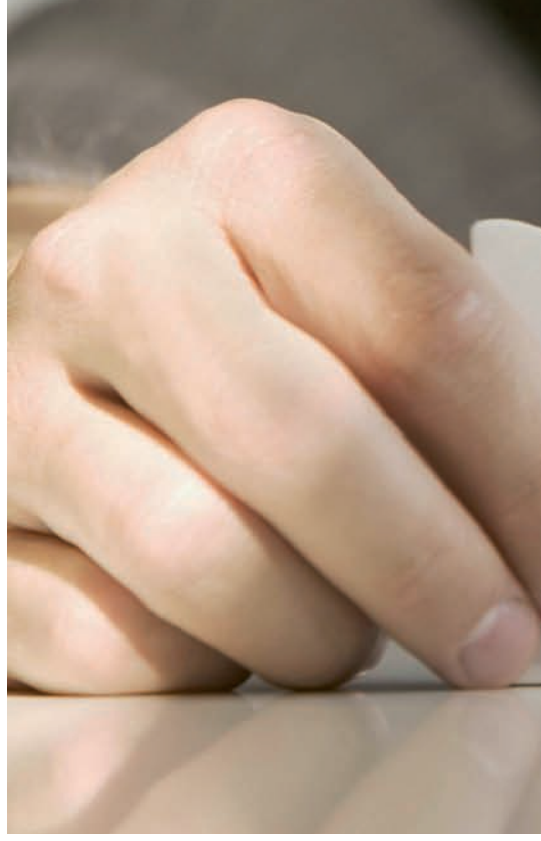

The principles of distribution fitting

Mikhail Makarov examines some of the challenges around fitting distributions properly to op risk models for the advanced measurement approach

AMA operational risk models require translating available information such as internal and external losses, scenarios, business controls and environment into frequencies and severities.

Distribution fitting is a technique often used for such a translation. In this article we will address three issues:

1 Definition of goodness of fit
2 Distributions that should be used for fitting
3 Methods that should be used for fitting

Goodness of fit distance

In order to define what a good fit is, one needs to introduce a measure of closeness for distributions. A well-known example is Kolmogorov-Smirnov (KS) distance:

$$d_{KS}(F,G) = \sup_x |F(x) - G(x)|$$

According to Kolmogorov-Smirnov distance, a cdf F is a good approximation of a cdf G, if the maximal vertical distance between the graphs of the two cdfs is small.

But is Kolmogorov-Smirnov distance a good definition of distribution closeness? The answer depends on the quantities being measured with the help of the fitted distributions. In practical problems, one is interested not in distributions, but in some statistics of the distribution such as mean, quantile, expected loss in a layer, shortfall, etc. From that point of view, a good approximation of the true distribution should give a good approximation to the statistics of interest.

For risk measurement application, computation of quantiles is of significant interest. Typically, one would like to guarantee that the relative error in measuring the quantiles is small. In order to measure goodness of relative quantile approximation, we introduce relative quantile distance:

$$d_{rq}(F,G) = \sup_q \left| \ln\left[F^{-1}(q)\right] - \ln\left[G^{-1}(q)\right] \right|$$

For example, if $d_{rq}(F,G) < 0.1$, then replacing distribution G with distribution F does not lead to more than 10% error in estimation of the quantiles. The relative quantile distance can be visualised by looking at Q-Q plot in log-log co-ordinates.

In frequency-severity models, one is interested in getting a relatively precise estimation for quantiles of the aggregate distribution S

$$S = \sum^{N} X$$

while the fitting is done for a single loss severity X. Fortunately, a relatively small error in estimation of quantiles for single loss distribution leads to a relatively small error in estimation of quantiles for aggregate distribution, ie:

$$d_{rq}\left(\sum^{N} X, \sum^{N} Y\right) = d_{rq}(X,Y)$$

A common mistake is to check goodness of fit for single loss distribution only up to 99.9% quantile. The confidence level of 99.9% should be used for aggregate loss distribution. The confidence level of single loss distribution for which goodness of test has to be tested can be – depending on the frequency – much higher.

Quantile is not the only statistic of interest in risk measurement application. One can be interested in measuring exposure, shortfall etc. Similarly to d_{rq}, one can introduce a distance that guarantees a (relatively) small error in measuring the statistics of interest, and use that distance to define goodness of fit.

Goodness of fit tests

KS distance allows for testing whether a fit *cannot* be used to model the data. One considers a hypothesis H_0 = {Fit is equal to the true distribution} and computes the KS distance between the fit and the empirical distribution. If the KS distance is large, one rejects the hypothesis. It is important to understand that if the distance is small, one cannot accept the hypothesis (the test does not have the power to accept). A big advantage of the KS test is that it does not require the knowledge of the distribution type for the true distribution. To summarise, the KS test can be used to exclude really bad fits to the data, but it cannot be used to conclude that a given fit is the right one.

A number of other tests (eg, Anderson-Darling) were designed to test hypothesis

H_0 = {Fit is equal to the true distribution}. Unfortunately, in order to use those tests, one needs to know the distribution type of the true distribution, so what those tests gain in power, they lose in flexibility.

For many applications one does not need to test for equality – in fact a much weaker hypothesis of approximate equality needs to be tested. For example, if G denotes the distribution and F denotes the fit, one could be interested in testing the following hypothesis:

$$H_0 = \left\{ d_{rq}(F,G) < 10\% \right\}.$$

Even for this weaker hypothesis, no tests are currently available.

It is worth mentioning that the graphical goodness of a test such as Q-Q plot should only be used for exploration. Even a person with a lot of experience can easily make a mistake judging goodness of fit from a graph. For example, it is not clear how to reject a fit by looking at Q-Q plot only.

Which distributions should be used for fitting?

Distribution fitting is performed in one of the following situations:

1 Distribution type is known and only parameters are to be determined.

2 Distribution type is not known; nevertheless a distribution is fitted with an aim to get a sufficiently good approximation to the true distribution.

The first case occurs when analysis on many similar data sets has been performed before, and the distribution type determined. Alternatively, there may be some theoretical considerations that lead to a selection of a particular distribution type: for example, it follows from a physical model, such as: the motion of a small particle should follow a Brownian motion.

The second case – the distribution type is not known – is much more common in practical applications, but is not well-studied. What if the distribution type selected for fitting is wrong? Can one still hope for a reasonable approximation of the true distribution? The only approach available at the moment that gives some answers to those questions is based on extreme value theory (EVT).

According to EVT, a generalised Pareto distribution (GPD) can be used to approximate tails of a wide class of distributions. As GPD can approximate almost any tail, it seems reasonable to take large losses and fit them with GPD. Surely, even if the loss distribution is not GPD, it will be well approximated by the fit. Unfortunately, the situation is not that easy. The problem lies in the definition of 'approximately' used by EVT: it guarantees a weak distribution convergence that is not sufficient to approximate quantiles, shortfalls and expected value.

Take as an example distribution:

$$G(x) = 1 - \frac{e}{x \ln^2 x}, \quad x \geq e$$

Given a threshold u, the tail of G is represented by the excess distribution G_u

$$G_u(x) = \frac{G(x+u) - G(u)}{1 - G(u)}$$

For a sufficiently high threshold, the excess distribution can be approximated by GPD distribution in the sense of Kolmogorov-Smirnov distance

$$\lim_{u \to \infty} d_{KS}\left(GPD, G_u\right) = 0$$

However, for any threshold, GPD cannot estimate high quantiles of the excess distribution G_u as

$$d_{rq}\left(GPD, G_u\right) = \infty$$

Therefore, GPD does not give sufficiently good approximation to high quantiles. A number of empirical studies indicate that in practice, GPD often overestimates quantiles, and suggest using more flexible distributions. Examples of more flexible distributions are generalised beta, loggamma, g-and-h. Using Taylor series as an analogy, one can think of GPD as the first (and most important) term in the series, while distributions such as g-and-h add a second term

in the series. Unfortunately there is no systematic classification of distribution beyond GPD (the first term in the series), so one is forced to test via trial-and-error with different distributions.

Which methods should be used for fitting?

Theoretically, if the type of the true distribution is known, then the maximum likelihood method (ML) provides the best estimation for the parameters. Under very general assumptions, ML leads to an estimator that is:

(a) Asymptotically unbiased, ie, bias tends to zero as number of samples increases.
(b) Asymptotically efficient, ie, has lowest mean squared error for unbiased estimators.
(c) Functional invariant, ie, ML for parameters leads to ML for any other statistics such as quantiles.

Clearly, the properties (a)–(c) indicate that ML is the best fitting method if the distribution type is known, and the data does come from the distribution.

Although ML is the best estimator under perfect conditions, it is not robust – it can produce completely wrong results for polluted data or in cases when the distribution type was not selected appropriately. To demonstrate the non-robustness of ML, consider fitting losses $x_1, \ldots, x_n$ with lognormal distribution. ML method leads to the following estimate of the first parameter:

$$\mu = \frac{1}{n}\sum_{k=1}^{n}\ln(x_k)$$

The problem with this formula is that if x_1 is equal to zero – or very close to it – then μ is not defined or is a very small number. Thus by including a very small number in the data set, one can completely change the estimates of the parameters and statistics such as quantiles.

Even if the original data set does not contain very small values, one can introduce them by selecting the fitting threshold very close to the smallest loss.

Due to the non-robustness of ML, one needs to verify the results by determining whether the data set contains outliers, perform bootstrapping to check if the confidence intervals for the parameters are unusually high, perform visual tests etc. For example, the data set may contain an unusually high loss (eg, a one-in-100-years event in the data set over five years). One needs to test whether the fitting results are sensitive to the high loss and if so, assign a smaller weight to the loss.

Thanks to Basel II and the effort of the banks with regard to data collection, the days when proper statistical techniques cannot be used are over

Several fitting methods more robust than ML have been developed. However, their properties are often not known, and ML after-data cleansing seems to be more appropriate.

A helpful approach for the ML method is an excess of threshold fitting. There are two reasons to fit distribution in excess of a threshold. First, many distributions are not flexible enough to fit huge data sets. Second, and more importantly, there is no reason to believe that small losses (high frequency, low severity) carry any useful information about large losses (low frequency, high severity).

If small losses are irrelevant for tail behaviour, there is no reason to use them for fitting. An excess of threshold fitting can not only simplify modelling, it can also be used to test the robustness of the model. Sensitivity of the model to small changes in threshold selection signals underlying problems, either with the data or distribution selection. One can test the sensitivity of the model to the threshold selection as a simple version of the bootstrap method.

Conclusion

Thanks to Basel II and the effort of the banks with regard to data collection, the days when proper statistical techniques cannot be used are over. Instead, we are faced with the challenge of developing sound methodology for data analysis. Most of the current papers that fit distributions to data encounter the three challenges raised in this article. First, often only visual techniques are used to check the goodness of fit. Second, distributions are selected using a trial-and-error approach. Third, distributions are often fitted using the ML method, without checking robustness. In order to get best results from the available data, practitioners need to further investigate the three issues raised: which goodness of fit, which distribution, and which fitting method.

Mikhail Makarov works for Switzerland-based EVMTech.
E-mail: mmakarov@evmtech.com