–Risk," *Journal of Banking and Finance* 44 (2014), 72–92.

Brownlees, C. T. and R. Engle, "SRISK: A Conditional Capital Shortfall Measure of Systemic Risk," (October 2015), Working Paper, NYU Stern School of Business.

Brownlees, C. T. and G. M. Gallo, "Comparison of Volatility Measures: a Risk Management Perspective," *Journal of Financial Econometrics* 8 (2009), 29–56.

Brunnermeir, K., M. and Y. Sannikov, "A Macroeconomic Model with a Financial Sector," *American Economic Review* 104 (2014), 379?–421.

Cont, R., "Model uncertainty and its impact on the pricing of derivative instruments," *Mathematical Finance* 16 (2006), 519–547.

Crockett, A., "Marrying the micro- and macro-prudential dimensions of financial stability," (2000), the General Manager of the Bank for International Settlements; `http://www.bis.org/review/rr000921b.pdf`.

Danielsson, J., "The Emperor has no Clothes: Limits to Risk Modelling," *Journal of Banking and Finance* 26 (2002), 1273–1296.

Danielsson, J. and C. G. de Vries, "Tail index and quantile estimation with very high frequency data," *Journal of Empirical Finance* 4 (1997), 241–257.

Danielsson, J., K. James, M. Valenzuela and I. Zer, "Can we prove a bank guilty of creating systemic risk? A minority report," (2015a), London School of Economics Working Paper.

Danielsson, J. and J. Morimoto, "Forecasting Extreme Financial Risk: A Critical Analysis of Prac- tical Methods for the Japanese Market," (2000), iMES Discussion Paper Series, E-8.

Danielsson, J. and H. S. Shin, "Endogenous Risk," in *Modern Risk Management — A History* (Risk Books, 2003), `http://www.RiskResearch.org`.

Danielsson, J., M. Valenzuela and I. Zer, "Learning from History: Volatility and Financial Crises," (2015b), London School of Economics Working Paper.

Drehmann, M. and N. Tarashev, "Measuring the Systemic Importance of Interconnected Banks," *Journal of Financial Intermediation* 22 (2013), 586–607.

Engle, R., "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business and Economic Statistics* 20 (2002), 339 – 350.

Gibson, R., ed., *Model Risk: Concepts, Calibration and Pricing* (Risk Books, 2000).

Girardi, G. and A. Ergun, T., "Systemic risk measurement: Multivariate GARCH estimation of CoVaR," *Journal of Banking and Finance* 37 (2013), 3169–3180.

Glasserman, P. and X. Xu, "Robust risk measurement and model risk," *Quantitative Finance* 14 (2013), 29–58.

Gray, D. and A. Jobst, *Systemic contingent claims analysis–Estimating potential losses and implicit government guarantees to the financial sector* (London: Edward Elgar, 2011), 143–185.

———, "Systemic Contingent Claims Analysis–Estimating Market-Implied Systemic Risk," Technical Report, IMF, 2013, iMF, WP/13/54.

Green, T., C. and S. Figlewski, "Market Risk and Model Risk for a Financial Institution Writing Options," *Journal of Finance* 54 (1999), 1465–1499.

Hansen, B. E., "Autoregressive Conditional Density Estimation," *International Economic Review* 35, (3) (1994), 705–30.

Hendricks, D., "Evaluation of Value-at-Risk Models Using Historical Data," Technical Report, FRBNY Economic Policy Review, April 1996.

Hill, B. M., "A simple general approach to inference about the tail of a distribution," *Annals of Statistics* 35 (1975), 1163–1173.

Huang, X., H. Zhou and H. Zhu, "A Framework for Assessing the Systemic Risk of Major Financial Institutions," *Journal of Banking and Finance* 33 (2009), 2036–2049.

———, "Assessing the Systemic Risk of a Heterogeneous Portfolio of Banks during the Recent Financial Crisis," *Journal of Financial Stability* 8 (2012), 193–205.

Hull, J. and W. Suo, "A Methodology for Assessing Model Risk and Its Application to the Implied Volatility Function Model," *Journal of Financial and Quantitative Analysis* 37 (2002), 297–318.

IMF, "Assessing the Systemic Implications of Financial Linkages," Technical Report, International Monetary Fund, April 2009.

J.P. Morgan, *RiskMetrics-technical manual*, third edition (1995).

Kuester, K., S. Mittnik and S. Paolella, M., "Value-at-Risk Prediction: A Comparison of Alternative Strategies," *Journal of Financial Econometrics* 4 (2006), 53–89.

Mainik, G. and E. Schaanning, "On dependence consistency of CoVaR and some other systemic risk measures," (August 2012), Working Paper.

Marimoutou, V., B. Raggad and A. Trabelsi, "Extreme Value Theory and Value at Risk: Application to oil market," *Energy Economics* 31 (2009), 519–530.

O'Brien, J. M. and P. J. Szerszen, "An Evaluation of Bank VaR Measures for Market Risk during and before the Financial Crisis," (2014), Finance and Economics Discussion Series, no.2014–21, Federal Reserve Board.

Poon, S. and C. W. J. Granger, "Forecasting Volatility in Financial Markets: A Review," *Journal of Economic Literature* 16 (2003), 478–539.

Segoviano, A., M. and C. Goodhart, "Banking Stability Measures," Technical Report, IMF, WP/09/4, 2009.

Suh, S., "Measuring systemic risk: A factor-augmented correlated default approach," *Journal of Financial Intermediation* 21 (2012), 341–358.

Tarashev, N., C. Borio and K. Tsatsaronis, "Attributing systemic risk to individual institutions," Technical Report, BIS, May 2010.