Blame the Models[1]

Jón Daníelsson
London School of Economics

Forthcoming *Journal of Financial Stability*

June 2008

Abstract

The quality of statistical risk models is much lower than often assumed. Such models are useful for measuring the risk of frequent small events, such as in internal risk management, but not for systematically important events. Unfortunately, it is common to see unrealistic demands placed on risk models. Having a number representing risk seems to be more important than having a number which is correct. Here, it is demonstrated that even in what may be the easiest and most reliable modeling exercise, Value–at–Risk forecasts from the most commonly used risk models provide very inconsistent results.

[1] I thank Max Bruche, Charles Goodhart and Con Keating for valuable comments. My papers can be downloaded from `www.RiskResearch.org`.

1 Introduction

As the financial system becomes more complex, the need for complicated statistical models to measure risk and to price assets becomes greater. Unfortunately, the reliability of such models decreases with complexity, so when we need the models the most they tend to be least reliable. Indeed, the credit crunch, which started in the summer of 2007, shows that risk models are of somewhat lower quality than was generally believed. This does not suggest that statistical models should not be employed. On the contrary, they play a fundamental role in the internal risk management operations of financial institutions. The main problem is unrealistic expectations of what models can do.

If we ask practitioners, regulators, academics, and especially model designers, how they judge model quality, their response is frequently negative. At the same time, many of these same individuals have no qualms about using models, not only for internal risk control but also for the assessment of systemic risk which is crucial for the regulation of financial institutions. To me this is a paradox. How can we simultaneously mistrust complicated models and advocate their use?

In order to demonstrate the high degree of uncertainty in risk forecasts, I forecast below Value–at–Risk (VaR) for IBM using the most common models and assumptions. While one can hardly find an easier modeling exercise, even in this case, the highest VaR forecast is double that of the lowest, with no satisfactory criteria for selecting the best one. This is not an unusual outcome. For smaller or less liquid assets we would expect more uncertainty.

It should, therefore, not come as a surprise that financial institutions have realized large losses on positions that were supposedly very low risk, as events since the summer of 2007 have demonstrated.

> "Wednesday is the type of day people will remember in quant-land for a very long time," said Mr. Rothman, a University of Chicago Ph.D. who ran a quantitative fund before joining Lehman Brothers. "Events that models only predicted would happen once in 10,000 years happened every day for three days."
> Wall Street Journal (2007)

These multiple 10,000 year events are however nothing compared to the enormous risks which Goldman Sachs observed:

> "We were seeing things that were 25-standard deviation moves, several days in a row," said David Viniar, Goldman's chief financial officer. "There have been issues in some of the other quantitative spaces. But nothing like what we saw last week."
> Financial Times (2007)

Under a normal distribution[2], a -25 sigma event happens with probability once every 10^{-140} years, which implies that Goldmans suffered a number of days losses, each of which their models predict occurs approximately once every 14 universes, using the current estimate of the age of the universe of approximately 10^{-10} years old.

It is, however, in the risk assessment of credit instruments, in particular CDOs, that the models have been found most lacking. As outlined below, the whole process from origination to purchasing the tranches has been found to be flawed. While the rating agencies take no responsibility for problems of liquidity or pricing, it is clear that, before the credit crunch in the summer of 2007, the ratings were produced by using overly optimistic input data, inappropriate modeling of dependence between assets, insufficient checking on the quality of data documentation, and permitting gaming of models.

Not only were there serious failures in the general approach to modeling risk in CDOs, but also basic mistakes. Indeed, as documented by the Financial Times (May 21, 2008), Moody's discovered in February 2007 that the models used to rate CPDO's had a mistake that provided ratings up to four notches higher than they should have been. "At the same time, the documents record that Moody's staff looked at how they could amend the methodology to help the rating". "The products remained triple A until January 2008." S&P provided the same AAA ratings.

A key problem in the modeling of CDOs was the incorrect assessments of correlations between the individual assets, but as noted by Duffie (2007), there is a serious lack of good models for estimating correlations. In addition, Coval et al. (2008), note that the particular prioritization rule which allows senior tranches to have low default probabilities, and get high credit ratings, also implies that the risk in senior tranches is particularly concentrated on systematically bad economic outcomes, implying they are effectively *economic catastrophe bonds*.

Clearly, this implies, as noted by Ubide (2008), that an AAA CDO tranche does not have the same risk characteristics as an AAA corporate. The average probabilities of defaults may have been similar, but the tails of the distribution are much fatter for CDOs. The differences between CDO and corporate risk characteristics became evident in how these different assets performed in the crisis. Credit spreads on high-grade corporate obligations tend to narrow in a crisis because of flight to quality. We see the opposite with CDOs.

[2] Without that assumption it does not make sense to talk about sigma events by themselves.

The rating agencies have evaluated corporate obligations for 80 years, and these measurements have given us a benchmark to access the ratings quality. Unfortunately, the quality of CDO ratings is different from the quality of the ratings of a regular corporation. An AAA for a SIV is not the same as an AAA for Microsoft. And as the events in the summer of 2007 and beyond indicate, the market was not fooled. After all, why would an AAA rated SIV earn 200 basis points above an AAA rated corporate bond? One cannot escape the feeling that many market participants understood what was going on, but happily went along. The pension fund manager buying such SIVs may have been incompetent, but more likely was simply bypassing restrictions on buying high-risk assets.

The current crisis took everybody by surprise in spite of all the sophisticated models, all the stress testing, and all the numbers. The financial institutions that are surviving this crisis best are those with the best management, not those who relied on models to do the management's job. Risk models do have a valuable function in the risk management process so long as their limitations are recognized. They are useful in managing the risk in a particular trading desk, but not in capturing the risk of large divisions, not to mention the entire institution. For the supervisors the problem is even more complicated. They are concerned with systemic risk which means aggregating risk across the financial system. Relying on statistical models to produce such risk assessments is folly. We can get the numbers, but the numbers have no meaning.

2 The fragility of models

It is, therefore, surprising that both financial institutions, and supervisors, insist on maintaining such a pre-eminent place for statistical risk and pricing models. I suspect that underpinning this view is the belief that sophistication implies quality. A really complicated statistical model *must* be right. That might be true if the laws of physics were akin to the statistical laws of finance. However finance is not physics; it is more complex.

Consider a simple physics model, Newton's second law of motion. An applied force on an object equals the time rate of change in its momentum. We know this holds in theory, and that any empirical deviations must be because of frictions. This gives us the confidence to build complicated engineering systems founded on correct behavioral equations, such as Newton's second law. Key to this confidence is the observation that the phenomena being measured do not generally change with measurement.

Financial models do not have such simplicity. Mathematical integrity in a financial model can just as easily lead us astray as provide an accurate solution. There are several reasons for this, including the following:

Endogenous risk In finance, the statistical properties of the phenomena being modelled generally change under observation since rational market

participants react to information, and by reacting, directly affect what is being observed. Outcomes in financial markets represent the aggregate strategic behavior of a large number of individuals, all with different abilities and objectives. We can only model aggregate behavior. Financial modeling changes the statistical laws governing the financial system in real-time, leaving the modelers to play catch–up. This becomes especially pronounced as the financial system gets into a crisis. This is a phenomenon that Danielsson and Shin (2003) call *endogenous risk*. Day–to–day, when everything is calm, we can ignore endogenous risk. In a crisis, we cannot. And that is when the models fail.

Quality of assumptions The purpose of models is to reduce the complexity of the world into a small number of equations. Therefore, the quality of the assumptions is of key importance. In financial modeling, there are a large number of potentially important factors and the model has to pick and choose, and the modelers tend to ignore what is difficult, not what is important. This has become over the past year when liquidity has been shown to be of key importance. Until then liquidity had generally been ignored in model design

Data quality Financial data have the annoying habit of being of short duration. The statistical properties of financial data change over time, often to a considerable extent, and are influenced by other financial variables as well as the general economy in ways that can be intuitively explained, but are often impossible to model.

I demonstrate these issues in two examples below. The first, one of the easiest risk modeling exercises, considers Value–at–Risk for a large liquid stock, and the second provides a glance at credit risk modeling.

2.1 Example 1: Market risk models

Consider one of the simplest possible risk modeling exercise, measuring 99% Value–at–Risk (VaR) for the stock price of a large corporation, i.e. IBM, using the most common methods for forecasting (VaR). For a survey of risk models used in practice, see the International Monetary Fund (2007).

In this exercise we have to make two subjective decisions:

1. Estimation method[3]

[3] The first two methods, historical simulation and moving window apply equal weight to each observation. In the former, the VaR equals the 1% smallest observation, while in the latter I calculate the VaR from the unconditional volatility using the normal distribution. The last three methods are all based on giving the most recent observation the highest weight in the calculation of VaR and the oldest one the lowest weight. Exponentially weighted moving average and normal GARCH use conditional volatility forecast and the normal

2. Number of observations or sample size[4]

The estimation methods chosen represent the most common risk forecast models used by the financial industry. They range from being very simple and easy to implement such as historical simulation to the highly sophisticated including a fat tail GARCH.

Daily 99% Value–at–Risk estimates for a $1000 portfolio of IBM for May 1, 2008

Sample size	1 year	4 years	10 years
Estimation method			
Historical simulation	$22.8	$17.7	$29.9
Moving window	$24.7	$18.8	$32.6
Exponentially moving average	$21.3	$21.3	$21.3
Normal GARCH	$25.5	$23.4	$24.0
Fat tail GARCH	$27.6	$20.6	$22.9

Even in this easiest case, we get a VaR estimates that differ by a factor of two ($17.7 to $32.6). Furthermore, there is no easy way to pick the best method. We can, of course, use backtesting, but that carries with it its own problems. First, we need a large sample size for a valid test, generally half a decade or more, and we also need to be confident that such old data represent current realities. Second, we need a testing procedure or a model to evaluate the VaR forecasts, and such procedures both lack robustness and carry with them their own model risk.

The model risk increases as we increase the number of assets. If we have a portfolio of equities trading on the same exchange, we need to make a third subjective decision:

3 Aggregation method

There are a number of methods to aggregate volatility/VaR across assets, most of which depend on very strong assumptions. In practice, most financial

---

distribution to calculate the VaR while the fat tailed GARCH is based on using the student t-distribution. For more information about these methods see e.g. Danielsson (2002).

[4] Each method is based on estimating VaR using a sample of past observation of returns on IBM. The smallest sample size is one year of daily data which is the minimum under the Basel Accords. Exponentially weighted moving average is not sensitive to sample size.

institutions use a factor model, which have the unfortunate side effect of underestimating the importance of changing correlations. Indeed, in a financial crisis, it is the large swings in correlations that are of key importance and using a model that does not allow for such changes is likely to be of limited use.

These problems become more pronounced when we incorporate assets traded on different exchanges, time zones and opening hours. This becomes quite challenging when combining assets trading in different continents such as in the UK, US, and Japan where there is no overlap between trading hours in all three markets and dates of bank holidays are quite different. There is no satisfactory method to address this problem. When we further need to incorporate different asset categories, these problems become even more pronounced. Aggregated VaR numbers for a big financial institution are essentially just random since model risk dominates

2.2 Example 2: SIVs and subprime

The modeling exercise in the previous example was relatively straightforward. Finding out the risk in structured credit products is considerably more difficult. The products are created out of multiple separate assets giving rise to the aggregation problem as discussed above, and the short data sets mean that we can not obtain the default probabilities simply by looking at past data, as one does for market risk. Instead, we need to model default probabilities. Therefore, the whole modeling process is much more complicated and harder to verify, which increases the risk of mistakes.
Consider an example of the arrangement at the heart of the current crisis, SIV/conduits containing subprime mortgages. For an interesting account of the problems of origination see, Bitner (2008).

- Most subprime mortgages were originated by brokers who had no financial stake in the quality of the mortgage. Their income came from originating between one and two mortgages a month, giving them on average 1-2% of the mortgage value, a number that increased with higher interest rates. They faced no financial risk if a mortgage subsequently failed. Their incentives were therefore only to originate mortgages, not to verify quality. So not surprisingly a large proportion of mortgage applications were found to be fraudulent to some extent.
- The borrower gets the money from a small bank (lender). This bank will typically borrow the money for the loan from somebody else, originate a sufficient number of loans, and then sell them on to large investors . From 2000-2002 the lenders may have been paid up to 5% for each loan sold on, a number that declined to less than 2% by 2005. The lenders have partial exposure to the mortgages, and will have to keep them on the books if either the borrower defaults before they can sell the loan on, or if the

borrower defaults or misses a certain number of payments in the first year. Their incentives are therefore to minimize the risk of serious payment difficulties in the first year. One way to ensure this, is to keep mortgage payments lower in the beginning, which both helps make the mortgages more attractive to the borrower and reduces the risk of default in the first year.

Essential to the process of securitization are credit ratings. The lack of quality of ratings of structured credit products is now well documented. For a bird's eye view of the problems encountered in the ratings of SIVs, see e.g. Lowenstein (2008) who documents how a particular SIV containing 2,393 mortgages with a total face value of $413 million, was rated by Moody's. Moody's had access to information about each loan but no direct access to the loan applications, but did not consider it to be their job to verify the information. As explained by the head of Moody's derivative group to Lowenstein: "We are structure experts, we are not underlying asset experts." Their approach was to use historical information on such loan characteristics to predict future default rates. According to Lowenstein they allocated one analyst for one day to process the credit data. This particular mortgage–backed security was then used as an input into a CDO. Their revenue may have been $200,000 for this particular security. Their initial estimates were that the CDO might have losses of 2% but the most recent estimate indicates losses of 27%. When Moody's reviewed this particular CDO, they found serious problems with the quality of the mortgage applications.

As *structure experts* in the Moody's terminology, the rating agencies seem to have had some serious deficiencies. The rating agencies underestimated the importance of the business cycle and the presence of a speculative bubble in housing markets. The subprime market took off in the early stages of the business cycle, under economic conditions that were generally improving. In such cases the employment prospects of the typical subprime borrower were improving over time, while at the same time the presence of additional ability to borrow money to buy property stimulated a property bubble. In this case, mortgage defaults are relatively independent events, reflecting individual difficulties rather than problems with the economy at large. At the height of the business cycle, both the default correlations and the rates of defaults of mortgages are low. Of course, even a cursory glance at history reveals that mortgage defaults become highly correlated in downturns. Unfortunately, the data samples used to rate SIVs often were not long enough to include a recession. It would be a straightforward modeling exercise to calibrate the default probabilities and correlations of subprime mortgages in an economic downturn, but it is unclear to what extent this was done.

3 Unrealistic demands

Statistical models play a valuable role both in the internal operations of financial institutions and in the regulatory process. Unfortunately, financial institutions, and especially regulators, have a tendency to place unrealistic demands on models.

Models are most useful where they are closest to the decision–making process in financial institutions, especially in internal risk management. There, the focus is on *frequent small events*, for which statistical models are best suited, because data sample sizes are generally adequate for reliable estimation of high probability events. Essential for a high quality modeling process is harmonization between probability levels, sample sizes, and testing. A risk model is designed for a particular probability level in mind, and will generally not perform well for other risk levels. For example, exponentially weighted moving average (EWMA) could be a good method for 95% VaR calculations, but not for 99% VaR.

Unfortunately, once a model has been estimated, it is generally straightforward to generate any combination of probabilities and outcomes from it. We may design a EWMA risk model to perform well at the daily 95% level, but there is nothing preventing us from using exactly the same model for more extreme risk levels, even for 99.9% annual, regardless of the reliability of such calculations. Similarly, once VaR has been estimated, it is straightforward to calculate other risk measures such as tail VaR. Indeed, there is a strong, but mistaken, belief in some quarters that such risk measures should be used instead of VaR.

Taken to the extreme, I have seen banks required to calculate the risk of annual losses once every thousand years, the so–called 99.9% annual losses. However, the fact that we can get such numbers does not mean they have any meaning. The problem is that we can neither realistically estimate models at such levels without access to data on such events nor can we backtest at such extreme frequencies.

Unrealistic demands on risk models fly in the face of a basic tenet of the scientific process, which is verification, and in our case backtesting. Since neither the 99.9% models, nor most tail VaR models, can be backtested, they cannot be considered scientific, regardless of all the mathematical sophistication that may go into them. There are certain exceptions to this, in particular methods based on extreme value theory and copulas, however, such methods are still mostly experimental, and not ready for day–to–day use.

We do, however, see increasing demands from supervisors for exactly the calculation of such numbers as a response to the current crisis. Of course the underlying motivation is the worthwhile goal of trying to quantify financial stability and systemic risk. However, exploiting the banks internal risk models

for this purpose is not the right solution. Internal models were not designed for this purpose and this type of calculation is a drain on the banks' risk management resources. It is the lazy way out. If we do not understand how the system works, generating numbers may give us comfort. However, the numbers do not equate with understanding. Furthermore, the problems facing the student of financial instability are more complicated than problem facing the financial institution because the estimation of systemic risk requires an integration of the risk of all financial institutions to take into account how they relate to each other. While models enabling such calculations exist, they are still in their infancy.

4 Financial regulation

Bank supervisors increasingly rely on models as a key component in their activities. Indeed, the Basel II Accord, is based on regulation by models. This dependence on statistical modeling is problematic. Model risk is not sufficiently appreciated; it creates perverse incentives; the regulatory processes becomes out of sync with financial markets development; it may destabilize and increases state intervention in the financial system.

Regulation based on models also carries with it the danger of the regulators' view of how to understand and model risk being out of sync with practices in financial institutions. At the beginning, the Basel II process may have incorporated common practices in financial institutions, but that was 10 years ago. Basel II represents the state–of–the–art 1998.

Regulation based on models also carries with it the direct danger that similar risk modeling methodology will be used throughout the financial system. While it is often claimed that there is considerable heterogeneity in risk modeling practices across industry, the survey of financial institutions made by the International Monetary Fund (2007) indicates otherwise. Indeed, harmonization of risk models is the logical outcome of regulation using models, because risk measurements become a competitive issue for financial institutions. If two banks get different risk estimates, and hence capital charges, on otherwise identical positions, the bank with a higher risk estimate is at a competitive disadvantage. Therefore, the incentive is to lower the risk estimate regardless of the actual underlying risk. After all, the purpose of risk management within banks is not to minimize risk but rather the opposite. Therefore, there could be a race to the bottom in model quality with the regulator having to step in to provide minimum standards. The latter, however, are not in a position to provide different standards on models to different institutions; hence, we are likely to see harmonization of regulatory risk models, resulting in even greater endogenous risk, see Danielsson and Shin (2003).

Ultimately, this implies that the role of the regulator, and hence of the state, in the financial system becomes unhealthily large since the regulator becomes

effectively the risk modeler of last resort. That means that the supervisory agency becomes responsible for model quality, and, therefore, will share the blame when things go bad. But in that case, one may ask if the state assumes such a responsibility in ensuring risk is measured and treated accurately, why does the state not share in the up side when things go well?

5 Conclusion

With all the issues facing statistical modeling and finance being better understood, it is incomprehensible why the supervisors are increasingly advocating the use of models in assessing the risk of individual institutions and financial stability. If model driven mispricing enabled the current crisis, what makes us believe future models will be any better?

One of the most important lessons from the crisis has been the exposure of the unreliability of models and the importance of management. The view frequently expressed by supervisors that the solution to a problem like the subprime crisis is Basel II is not really true. What is missing is for the supervisors to understand the products being traded in the markets, to have an idea of the magnitude and potential for systemic risk, and of the interactions between institutions, endogenous risk, coupled with a willingness to act when necessary.

References

Bitner, R. (2008). *Greed, Fraud & Ignorance: A Subprime Insider's Look at the Mortgage Collapse*.

Coval, J. D., Jurek, J. W., and Stafford, E. (2008). Economic catastrophe bonds. mimeo, Harvard University.

Danielsson, J. (2002). The emperor has no clothes: Limits to risk modelling. *Journal of Banking and Finance*, 26(7):1273–1296.

Danielsson, J. and Shin, H. S. (2003). Endogenous risk. In *Modern Risk Management — A History*. Risk Books.

Duffie, D. (2007). Innovations in credit risk transfer: Implications for financial stability. mimeo, Stanford University.

Financial Times (2007). *Goldman pays the price of being big*. Financial Times, August 13, 2007 edition.

Financial Times (2008). *CPDOs expose ratings flaw at Moody's*. Financial Times, May 21, 2008 edition.

International Monetary Fund (2007). Global financial stability report: Financial market turbulence: Causes, consequences, and policies.

Lowenstein, R. (2008). Triple–a failure. *New York Times*.

Ubide, A. (2008). *Anatomy of a modern credit crisis*. Forthcoming Financial Stability Review, Bank of Spain.

Wall Street Journal (2007). *One 'Quant' Sees Shakeout For the Ages – '10,000 Years'*. Wall Street Journal, August 11, 2007 edition.