Loss Distribution Estimation, External Data and Model Averaging

Ethan Cohen-Cole* and Todd Prono**

This paper discusses a proposed method for the estimation of loss distribution using information from a combination of internally derived data and data from external sources. The relevant context for this analysis is the estimation of operational loss distributions used in the calculation of capital adequacy. A robust, easy-to-implement approach that draws on Bayesian inferential methods has been presented. The principal intuition behind the method is to let the data itself determine how they should be incorporated into the loss distribution. This approach avoids the pitfalls of managerial choice on data weighting and cut-off selection and allows for the estimation of a single loss distribution.

Introduction

This paper aims to estimate the operational loss distributions used in the calculations of capital adequacy. Capital adequacy may be either regulatory capital, as required under Basel II guidelines or economic capital. Throughout the paper, it has been referred to capital adequacy broadly as the methods and approaches here are generalizable. Under new Basel requirements, a number of institutions in the United States are likely to be required to use the so-called Advanced Measurement Approach (AMA) and a handful more look likely to 'opt-in'.[1] This AMA approach requires the inclusion of information from four 'elements': internal data, external data, scenarios and business environment and internal control factors. The combination of internal and external data, are addressed though the methods could in principle be extended to include information from the other two.

External data are often used to 'fill in' information on very low frequency, high severity losses where data are unavailable within the institution. By their nature, it is this group of losses that are both of issue in capital adequacy and the most difficult to measure. Measurement difficulties derive from the very fact that they are of such low frequency.

Why have large banks had such a difficult time in measuring large events and incorporating them into capital estimates? This paper focuses here on the latter question and return to the issue of data collection briefly in the conclusion. Assume that a bank has completed the exercise of gathering external data on large, low frequency losses. Its task is now to find a way to mix this information into the set of existing losses in some logically and statistically consistent way. This is a non-trivial exercise for a variety of reasons.

* Financial Economist, Federal Reserve Bank of Boston, Boston, USA. E-mail: ethan.cohen-cole@bos.frb.org
* Financial Economist, Federal Reserve Bank of Boston, Boston, USA. E-mail: todd.prono.@bos.frb.org

[1] See the Notice of Proposed Rulemaking, Federal Reserve Board of Governors. [http://www.federalreserve.gov/GeneralInfo/Basel2/NPR%5F20060905/].

First, there is no reason to believe that the quantity of available external data is related to the quantity of available internal data; i.e., it is not possible to simply add this data to the existing data set without some attempt at accounting for the implied frequencies of the two samples. Second, once the relative frequency matter has been handled, a larger problem crops up. How relevant are the individual observations? Are all observations good? Should some of the data be weighted or scaled? Should some be tossed out? These questions are typically answered with a range of managerial decisions from the 'weight' to ascribe to internal data to the appropriate 'cut-off' point to distinguish between the body of a distribution and its tail. In theory, one wants to find a way to gather and use the information contained in the external data. Specifically, one wants to use only that information which is of relevance, while ignoring or down-weighting observations that are of less importance or accuracy.

The goal of this paper is to present a methodology for the incorporation of external data into internal loss distributions. A robust, easy-to-implement approach that draws on Bayesian inferential methods has been presented. The principal intuition behind the method is to let the data itself determine how they should be incorporated into the loss distribution. The approach avoids the pitfalls of managerial choice on data weighting and cut-off selection and allows for the estimation of a single loss distribution.

Background

This paper suggests that one can consider this a problem of 'model uncertainty' and that Bayesian methods of model averaging can resolve the persistent difficulties. 'Models' in this context, mean a set of assumptions about the set of appropriate external data points to include in an analysis, their weighting relative to internal data, etc.

The fundamental problem that underlies the disparate methods used across the industry is that individual approaches reflect unspecified, yet highly specific, assumptions about the appropriate weighting of data, distribution, cut-off points, etc., on the part of the analysts. These assumptions can reflect attempts to manage the number produced by the model (choosing to under-weight external data) and can have large effects on the conclusions of a particular data analysis. However, one is hard pressed to make a compelling argument that inclusion of a given point over another is the crucial decision in the formation of a study. This is based on the fact that these assumptions are themselves typically not falsifiable. As a result, regulators and bank managers—both of whom may have developed conceptually valid and potentially correct 'models'—can reach quite distinct conclusions. The resulting lack of certainty over the appropriate data set and a suitable weighting procedure, or 'model uncertainty', is the technical focus of this note. In particular, model averaging is utilized, which is a method that enables a researcher to take the weighted average of findings across possible 'models' to develop inferences that are not dependent on the assumption that one of the models is true.

A growing literature has emerged in the use of this method known as model averaging. It seeks to avoid researcher bias in the determination of variable choice, specification

choice, etc., by including a large set of possibilities into a single analysis. The earliest discussion of model averaging is Leamer (1978). After a long period in which the methods were not widely used, the mid-1990s saw a resurgence of interest. The re-initiation of interest appears to be due to Draper (1995), Raftery (1995) and Raftery *et al.* (1997). Useful introductions are available in Hoeting *et al.* (1999), and Wasserman (2000). Model averaging has been advocated and employed in various fields, see Brock and Durlauf (2001), Fernandez *et al.* (2001), Brock *et al.* (2003), Sala-i-Martin *et al.* (2004) and Cohen-Cole *et al.* (2007).

Theory

In the analysis here, this paper follows the model averaging literature as the mechanism for dealing with model uncertainty. To accommodate the particulars of the case at hand, the general framework that has been developed in the statistics literature is adapted; however, standard frequentist estimators are used.[2] This frequentist approach to model averaging is described in Sala-i-Martin *et al.* (2004) and Brock, Durlauf and West (2003). The basic idea of model averaging is straightforward. Consider an object of interest—in this case the appropriate parameters of a loss-distribution—and take a weighted average of the findings across all possible 'models' that can explain the effect. Each model's effect is weighted based on its ability to explain the data.

To see how this procedure works intuitively, imagine a set of external data points and the question of whether or not to include a particular large fraud observation. One school of thought argues that it represents the existing risks to the institution and another that the case was non-representative and should be excluded. Inclusion or exclusion of that particular data point leads to large differences in capital requirements. One can think of this disagreement as reflecting a simple form of model uncertainty in that the model space has only two elements: a set of data that includes the point and one that does not. How would we propose adjudicating the disagreement? It is argued that one should average the estimates from the two studies by taking a weighted average of the results from each, where the weights are posterior model probabilities.[3] By setting these two cases alongside one another, it is possible to evaluate the relative probability that each can describe the data used.

More generally, the structure of model averaging can be understood as follows. Suppose one wishes to produce an estimate of some object of interest δ. In the context of operational risk and capital measurement, δ tends to be the parameters of a statistical distribution that reflects the losses of a given institution. Conventional statistical methods may be thought of as calculating an estimate that is model specific, $\hat{\delta}_m$. In the

[2] Within the statistics literature, model averaging is usually done in Bayesian contexts. However, most readers will be unfamiliar with the interpretation and implementation of Bayesian estimators. A full discussion of the difference between Bayesian and Frequentist approaches is beyond this paper.

[3] This can be thought of simply as the conditional probability that a given model describes the data. This as discussed at greater length below.

model averaging approach, one attempts to eliminate conditioning on a specific model. To do this, one specifies a set or space of possible models M. The true model is of course unknown, so from the perspective of the researcher, each model will have some probability of being true.[4] These probabilities depend on the relative goodness of fit of the different models, given available data *D* as well as the prior beliefs of the researcher (discussed below); hence each model will have a posterior probability: $\mu(m|D)$. This can be read as the probability of a given model (*m*) conditional on data available, *D*. These posterior probabilities allow us to average the model-specific estimates:

$$\hat{\delta} = \sum_{m} \hat{\delta}_m \, \mu(m/D) \qquad \ldots(1)$$

The estimate $\hat{\delta}$ thus accounts for the information contained in each specific model about δ and weights this information according to the likelihood that of the model being correct one. As suggested above, in the case that a single model is true, that is, it perfectly fits the distribution in question, it will receive a weight of 1. Brock, Durlauf and West (2003) argue that the strategy of constructing posterior probabilities that are not model-dependent is appropriate when the objective of the statistical exercise is to evaluate alternative policy questions such as level of capital. Notice that this approach does not identify the 'best' model; instead, it studies an outcome (a distribution and a capital number) conditional on the data. Notice that in theory, there is no reason for one to choose a particular set of external data points as the 'correct' ones. Instead, one wants the aggregated information content of the external data.[5]

Thus, while the exercise could theoretically yield a single model with a weight of 1, this is unlikely. In practice, a finding that a given model, *m**, within some space M, that has the highest conditional probability of describing the data is not a recommendation to select that model. Instead, it merely identifies that model as being the one that has the largest contribution to $\hat{\delta}$.

Notice that averaging across models means that a key role is played by the posterior model probabilities. Using Bayes rule, the posterior probability may be rewritten as

$$\mu(m|D) = \frac{\mu(D|m)\,\mu(m)}{\mu(D)} \propto \mu(D|m)\,\mu(m) \qquad \ldots(2)$$

The calculation of posterior model probabilities thus depends on two terms. The first, $\mu(D|m)$ is the probability of data given a model (again, in our setting, a set of external data points). Raftery (1996) has developed a proof to illustrate that the probability of a model, given a dataset, can be calculated using the ratio of a given model's likelihood to the sum of the likelihood of all models in the space M. This derivation allows us to use

[4] If one model is true then this method produces a probability of 1 for that model and zero for all others.

[5] It is acknowledged that there is some value in the identifying a set of 'correct' scenarios, particularly for business planning and risk control establishment. However, it is to distinguish that exercise from this one – the identification of an appropriate and accurate loss distribution.

the likelihood of a given model for $\mu(D|m)$. The second term, $\mu(m)$, is the prior probability assigned to model m. Hence, computing posterior model probabilities requires specifying prior beliefs on the probabilities of the elements of the model space M (see below).

Calculating Variance of $\hat{\delta}$

While one does not traditionally bound capital estimates to account for uncertainty in estimation, the methods here allow one to do so. The method for completeness is presented. Applying the concepts above, one can compute the uncertainty, i.e. variance, associated with an estimated policy effect when one avoids conditioning on knowing the true model. This variance is written as:

$$var(\hat{\delta}) = \sum_{m \in M} \mu(m|D) var(\hat{\delta}_m) + \sum_{m \in M} \mu(m|D)(\hat{\delta} - \hat{\delta}_m)^2 \quad \ldots(3)$$

This formula illustrates how model uncertainty affects the overall uncertainty one should associate with given parameter estimates. The variance of $\hat{\delta}$ consists of two separate parts. The first, $\sum_{m \in M} \mu(m|D) var(\hat{\delta}_m)$, is a weighted average of the variances for each model and is effectively the same construction as the estimate itself. The second term however, reflects the variance across models in M; this reflects that the models themselves are different. This term, $\sum_{m \in M} \mu(m|D)(\hat{\delta} - \hat{\delta}_m)^2$, is not determined by the model-specific variance calculations. In some sense, it captures how model uncertainty increases the variance associated with a parameter estimate relative to conventional calculations. To see why this second term is interesting, suppose that $var(\hat{\delta}_m)$ is constant across models. In general, one should not conclude that the overall variance is equal to this same value. If there is any variation in $\hat{\delta}_m$ across models, then $var(\hat{\delta}_m) < var(\hat{\delta})$; the cross model variations in the mean increase the uncertainty (as measured by the variance) that exists with respect to δ.

Road Map

The four straightforward steps to implement the theory are described. The basic steps are outlined first and then each of them is discussed.

Process Summary

- Specify distribution
- Specify space of models
- Specify priors for models
- Estimate

Specify Distribution: As with any operational risk loss distribution analysis, one must specify a choice of distribution or class of distributions. This method can in principle be broadened to include uncertainty over the choice of models, but to date the necessary theory has not been developed.[6]

Specify Space of Models: There are two issues to consider in this step. First is the collection of external data and second is the specification of the set of external data to use for estimation. The criterion for suitability is that the data collected can be considered unconditional, draws from the same loss distribution that the internal data was derived. That is, one needs to make the assumption that the externally derived data come from the same 'class' of distributions that characterize the internal data.

The latter of these is straightforward. The appropriate 'space' of models here is the combination of all possible subsets of the set of external data. Begin by identifying a set of external data points, *S*. Then, for each component subset $s \in S$ (e.g. the first, fifth and tenth data point) one has a model, *m*. The collection of all models, *m* comprises the space M. Since the number of models can quickly explode, one may wish for computational reasons to specify that only a subset of models, *m*, with more than a certain number of elements be included.

Specify Priors for Models: For the approach to work, one needs to specify $\mu(m)$ for each model. A value of this approach is that the results, in general, are not sensitive to the selection of priors on the models. Since the determination of the final distribution parameters are done by weighting results by their posterior probabilities, the role of prior specification is minimized. Consider the case of a very unlikely model that is erroneously given a high prior probability. The posterior of this model will nonetheless be very low. The converse is also true. A 'true' model that is given a low prior will nonetheless receive a posterior that approaches 1.

A baseline approach is to specify a 'flat' prior, i.e. one that gives a prior probability of $\frac{1}{n}$, where *n* is the number of models in the space M. From this baseline, one can make some progress toward the determination of appropriate priors, even if one cannot be perfectly precise. An approach that can deliver some added value in this context is the hierarchical priors determination, outlined in Brock *et al.* (2003). They specify that groups of similar models be 'nested' in a hierarchical structure in the determination of priors. As an example, one might want to specify a flat prior across the data points at the risk type level and a corresponding flat prior at the lower level. If the number of risks within each risk type is not the same, this method produces priors that differ from a flat structure.

Estimate: Once the above steps are complete, estimation itself is a simple process. For each model, *m*, contained in the set, M, one simply estimates the parameters of the given

[6] The technical issues are beyond the scope of this paper, but the intuition is that if one wishes to compare and average across distributions, one must find a way to compare the likelihood of a given set of data describing the various distributions. To date, the methods for this are not in place.

distribution using standard maximum likelihood methods. From this process, one obtains not only distributional estimates, but also a log-likelihood value. Once all models, m, have been estimated, the posterior probability can be calculated using equation (2) and the weighted average object of interest δ, (the distribution parameters), can be computed using equation (1).

Conclusion

This note has provided a methodology and process for a robust method of including external data into the calculation of loss distributions. To date, institutions have struggled with notions of how to combine the information included in scenarios, internal and external data into the capital calculation process for operational risk. To the common approach of weighting data using *ad hoc* assumptions on weight values, this paper provides a feasible, theoretically justified alternative.

The method proposed here has associated recommendations relating to data collection. We can summarize the implications for data collection in two components. First, the method suggests that one does not need to find a 'perfect' external dataset under the proposed approach. This derives from the fact that the external data are not being used to determine a distributional form, and their importance in the determination of parameters is dependent on the accompanying data. Second, one wants to collect a wide variety of data without modifications or adjustments. The practice of calibrating external data to fit with internal data trends or patterns is incorporated into the econometric methodology; as such, *ad hoc* adjustments in effect reduce the available information.

Acknowledgment: *Many thanks are due to Jonathan Morse for outstanding research assistance. The views expressed in this paper are solely those of the authors and do not reflect official positions of the Federal Reserve Bank of Boston or the Federal Reserve System.*

References

1. Brock W and Durlauf S (2001), "Growth Economics and Reality", *World Bank Economic Review,* Vol. 15, pp. 229-272.
2. Brock W, Durlauf S and West K (2003), "Policy Evaluation in Uncertain Economic Environments", *Brookings Papers on Economic Activity,* No. 1, pp. 235-322.
3. Cohen-Cole E, Durlauf S, Fagan J and Nagin D (2007), "Reevaluating the Deterrent Effect of Capital Punishment: Model and Data Uncertainty", Federal Reserve Bank of Boston, Manuscript.
4. Draper David (1995), "Assessment and Propagation of Model Uncertainty", *Journal of the Royal Statistical Society,* Series B, Vol. 57, pp. 45-70.
5. Fernandez C, Ley E and Steel M (2001). "Model Uncertainty in Cross-Country Growth Regressions", *Journal of Applied Econometrics,* Vol. 16, pp. 563-576.

6. Hoeting J, Clyde M, Madigan D and Raftery A (1999), "Bayesian Model Averaging: A Tutorial", *Statistical Science,* Vol. 14, pp. 382-401.

7. Leamer E (1978), *Specification Searches*, John Wiley & Sons, New York.

8. Raftery A (1995), "Bayesian Model Selection in Social Research (with discussion)", *Sociological Methodology,* Marsden P (Ed.) Cambridge, MA, Blackwell, pp. 185-195.

9. Raftery A (1996), "Approximate Bayes Factors and Accounting from Model Uncertainty in Generalised Linear Models", *Biometrika,* Vol. 83, pp. 251-266.

10. Raftery Adrian E, David Madigan and Jennifer A Hoeting (1997), "Bayesian Model Averaging for Linear Regression Models", *Journal of the American Statistical Association,* Vol. 92, pp. 179-191.

11. Sala-i-Martin X, Doppelhofer G and Miller R (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach", *American Economic Review,* Vol. 94, No. 4, pp. 813-835.

12. Wasserman L (2000), "Bayesian Model Selection and Bayesian Model Averaging", *Journal of Mathematical Psychology,* Vol. 44, pp. 97-112.

Reference # 37J-2007-12-05-01