# Model Risk of Systemic Risk Models*

Jon Danielsson
London School of Economics

Kevin R. James
Financial Markets Group LSE

Marcela Valenzuela Ilknur Zer
London School of Economics

November 2011
First Version

**Abstract**

Statistical systemic risk measures (SRMs) have been proposed by several authors. Those generally depend on established methods from market risk forecasting. The two most common SRMs, MES and CoVaR, along with VaR, are compared theoretically and then critically empirically analyzed. They are found to contain a high degree of model risk so that the signal they produce is highly unreliable. Finally, the papers discusses the main problems in systemic risk forecasting and proposed evaluation criteria fur such models.

*Corresponding author Jon Danielsson, `j.danielsson@lse.ac.uk`. We sincerely thank 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. Updated versions of this paper can be found on `www.RiskResearch.org` and the Webappendix for the paper is at `www.RiskResearch.org/sysrisk` with extended results and computer code.

# 1 Introduction

Systemic risk, having been a neglected backwater in the empirical literature before the crisis, has emerged as a major public concern, with policy makers under considerable political pressure to develop methodologies for measuring and containing systemic risk. In this paper, we approach the issue of systemic risk measurement from the point of view of model risk, and empirically analyze the most common approaches. Our results are essentially negative, existing empirical methods for systemic risk come up short and do not provide a reliable signal to policymakers.

Most recent systemic risk initiatives in the literature are theoretic in nature and it is usually not possible to bring such models to data. There are several exceptions where theory motivated models lend themselves naturally to the question of empirical systemic risk. Of these, the most prominent are CoVaR proposed by Adrian and Brunnermeier (2010), MES suggested by Acharya et al. (2010); Brownlees and Engle (2011) and the Shapley value (SV) approach of Tarashev et al. (2010).

These methods, along with other similar approaches, are fundamentally based on using daily price data to forecast value–at–risk (VaR) as a first step in the calculation of the systemic risk measure, perhaps with expected shortfall (ES) as the next intermediate step. Because VaR is the first step in the calculations, it is the cornerstone upon which all subsequent calculations are based.

Surprisingly, in spite of its prominence, very little formal model risk analysis has been done on VaR. Of course, various authors have studied the empirical properties of different VaR measures, but not usually from the point of view of model risk. This means that the empirical properties of most riskometers,[1] like VaR, are not well understood.

The question of model risk in systemic risk riskometers is important, because as concluded by Danielsson, James, Valenzuela and Zer (2011) a noisy riskometer may be worse than no riskometer at all, that is, for a riskometer to be useful it needs to be of a high–quality, not just weakly better than noise.

We consider the two most commonly used systemic risk measures (SRMs) in detail, MES and CoVaR, but our conclusions likely would apply to other SRMs like SV. We employ a sample of daily returns from the 92 largest financial institutions in the US from January 1997 to December 2010 for the VaR

[1] In the terminology of Daníelsson (2009).

and MES analysis and a subsample of those for CoVaR. We present representative results in this paper, with full results, along with the computer code to implement all methods, relegated to the Webappendix, `www.RiskResearch.org/sysrisk`.

MES and CoVaR derive from the joint distribution risk of the system and the risk of institution. CoVaR is the conditional risk of the system given the institution whilst MES is the conditional risk of the institution given the system. CoVaR directly uses VaR whilst MES is based on ES, which then itself is based on VaR. We could just as easily have defined CoES and MVaR.

We start by focusing on VaR model risk and forecast daily 1% VaR with six of the most commonly used empirical methods and three estimation windows. Our main tool of analysis is the ratio of the highest to the lowest VaR forecasts, *VaR ratio*. In the 4 institution subsample presented in this paper, the median VaR ratio is 3, with the smallest 1.4 and largest 13.3. The lowest values come from normal or conditionally normal approaches whilst the highest provided by fat tailed or historical simulation based methods. The VaR ratio is highest during market turmoil. This suggests to us that the model risk in VaR is quite high and increasing with market uncertainty. This cost doubts about the wisdom of basing SRMs on VaR.

The first SRM we study is 5% MES which is based on ES and in turn derived from VaR, where we investigate the model risk of MES employing the same VaR forecast methodologies. There are two reasons for using 5% in spite of its rather none–extreme nature. First, MES was proposed by Acharya et al. (2010) who use 5% probability, and second it is not possible to estimate MSE at more extreme probabilities except in special cases. We find similar results as for VaR, MES estimates highly depend on the chosen model, especially during the crisis period, and model risk is very high.

One gets much more accurate risk forecasts in the center of the distribution compared to the tails, and therefore 5% risk forecasts are more accurate than 1% risk forecasts. This however does not mean the the dynamic behavior of 5% risk forecasts says much about the dynamic behavior of 1% risk forecasts, especially when one compares non–crisis periods to crisis periods. The reason is that in a crisis, the shape of the return distribution changes and different quantiles are affected differently. For example, we may expect much bigger relative changes in the more extreme quantiles. ultimately this means that less extreme risk forecasts, such as at 5% are not all that informative about more extreme risk, at 1% and beyond.

Finally, we consider CoVaR, and especially $\Delta$CoVaR which is the difference between the CoVaR conditional on a institution under distress and CoVaR calculated in the median state of the same institution. In order to understand the empirical properties of ΔCoVaR we focus on both the cross–sectional and time–series relationship between ΔCoVaR and VaR. We find that there is a considerable difference between the two measures cross–sectionally, suggesting that even when the VaR of two institutions is similar, their contribution to a systemic event, i.e. their ΔCoVaR can be different, confirming the results in Adrian and Brunnermeier (2010). We then estimate the confidence intervals in of the ΔCoVaR — VaR measures and find them to be quite large. Finally, mathematical and time–series analyses suggest that ΔCoVaR is more than 99% correlated with VaR implying that it may be sufficient just to use VaR.

Taken together, the results provide a rather dismal view of the current crop of SRMs. The main reason is that these methods depend on established methods in market risk forecasting, with strong assumptions on the stochastic processes governing market prices necessary. Since different statistical methods depend on different assumptions, there is no surprise that model risk is very high. This problem is made worse by the stochastic processes of market prices being quite different in crisis and non-crisis periods, so that models estimated before a crisis are likely to give very misleading indications of risk during crisis.

The SRMs suffer from two additional problems. First, they are most needed during crisis periods but at that time they are least reliable both because the data has undergone a structural break and also because of the relative scarcity of crisis data. Secondly, because they depend on additional model assumptions beyond the underlying VaR, for example in estimating ES and other measures that are conditional on VaR but in addition rely on other assumptions as well, model risk accumulates, making them less reliable than the basic market risk models.

Ultimately this means we have serious reservations about the use of currently available SRMs for policy purposes. However, there are some models on the drawing board that may provide a useful indicator of risk during crises, and we discuss some those below.

Finally, because of the importance attached to any SRM, we feel that they should be subject to a set of evaluation criteria before use. First, point forecasts are not sufficient, confidence intervals, backtesting and robustness analysis should be required. Second, models should not only rely on observed market prices, instead, ought to aim at capturing the pre–crisis built-up of risk as well. Finally, the probabilities should correspond to actual crisis probabilities. SRMs usually focus on 1% or 5% daily events, those that happen once every 5 months or once a month respectively. That is hardly systemic. Extreme market turmoil happens perhaps once a decade, so the relevant probability is less than 0.1%.

The outline of the paper is as follows. In Section 2 we start by providing a survey of the literature, followed by a systematic attempt at identifying systemic risk methodologies, especially those with an empirical bent in Section 3. In Sections 5, 6, 7 we analyze in detail three riskometers, VaR, SES/MES and CoVaR respectively. This is followed by Section 8 analyzing the results and proposing criteria for systemic riskometers. Section 9 concludes.

# 2 Related Literature

Following the financial crisis, a number of studies are proposed for embedding risk measurement — typically of the systemic risk type — within the macroprudential regulatory structure. Such proposals generally aim to measure an individual bank's contribution to overall financial system risk, aiming to capture phenomena such as extreme risk and how extreme risk is related between bank, the causality between an individual bank risk and the risk of the system and the joint risk of banks and the system.

The current empirical literature on systemic risk can split into two, as noted by Tarashev et al. (2010). One branch of the literature explicitly treats the financial system as a portfolio of banks. Among others, Segoviano and Goodhart (2009) propose a set of quantitative measures of the financial stability of a portfolio of banks (banking stability measures) by considering the dependence among the banks. Zhou (2010) proposes two measures as an extension to their work: a systemic impact index (SII) and a vulnerability index (VI). The SII measures the distress in the overall system given one particular bank fails, whereas VI is the reversed measure, the probability of a particular bank failure given that there exists at least one another failure in the system.

The other branch of the literature focuses on how an individual bank contributes to systemic risk. Several different methods have been proposed, for example the conditional value at risk (CoVaR) method of by Adrian and Brunnermeier (2010) which aims to measure the systemic importance of individual banks and distress in one bank given that another bank is in financial distress.

Other methods consider the reverse problem of a bank risk given system risk. Recent examples include Acharya et al. (2010), who propose marginal expected shortfall (MES) as a measure to a bank's contribution to systemic risk. Banks with higher MES are the ones that contribute the most to the market decline, hence they are more likely to be systemically risky. By employing TGARCH and DCC models, Brownlees and Engle (2011) investigate time series dynamics of MES. They conclude that higher volatility of the institution and less diversification with respect to the market induces higher MES, i.e. higher systemic risk contribution.

Huang et al. (2010) construct a systemic risk indicator using the structural model of Vasicek (1991). They assess the individual banks' marginal contribution to systemic risk by credit risk models where the probability of default, size and asset return correlations, are estimated from banks' CDS spreads.

Tarashev et al. (2010) propose a Shapley value (SV) methodology which is a game theoretical concept first introduced by Shapley (1953). It aims to allocate the total value generated by a coalition to individual players. As an application to this game theoretical approach on systemic risk, SV methodology allocates the total risk of the aggregated financial system, measured by both VaR and ES, to individual institutions. Those allocations are based on each institution's marginal contribution to the overall risk.

The SV approach diverges from both SES and CoVaR in that SES and CoVaR do not try to decompose the overall systemic risk, they rather focus on individual institutions, whereas SV decomposes the systemic risk based on a characteristic function. This means that the summing up of CoVaRs for all individual institutions in the system does not give the overall systemic risk, summation of all SVs does.

All of the studies above assume that the the shocks to the banking system and individual banks are essentially exogenous. This means the models assume away for the most parts the potential for domino effects in defaults or risk transmission. Addressing this problem, Drehmann and Tarashev (2011) take into account the interbank linkages by considering the system as a network of institutions using a SV approach similar to Tarashev et al. (2010). In their sample of 20 global large banks they conclude that the contribution of interconnected banks to systemic risk is higher than the ones without any interconnection. Moreover, systemic importance of an institution is increasing with its level of interbank market activity.

Most empirical implementations of the methods discussed are based on using market price data, perhaps coupled with other information like balance sheet data, to construct the systemic risk measure (SRM). In most cases, the fundamental building block of the SRM is daily VaR. While there are many studies considering the forecast properties of the various VaR implementations, not many papers have formally used model risk analysis to study VaR.

One exception is Daníelsson (2002), and a more recent paper by Boucher et al. (2011) who propose a formal environment for understanding model risk in riskometers.

## 3 Systemic risk measures

Amongst the many systemic risk measures (SRMs) that have been proposed, two have gained particular attraction; SES and CoVaR. They are closely related to each other and in turn derived from VaR.

Let $R_i$ indicate the risky outcomes of a financial institution $i$ on which the risk measures are calculated. This could be for example, daily returns of such an institution. Similarly, we denote the risky outcomes 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 normalizing constant, we get by Bayes theorem:

$$f\left(R_i|R_S\right) \propto f\left(R_S|R_i\right) f(R_i).$$

Suppose we use VaR as a risk measure, we arrive at CoVaR. Define $Q$ as the event such that:

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

where $Q$ is some extreme negative quantile and $p$ is the probability. Then, $\mathrm{CoVaR}_i$ is the value at risk of the financial system given that the institution $i$ is under financial distress. i.e.,

$$\mathrm{CoVaR}_i = \mathrm{pr}[R_S \leq Q_S | R_i \leq Q_i] = p$$

If alternatively we use ES:

$$\mathrm{E}[R | R \leq Q]$$

we get MES; institutions expected equity loss in tail market outcomes:

$$\mathrm{MES}_i = \mathrm{E}[R_i | R_S \leq Q_S] \tag{1}$$

We could just as easily have defined MVaR as

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

and CoES as

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

To summarize;

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

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 $\theta$ and the system $S$. If we choose $\theta$ 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 the same definition of MES.

Ultimately, this suggests that regardless of the risk measure or conditioning, the main determinant of the empirical performance of each measure is VaR. It would consequently be of benefit to study the empirical properties of VaR in detail.

# 4 Data

Since our focus is on systemic risk, it would be natural to consider a sample of financial institutions. In order to keep the estimation manageable and avoid problems of holidays and time zones, we focus on the largest financial market in the world, the US. Similar to the sample used in Acharya et al. (2010), we pick the 101 largest publicly traded financial institutions, each with a market capitalization in excess of $5 billion as of June 2007. However, as we would like to use a relatively long time period for the analysis, we need to drop 9 institutions since they have fewer than 1501 return observations over the period of January 1997–December 2010. Therefore, for all of the three risk measures discussed in the following section, the sample consists of 92 financial institutions. Not all institutions are present throughout the entire sample, but because of the nature of our study, we do not need to worry about survivorship bias.

The daily market equity data is obtained from CRSP, in a total return form, i.e., adjusted for issues such as splits, dividends and buybacks. To capture the overall financial system, following Acharya et al. (2010) we use the value-weighted market return from CRSP.

Adrian and Brunnermeier (2010) in their proposal for CoVaR use a quantile regression method that depends on the presence of certain state variables. The list of all the state variables and data sources can be found in Appendix.

For the sake of brevity, the analysis in the following sections is based on four representative stocks: Bank of New York Mellon (BK), JP Morgan Chase& Co. (JPM), State Street (STT) and US Bancorp (USB). However, entire set of results can be found in the Webappendix.

# 5 VaR

The two main systemic risk measures studied in detail below fundamentally dependent on VaR. It makes no difference whether they are based on VaR directly like CoVaR or indirectly like MES. The same applies to the Shapley measures.

Empirical analysis of VaR will therefore provide useful guidance on how we can expect the systemic risk measures to perform. After all, their performance will necessarily be worse than VaR.

## 5.1 Estimation

We employ the most common VaR forecast methodologies used by industry: normal GARCH (G), Student–t GARCH (tG), moving average (MA), Student–t moving average (tMA), exponentially weighted moving average (EWMA) and historical simulation (HS) along with 3 estimation window sizes; 500, 1000 and 1500 days. The probability is 1%. In turn, this generates 18 VaR forecasts for each institution each day.

It was not possible to obtain VaR forecasts for every estimation method and institution each day. In some cases, the nonlinear optimization methods would not converge, usually for tGARCH, but even in some cases happened for GARCH, always on days with extreme market outcomes. In other cases, the optimizer did converge but the results were nonsensical, where the VaR exceeded the portfolio value.

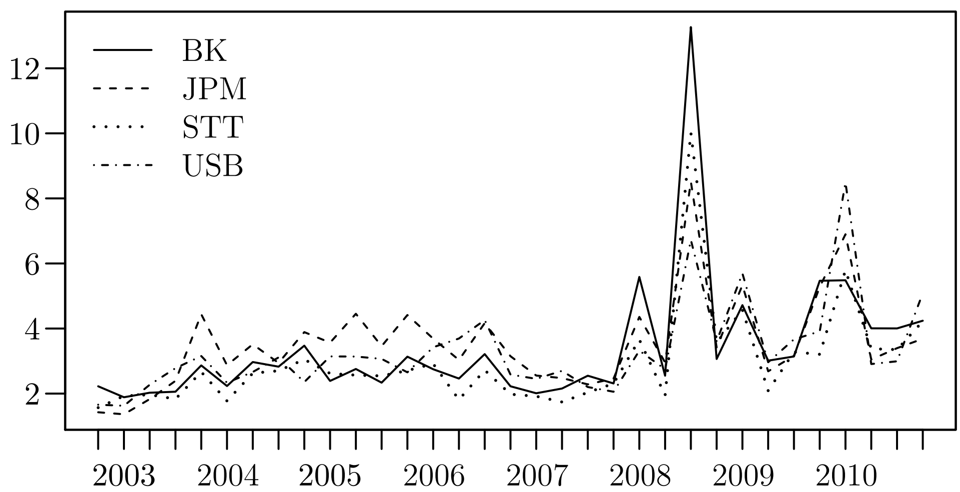
While one might be tempted to use different optimizer, our investigation showed that the optimization failed because the model was badly mis–specified given some of the extreme outcomes. In particular, the models were unable to simultaneously find parameter combinations that work for typical market outcomes and the extreme outcomes. Using a different optimizer would not solve that problem.

For this reason, for cases where the optimizer did not converge, we report a NaN. For the plots in the Webappendix we imposed an additional magnitude restriction on the VaR.

## 5.2 Results

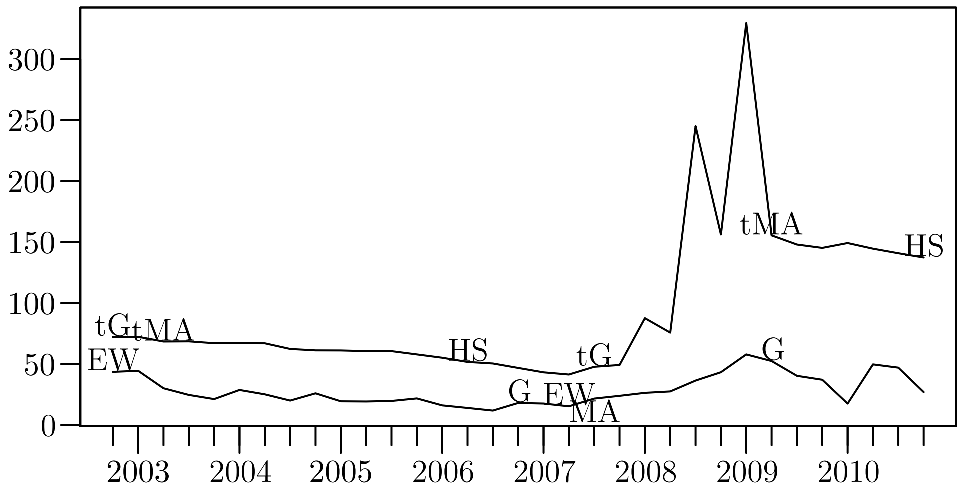
In order to present representative results, in the figures below we only present end of quarter results. As always, the full results can be found in the Webappendix. We present the ratio of the highest to lowest VaR across all methods in Figure 1.

Figure 1: Ratio of highest to lowest daily 1% VaR. End of quarter results. Probability is 1%




The most quiet part of the sample is at the beginning and at this also where the highest and lowest VaR forecasts are most similar, with a ratio below two for all but one asset. As we move into the 2007 crisis, the ratio increases sharply, exceeding 13 for one of the banks during the extreme turmoil of 2008. Taken together, these results suggest that there is a considerable difference between the forecasts, with little guidance on which to employ.

Figure 2: JP Morgan highest and lowest daily VaR and the VaR methods. Every time the VaR method changes, the label changes. End of quarter results. Portfolio value is $1000 and probability 1%




In order to investigate this further, Figure 2 focuses on JP Morgan and shows which VaR method produced the highest and the lowest forecast. In order to minimize clutter, we only indicate the method if it changes from the previous quarter. As expected, the fat tailed methods, HS, tGARCH and tMA produce the highest forecasts whilst the relativity thin tailed methods of EWMA, MA and GARCH result in the lowest forecasts.

## 5.3 Analysis

The empirical results suggest that model risk is quite high. By using the most commonly used methods and assumptions employed by the financial industry and academics, the method producing the highest forecasts results in forecasts that are on average times higher than the lowest. Unfortunately, there is no obvious way to identify which method is best.

One could employ back testing, but that is at the best of times limited, and especially likely to fail during the structural break of 2007. Suppose one had found the best model to use at the end of 2006. A financial institution choosing to implement that model cannot easily change it in the future, and similarly a SRM may not be changed. After the model choice has been made, the crisis arrives along with a structural break in the stochastic process for returns. In order to identify a better model, the back testing procedure would need to have sufficiently large data sample after the crisis started. However, by that time the crisis might be over. For this reason, one cannot rely on back testing to pick the best model.

The end result is that we are faced with a wide range of possible VaR forecasts with no reliable method for identifying the best.

# 6 MES

The first SRM we study is MES as proposed by Acharya et al. (2010). It is defined as the institution's expected equity loss given that the system is in a tail event. Its formal definition is given by (1), i.e.

$$\mathrm{MES}_i = \mathrm{E}[R_i | R_S \leq Q_S]$$

where $R_i$ is the institution's risky outcome on which the VaR is calculated, $R_S$ is the system outcome and $Q$ is some extreme negative quantile.

Hence it is an expected shortfall (ES) estimate modified to use a threshold from the overall system rather than the returns of the institution itself. It is therefore dependent on the system VaR.

## 6.1 Estimation

In order to investigate the model risk of MES, we employ the same VaR forecast methodologies used in Section 5 with a 500 day estimation window size for 5% probability level. There are two reasons for using 5% in spite of its rather none–extreme nature. First, MES was proposed by Acharya et al. (2010) who use 5% probability, and second and since it is not possible to estimate MSE at more extreme probabilities except in special cases.

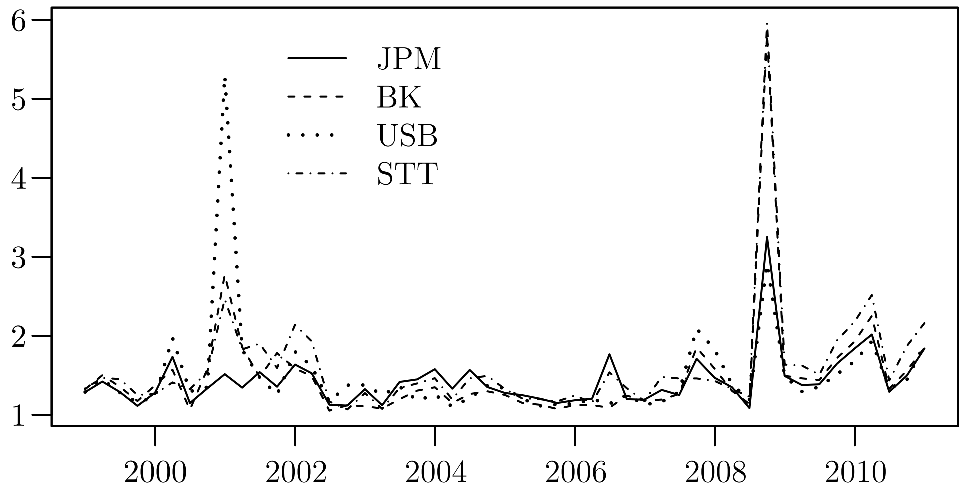
As a first step, the VaR of the overall financial system is estimated. Then, we obtain the days in which the system is in distress compared to its VaR level. Finally, using the definition of MES, for those given days, we calculate the ES of each institution. For 92 stocks in our sample, we end up six different MES estimates for each methodology employed for the overall testing window.[2]

## 6.2 Results

Figure 3 illustrates the ratio of the highest and lowest MES estimates within the six methods for the last day of each month in our sample period.

[2]Acharya et al. (2010) employ historical simulation.

Figure 3: MES model risk. Ratio of maximum and minimum daily 5% MES. End of quarter results




Results shows that the ratio fluctuates significantly for different methodologies employed. It even increases sharply to over 5 for BK in 2008. During the crises period the ratio is usually equal to about two. Note that these ratios are closer to one than the VaR ratios in Figure 1. This is because estimation is much more robust at the 5% probability level than at the 1% level.

## 6.3 MES peculiarities

The results above are based on a 5% probability level following Acharya et al. (2010), which is an event that happens about once a month on average, hardly an extreme event. In order to ascertain how MES performs in more uncommon outcomes, we repeated the analysis for $p = 1\%$, keeping the estimation window size at 500. Those results demonstrate a particular weakness of the MES approach.

In order to calculate the MES, we first need to identify the days in which system VaR is violated based on different methods employed. Since we are interested in in–sample violations, HS will always give $p \times E_W$ number of violations, where $E_W$ is the estimation window size. Any other method could in principle result in the number of violations being any number from 0 to the full window size.

In our sample, for some days VaR is violated for only one of the days when we use GARCH, tGARCH or EWMA. This does not cause problems for VaR or ES. However when we calculate MES, the ratio of the highest to the lowest forecasts can be very high.

As a example, we present detailed analysis for January 10, 2006, a day chosen at random. Given the VaR, HS identifies 5 days as VaR violation, whereas according to GARCH, system VaR is violated only on a single day. Table 1 shows the ES and MES estimates for the system and BK respectively.

Table 1: MES and ES comparison for BK, January 15, 2004 to January 09, 2006 ($E_W = 500$). Portfolio value is \$1000 and probability 1%

| Method | ES | $MES_{BK}$ |
|---|---|---|
| HS | $-16.1$ | $-18.02$ |
| MA | $-16.2$ | $-15.25$ |
| tMA | $-16.12$ | $-15.25$ |
| EWMA | $-15.6$ | $-17.98$ |
| GARCH | $-16.8$ | $-0.35$ |
| tGARCH | $-16.4$ | $-12.12$ |
| Ratio of highest to lowest | 1.08 | 52.23 |

As Table 1 shows, although the number days that VaR is violated is different under two methods, the resulting ES figures are close because the ES calculation is based on the expected returns of the system given that the system is in the tail. On the other hand, in order to estimate the MES for BK, we need to calculate the expected returns of BK, not the system, for the days in which our methods identify as the system is in distress. In that case the return in a single day can be very close to 0, as in this case, making the ratio very far from 1.

This suggest us that estimating MES for 1% or even 5% but with a small sample does not make sense. Because depending on the method employed, one can end up very misleading results.

## 6.4 Analysis

The MES results indicate that the estimates are strongly model dependent with the model risk worse during the crises period. It would be even harder to backtest MES than VaR since it depends on ES.

# 7 CoVaR

The other SRM we study in detail is CoVaR as proposed by Adrian and Brunnermeier (2010), in particular ΔCoVaR which is the marginal contribution of a particular institution to the systemic risk. In what is perhaps their key result, they 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 regrding institutions' risk.

$\text{CoVaR}_i$ is defined as the value at risk of the financial system given that the institution $i$ is under financial distress;

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

where $R$ is the institution's risky outcome on which the VaR is calculated, $Q$ is some extreme negative quantile and $p$ is the probability.

## 7.1 Estimation of CoVaR

Following Adrian and Brunnermeier (2010), $R$ is defined as the growth rate of marked-valued total assets. The time-varying CoVaR is estimated by employing the following quantile regressions:

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

where $M$ denotes the set of state variables that captures the risk over time. A list of the state variables used in above regressions can be found in the Appendix.

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

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

## 7.2 ΔCoVaR

The marginal contribution of an institution, ΔCoVaR, is the difference between the CoVaR conditional on the institution is under distress and CoVaR calculated in the median state of the same institution. i.e.,

$$\Delta\text{CoVaR}_{t,i} = \text{CoVaR}_t^{R_i=\text{VaR}_i^p} - \text{CoVaR}_t^{R_i=\text{VaR}_i^{50\%}}$$

which implies that ΔCoVaR is equal to

$$\Delta\text{CoVaR}_{t,i}(p) = \widehat{\beta}_{S|i}\left[\text{VaR}_{t,i}(p) - \text{VaR}_{t,i}(50\%)\right]. \tag{2}$$

Given that the financial returns are (almost) symmetrically distributed, one expects VaR calculated at 50% is equal to zero. Hence (2) suggests that ΔCoVaR is simply a linear function of VaR.

## 7.3 Results

We replicate a subset of the results for 12 stocks to focus on the relationship between ΔCoVaR and VaR both in cross-sectional and time-series.[3] The sample financial system is composed of 12 stocks. However the results are qualitatively similar when the system consists of 92 stocks instead of 12. Full results are in the Webappendix.

### 7.3.1 Cross–sectional results

We found that there is considerable difference between VaR and ΔCoVaR cross–sectionally, as can be seen in Figure 4, confirming the results of Figure 1 in Adrian and Brunnermeier (2010).

To get the idea of the noise embedded in this estimation, for each of the four stocks we re-run quantile regressions for 1000 times by reshuffling the error terms and estimate VaR, CoVaR and ΔCoVaR for each trial. Figure 5 shows 99% confidence intervals of the bootstrapped estimates. It suggests that a conclusion of JPM being systemically riskier than STT requires a caution since the confidence intervals overlaps in quite wide ranges.

### 7.3.2 Time–series results

When we consider the time–series relationship between VaR and ΔCoVaR, we see that the two measures are almost indistinguishable. Their unconditional correlation mostly exceeds 99%. Figure 6 shows the time–series relationship between VaR and ΔCoVaR for the same four institutions, suggesting that the scaled signal provided by ΔCoVaR is very similar to the signal provided by VaR. This means that the results in Section 5 apply to ΔCoVaR as well.

[3] American Express (AXP), Bank of America (BAC), Bank of New York Mellon (BK), Fifth Third Bancorp (FITB), JP Morgan Chase & Co. (JPM), Keycorp (Key), Pnc Finl.Svs.Gp. (PNC), Regions Finl.New (RF), State Street (STT), Suntrust Banks (STI), Us Bancorp (USB), Wells Fargo & Co (WFC).

Figure 4: VaR and ΔCoVaR in Q4 2006

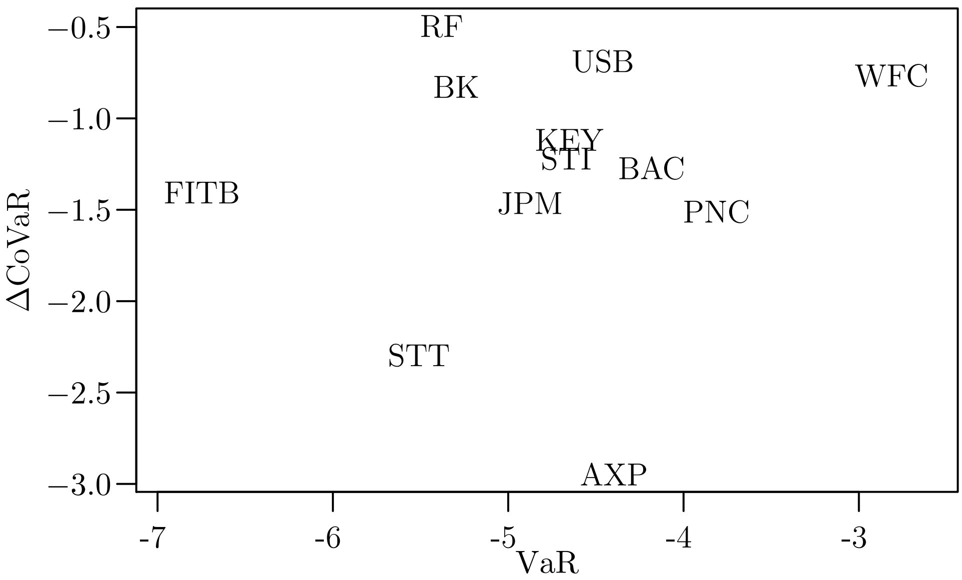
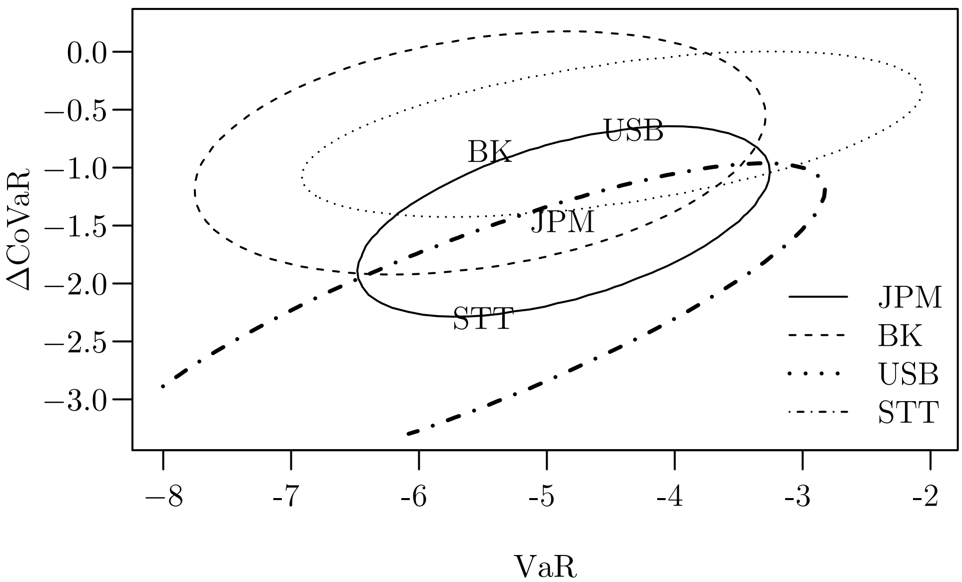


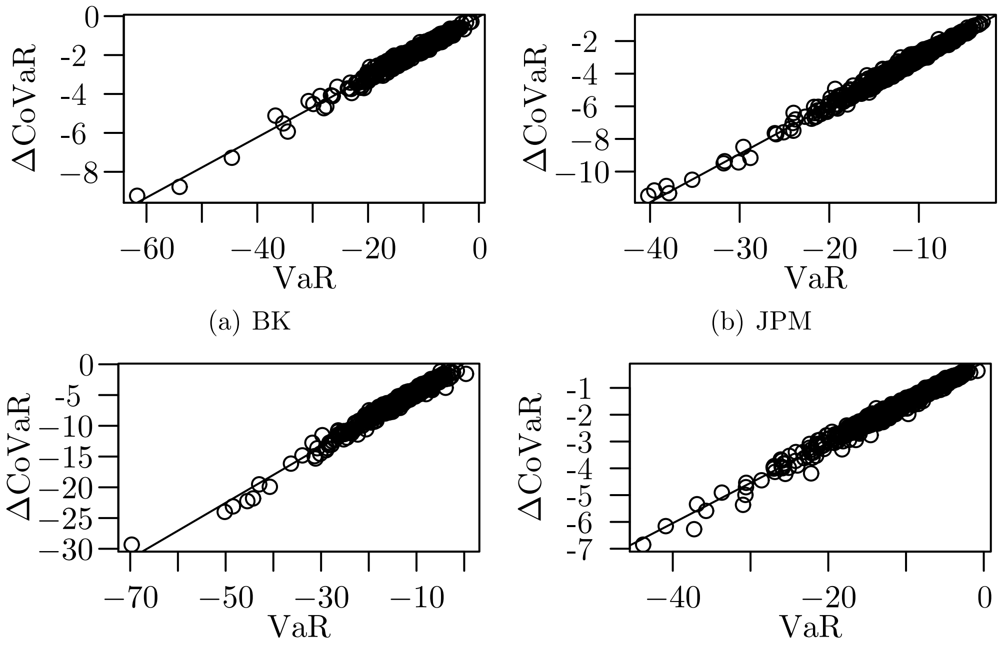

Figure 5: 99% confidence intervals for the institutions in Figure 4




### 7.4 Analysis

How useful is ΔCoVaR as a macroprudential tool? Our results suggest not very. For example, suppose that a supervisor would like to generate a tax scheme to discourage the risk taking behavior of systemically important

Figure 6: Time Series VaR vs ΔCoVaR




(a) BK (b) JPM

(c) STT (d) USB

banks, where systemic contribution is calculated based on ΔCoVaR. But then, our time–series analysis suggests that employing ΔCoVaR might not bring any advantages to simply using VaR, since they are giving almost the same signal.

On the other hand, one can argue that ΔCoVaR and VaR are different cross–sectionally, so ΔCoVaR seems useful to generate a relative regulation scheme, a tax scheme proportional to the relative systemic importance of the institutions. However, this makes sense only if ΔCoVaR is robust. Cross–sectional analysis suggests the reverse though. The bootstrapping exercise finds that the confidence intervals underlying ΔCoVaR estimates are quite large, so it is not possible to conclude which institution is systemically riskier than the other. It is worth to note here that this represents the best case scenario since it is conditional on the model being correctly specified. In reality, the confidence intervals would be even larger.

# 8 Can we measure systemic risk?

Taken together, the results above provide a rather dismal view of the current crop of SRMs. This leaves the question of how we should address the problem of systemic risk measurements and how we should interpret the results suggested by those riskometers.

## 8.1 Relevance of data

The most obvious approach to measure systemic risk is the one taken by the papers discussed above which is to adopt established methodologies in market risk, in particular daily 99% or 95% VaR. There are however particular reasons why such an approach is likely to fail.

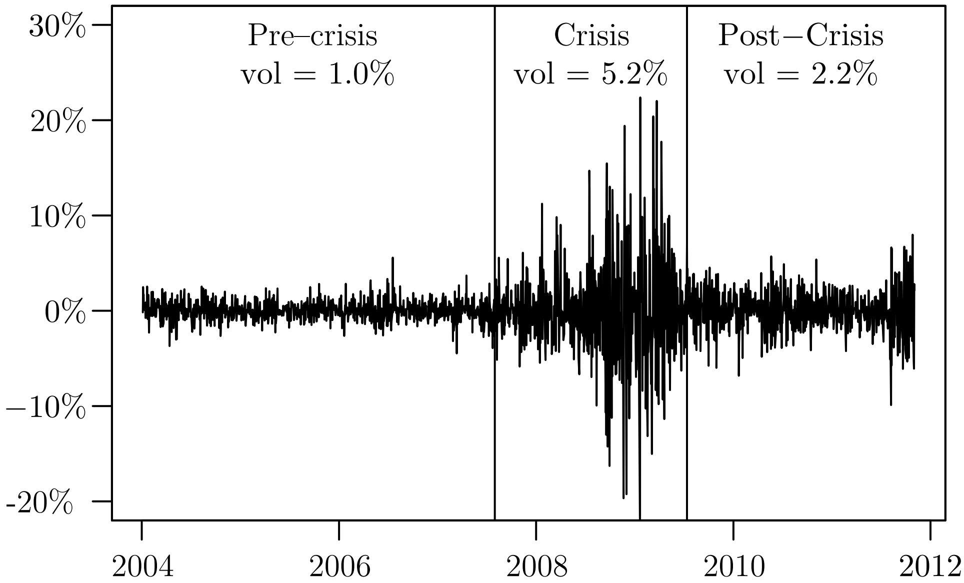
First, we are facing the problem of lack of data. There is of course no shortage of financial data, with almost an infinite number of observations on key market variables. However, this leaves open the question of how relevant such data is for measuring systemic risk. Such events are by definition very uncommon, and it is hard to see how observed data relate to such uncommon events, especially ex–ante. It is important not to confuse data availability with data informativeness.

Over the past half a century, we have observed less than 10 episodes of extreme market turmoil in global international markets, a few more if we consider regional crisis. None of these events can really be considered systemic in the commonly understood sense of the term. 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. It requires a leap of faith to believe that price dynamics during calm time have much to say about price dynamics during crisis, especially when there is no real crisis to compare the forecast to.

Second, our empirical analysis suggests that the riskometers are subject to high degree of model risk and it increases significantly in the days of financial crises. The main reason for this is that such models embed strong assumptions about the stochastic processes governing market prices, assumptions that are likely to fail when the economy transits from a calm period to turmoil or even crisis. From the point of view of statistical modelling, the stochastic processes governing market prices appears to be different in and out of crisis. At the very least, this implies that a reliable method would need to consider the transition from one state of the world to another.

The challenges in applying market data can be seen by a graph of returns

Figure 7: Daily JP Morgan returns before, during and after the last crisis, along with daily volatility




on JPM, before, during and after the crisis as seen in Figure 7. There are three distinct regimes, volatility and extreme outcomes before the crisis do not seem to indicate the potential for future crisis event, and similarly, data during the crisis would lead to the conclusion that risk is too high after the crisis. In other words, we get it wrong in all states of the world, risk assessments are too low before the crisis and too high after the crisis. Ultimately, there is no way to get from the failure process in normal times to the failure process in crisis times.

This also leaves the question of the relevant event probabilities. The systemic riskometers discussed above are estimated with publicly available data, reporting non–extreme outcome probabilities. This however leaves open the question of how reliable such estimation is for the intended purpose. For example, focusing on 99% daily or weekly events would seem to be rather silly if systemic risk events have a lot lower probabilities of occurrence. Furthermore, even if we learn from a previous crisis episode, it might seem rather useless to claim that data predicting the last crisis will predict the next, especially when the causes of crises are different.

One gets much more accurate risk forecasts in the center of the distribution compared to the tails, and therefore 5% risk forecasts are more accurate than 1% risk forecasts. This however does not mean the the dynamic behavior of 5% risk forecasts says much about the dynamic behavior of 1% risk forecasts, especially when one compares non–crisis periods to crisis periods. The reason is that in a crisis, the shape of the return distribution changes and different quantiles are affected differently. For example, we may expect much bigger relative changes in the more extreme quantiles. ultimately this means that less extreme risk forecasts, such as at 5% are not informative about more extreme risk, at 1% and beyond.

To sum up, the main reason why currently available risk models fail when confronted with extreme events is because they assume that currently observed market prices provide a good guidance as to future risk. While this is true for common events, it is not true for systemic crisis. The reason was suggested by the former head of the Basel committee, 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 you upswings, as financial imbalances build up, and materialising in recessions."

## 8.2 Quality control for SRMs

Statistical models used by the supervisors for policy purposes ought to be subject to stringent quality controls. After all, if the policymakers want to use such models to supervise or impose costs on the economy, they must be able to demonstrate robustness of the underlying statistical procedures. This suggests a set of criteria for the evaluation of SRMs:

1. Point forecasts are not sufficient, confidence intervals incorporating both the uncertainty from a particular model as well as model risk should be provided along with any point forecast. Methods need to be analyzed for robustness, including both uncertainty in the model and also model risk;

2. Data should be predictive and not reactive. A large number of time series coincide in signaling market turmoil. However, most of these show market turmoil only after it has occurred. It is therefore important that a systemic riskometer be properly predictive;

3. Statistical method needs to include backtesting. Simply proposing a riskometer without demonstrating how it has performed over historical periods of market turmoil is not sufficient;

4. Event probabilities need to correspond with the probability of actual market turmoil. If such events happen once every 10 years, 99% probabilities of the events, which happen 2.5 times a year are of little relevance.

## 8.3 Is a bad SRM better than none?

When the pros and cons of SRMs are discussed, few would argue that they are without flaws, but a common view is that a bad SRM is better than none. This leaves open two key questions.

How bad are the SRMs? Our results above suggest they are quite bad, perhaps indistinguishable from random noise or at best weakly better in prediction. One must be careful of not falling into the fallacy of requiring a number for decision–making regardless of the number quality. If the methods are as bad as we find them to be, the immediate conclusion should be to go back to the drawing board and not to use them for policy purposes.

Secondly, is a bad SRM better than none? The idea behind SRMs is that they be used for policy purposes, perhaps determine capital for systematically important institutions, or in the design of financial regulations. In such applications there is a high cost of using an incorrect method and because methodological approaches tend to become a part of the regulatory approaches, once they are a part of the regulatory structure are likely to remain there. Witness how resilient the 1% VaR is, virtually unchanged from its initial proposal in 1996. If they are proven to be wrong, the economic consequences can be disastrous. For this reason, a bad SRM should not be acceptable for policy purposes, it actually should be of a proven quality.

## 8.4 Proposals

While much of the current crop of SRMs does not seem to be fit for purpose, it does not necessarily imply that the task of designing a reliable SRM is impossible.

Such a method needs to take into account the insight provided by Andrew Crockett above and go beyond focusing on real–time indicators. After all, we learned in 2007 that all indicators trigger at the same time. Instead, such a method could take a leaf from theories of endogenous risk and explicitly model the slow buildup of hidden risk. Several such approaches have recently been emerging, either from general equilibrium approaches or agent based modeling.

Creating such theoretically based model would be quite challenging, but alternative empirical approaches, perhaps based on Markov switching type methodologies, might provide a partial answer where the switching could be based on the buildup hidden risk before a crisis.

# 9 Conclusion

This paper has focused on the problem of systemic risk measurements (SRMs) and studied in detail the two main approaches to SRM analysis, CoVaR and MES. Since both of those methods, as well as some others proposed in the literature, such as Shapley values, are fundamentally based on daily VaR, analysis of the quality of VaR risk forecasts provides a baseline for the quality of forecasts from the SRMs.

The results are essentially negative. The models are subject to a high degree of model risk and parameter uncertainty, suggesting that any systemic risk forecast out of these models is unreliable.

## Appendix 1: State Variables

This appendix lists the state variables used in the time–varying CoVaR analysis. State variables have been selected in order to capture the time–varying risk factors.

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 since 1990.

2. *Short–term liquidity spread:* Calculated as the difference between 3M-US repo rate and 3M Treasury bill rate. 3M-US repo rate is available in Bloomberg since 1991 whereas bill rate is from the Federal Reserve Board's H.15 release.

3. The change in the 3M Treasury bill rate.

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

5. *The change in the slope of the yield curve:* The change in difference of the yield spread between the 10Y Treasury rate and the 3M bill rate.

6. S&P500 index.

7. Dow Jones real estate sector stocks return.

# References


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

Adrian, T. and M. K. Brunnermeier, "CoVar," Technical Report Staff Report no. 348, Federal Reserve Bank of New York, 2010.

Boucher, C. M., J. Danielsson, P. S. Kouontchou and B. M. Maillet, "Risk Models–at–Risk," (2011).

Brownlees, C. T. and R. Engle, "Volatility, Correlation and Tails for Systemic Risk Measurement," (June 2011).

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`.

Daníelsson, J., "The Emperor has no Clothes: Limits to Risk Modelling," *Journal of Banking and Finance* 26(7) (2002), 1273–1296, download from `http://www.RiskResearch.org`.

———, "The myth of the riskometer," *VoxEU.org* (2009), `http://www.voxeu.org/index.php?q=node/2753`.

Danielsson, J., K. R. James, M. Valenzuela and I. Zer, "Dealing With Systemic Risk When We Measure Systemic Risk Badly," (2011), download from `http://www.RiskResearch.org`.

Drehmann, M. and N. Tarashev, "Measuring the Systemic Importance of Interconnected Banks," Technical Report, Bank for International Settlements, 2011.

Huang, X., H. Zhou and H. Zhu, "Measuring Systemic Risk Contributions of Large and Complex Financial Institutions," (2010).

Segoviano, M. A. and C. Goodhart, "Banking Stability Measures," *IMF working papers* 4 (2009).

Shapley, L. S., *A Value for n-person Games. In Contributions to the Theory of Games*, volume II (Princeton University Press, 1953).

Tarashev, N., C. Borio and K. Tsatsaronis, "Attributing systemic risk to individual institutions," Technical Report Working Papers No 308, BIS, May 2010.

Vasicek, O. A., "The Limiting Loan Loss Probability Distribution," KMV Working Paper (1991).

Zhou, C., "Are banks too big to fail? Measuring systemic importance of financial institutions," *International Journal of Central Banking* 6 (2010), 205–250.