Institut canadien des dérivés
Canadian Derivatives Institute

L'Institut bénéficie du soutien financier de l'Autorité des marchés financiers ainsi que du ministère des Finances du Québec

Note technique

NT 16-01

Big Data and Risk Management in Financial Markets: A Survey

Avril 2016

Cette note technique a été rédigée par

Francesco Corea, LUISS Guido Carli University

Big Data and Risk Management in Financial Markets: A Survey

Francesco Corea[1]

Department of Economics and Finance, LUISS Guido Carli University

Abstract

Big data is a buzzword that indicates data that do not fit traditional database structure. Their potential is enormous for many fields, and risk management is within the ones that could benefit the most from new sources of unstructured data. This paper introduces the big data framework, terminology, and technology, in order to understand the upsides and challenges that they pose to financial markets. A review of standard methods and tools in risk management is then provided, in order to be able to understand the revolution brought into the environment by big data. Simulation and forecasting are the two areas that are affected the most, and therefore the ones of interest for this study.

Keywords: large datasets, XVA, big data, Monte Carlo simulation, forecasting

JEL Classification: B23, C58, C82, G14, G17, G32

[1] Corresponding author; email: fcorea@luiss.it.

1. Introduction: Why big data are relevant to risk management

The exponential increase in the amount of data created in the last few years impacted every sector. Big data, i.e., datasets so large to not be eventually fit into traditional databases, revolutionized the way we deal with every decision making process, as well as approach to several business and research questions. It has been proved how big data can be used in context such as contagion spreading (Culotta, 2010), music albums success predictions (Dhar and Chang, 2009), or presidential election (Tumasjan et al., 2010). Even though the list of innovative applications can fill many pages, one of the greatest changes brought by big data concerns financial markets. For example, many works have been implemented using sentiment analysis, i.e., whether a piece of text reflects a positive/negative sentiment (Bollen et al., 2011; Corea and Cervellati, 2015), and several algorithmic trading companies have been started with the help of massive datasets. Hence, even though many innovative steps have been undertaken in financial markets, not all the categories have been affected in the same way: in particular, new information and stack of technologies did not bring as many benefits to the risk management as they did to credit fraud detection for example.

Risk is usually addressed from an operational perspective, from a customer relationship angle, or with a specific focus on preventing fraud and credit scoring. However, applications strictly related to financial markets are still not so common: even if in theory more information should entail a higher degree of accuracy, in practice it also exponentially augments the system complexity, and makes extremely hard to

identify and analyze timely unstructured data that are valuable. The Deposit Trust Clearing Corporation had recently identified many potential risks in financial markets, and a vaster amount of data can help institutions and banks in addressing them: high-frequency trading risk, liquidity and credit risks, collateral risk, counterparty risk, only to name a few. Markets are always more interconnected, which also increases the risk of a network systemic failure: more and more data can help central institutions and regulators to predict in real-time symptoms of a future crisis, and acting on time to prevent it or weaken it.

A particular field of interest regards the market and counterpart risks. The huge complexity introduced in the market made the common pricing techniques obsolete and slowly reactive, and required a more comprehensive pricing approach than a net discounted value of derivative's legs. This is the reason why banks and financial institutions need (but struggle) to simulate a single portfolio a hundred thousand times, or because an accurate fast forecast is considered a breakthrough achievement.

The purpose of this paper is to give an overview of common risk management models, and of how simulation and forecasting models are modified and improved by a larger amount of data. The next section presents a primer on the technologies used in the big data space, while the following one deals with traditional risk management models, metrics and tools. Section 4 shows some advancement in big data space, with a focus on simulation and forecasting. Section 5 finally sums up and concludes.

2. Big data technologies

Technology plays an important role into this field, and the transformations we have taken part of are outstanding: the data traffic has been redirected to the cloud, through a shared pool of connected storage devices; parallel computing method helped in computing larger amount of data, reaching a greater accuracy for the same costs; and finally, unstructured data have been taken into account, thanks to the implementation of new developments in the data architecture space. In particular, few technologies seem to be commonly used across financial markets player: Hadoop in undoubtedly the most used platform for managing unstructured data in a parallel computing setting. Along with MapReduce, it allows for an efficient data cleansing and massaging, as well as for complicated or CPU-demanding computations: the task is indeed split between many servers (independent machines with their own memories and operating system), computed locally, and the results are reaggregated afterwards. A second important tool in a non-relational database named NoSQL, which does not use structured query language and allows to consider data that would not usually fit a standard table.

Those technologies allow for improvements in timing and predictive intelligence into the risk management field. The degree of innovation these tools and techniques are bringing is remarkable, especially when it comes to real-time simulation and large–volume forecasting. Section 3 and 4 show how big data are actually used in those applications.

3. Traditional risk management methods

We briefly survey the most common techniques and methods contained in a risk manager toolbox. The initial approach to begin with is called Value-at-Risk (VaR), which is used to assess the market risk. In a nutshell, the VaR is a statistical technique used to measure the level of risk of a portfolio given a certain confidence interval and within a fixed time frame.

From a more technical perspective, the VaR is a threshold value such that the potential loss over a specified time period is equal to a given probability. In other words, given the confidence level α, the VaR is that number k that makes the probability of a loss L greater than k be exactly equal to $1 - \alpha$:

$$VaR_{\alpha} = inf\{k \in \mathbb{R}: \Pr(L > k) \leq 1 - \alpha\} \quad (1)$$

Many different variations have been proposed over the years, and a particular attention has to be devoted to two *coherent risk measures* alternatives, i.e., the conditional VaR (CVaR) and the entropic VaR (EVaR). The CVaR, also called *expected shortfall*, indicates for a certain probability level the expected return of the portfolio in the worst scenarios:

$$CVaR_{1-\alpha} = \frac{1}{\alpha}\int_0^{\alpha} VaR_{1-\gamma}(X)\, d\gamma \quad (2)$$

The EVaR (Ahmadi-Javid, 2011) instead represents the upper bound for both VaR and CVaR, and its dual representation is related to the concept of relative entropy:

$$EVaR_{1-\alpha} = inf_{z>0}\left\{\frac{1}{z}\ ln\frac{M_L(z)}{\alpha}\right\} \quad (3)$$

where $M_L(z)$ is the moment-generating function of the loss.

The VaR, regardless of the type, is usually computed through either historical method, the Delta-Normal one, or Monte Carlo simulation. The first method just lists historical returns in ascending order, while the Delta-Normal technique looks back in the time series, computes mean, variance, and correlation, and finally obtains the portfolio risk through a combination of linear exposure to factors and the covariance matrix (Jorion, 2006). The last method, i.e., the Monte Carlo simulation, is probably the most used nowadays, as well as the most interesting from a big data perspective. It actually requires to develop a model for the stock price/returns trajectories, then runs a multitude of simulated trials and averages the results obtained.

The Monte Carlo is then a repeated sampling algorithm that could be used for solving any problem that may be stated through a probabilistic lens, and which is often exploited for pricing extremely complex derivatives.

More recent advances in risk management tools are related to credit counterparty risk. The framework known as the X-Value Adjustment (XVA) includes credit valuation adjustment (CVA), debt valuation adjustment (DVA), and funding valuation adjustment (FVA), and respectively deals with the risk of the counterparty, the risk of the entity itself, and the market value of the funding cost of the instrument

(Hull and White, 2014; Smith, 2015). According to Albanese et al. (2011), the CVA is defined as

$$CVA = E_0\left[\sum_n \int_0^\infty e^{-\int_0^t r_s ds}\,(P_t^n) + d\pi_t^n\right] \qquad (4)$$

where *P* is the price process of the portfolio, *n* is the index for netting sets, *r* is the short rate and π is the process of the cumulative probability of default. The intuitive interpretation for the CVA is that it represents the market value of counterparty credit risk, and it is obtained as the difference between the risk-free portfolio and the portfolio that embeds a potential counterparty's default. The DVA is instead defined in Smith (2015) as the expected loss of the firm if the bank defaults - contrarily to the CVA, which represents the expected loss in case of counterparty's default. Hence, the value of the transaction is affected by imbalances between those two measures.

Finally, the FVA represents the difference between funding costs and benefits, or, alternatively, the difference between the value of a portfolio of uncollateralized transactions calculated with the risk-free rate and that same value calculated with the bank average funding cost (Hull, 2015).

The problem with XVA measures is that they require a huge amount of computation power to be calculated effectively. Even though for a standard portfolio CVA calculation a reasonable number of simulations are required, banks might need to run many more deals if they want to take into account all the path-dependent

derivatives in their portfolios - the number would be indeed close to several hundred thousand simulated paths (Green, 2015; Veldhoen and De Prins, 2014).

It is clear then from this short survey of methods that two issues arise from the standard techniques: it is hard to predict the VaR because of parameters estimation, and the simulations are hard to handle because of the tradeoff between accuracy and computational effort.

4. Big data methods: The econometrics of large datasets

Big data are everywhere. Social media provide for example an endless source of information for financial market, because the market moves following the actions of the crowd. Veldhoen and De Prins (2014) claim that different data affect different risks with a distinctive intensity. The following table summarizes their findings rating from 1 (the feature with the strongest impact) to 4 (the weakest benefit) the impact of each characteristic on each risk (each cell is evaluated independently from the others):

Risk Area	Volume	Velocity	Variety	Veracity
Credit Risk	1	4	3	4
Market Risk	3	3	4	4
Operational Risk	3	4	4	3
Compliance	2	3	2	2
Asset-Liability risk management	2	4	3	4

Table 1. Impact of big data features on risk management (Veldhoen and De Prins, 2014).

The availability of those endless data cannot solve every single problem, and in fact big data poses as many technical challenges as well as opportunities for organizations and regulators (Hassani and Silva, 2015), such as lack of skills, issues related to hypothesis/testing/model, or hardware/software challenges. Silver (2013) identifies as the main challenge the increase of noise into the signal ratio, to the detriment of the actual predictive power of the additional data. The forecasting techniques then have to be able to filter down that noise and leave the model with only the variables and data that matter, and at the same time to provide accurate out-of-sample forecasts without abusing of a large number of predictors (Einav and Levin, 2013). In addition, according to Varian (2014), conventional statistics techniques face two additional issues when big data are added to the equation: a higher degree of data manipulation is required, because every data problem is exponentially amplified, and large data allow for different relationships than linear ones.

Hence, the goal of the next two subsections is to provide a summary of models that prevent over-fitting and that are able to manage large datasets efficiently. In general, simpler models work better for out-of-sample forecasts, and excessive complexity should be avoided.

4.1 Big data simulation

Scenario simulations considering huge data amounts allow for an efficient realization of risk concentrations and quicker reactions to new market developments. In particular, Monte Carlo simulation is a powerful and flexible tool, and the challenge

with that is finding the optimal number of paths to match speed and accuracy. A higher accuracy is achieved by the larger amount of simulation the model can project, but it has been always bounded by a lower processing speed as well as machine memory. Even though a set of techniques have been used to handle this burden, the only solution lies in splitting the data between many different *workers*.

Luckily, parallel computing is gaining popularity, and many algorithms for making it less expensive have been developed in the last few years (Scott et al., 2013). Hence, two main methods may be used in order to relief a single terminal from a great data burden: either it can be divided into different cores on the same chip, or it can be divided through different machines. In the first case, the splitting can be made on multi-core CPU, or on parallel GPU (Scott et al., 2013). In any of the two cases, few problems arise: difficulty in writing the splitting configuration, absence of positive effect on memory, and difficulty in abstraction make those methods cumbersome to be used. The second alternative instead is much more scalable: dividing data into different machines increases the processing power and efficiency, although it comes with a higher cost. A solution to this problem has been proposed in Scott et al. (2013), called consensus Monte Carlo: this new model runs a separate Monte Carlo algorithm in each terminal, and then averages individual draws across machines. The final outcome resembles a single Monte Carlo set of simulations run on a single machine for a long time.

The future for Monte Carlo methods presents many possible developments. According to Kroese et al. (2014), at least three different elaborations can be pursued:

quasi Monte Carlo, rare events, and spatial processes. Quasi Monte Carlo uses quasi-random number generators especially in multi-dimensional integration problems; rare events will instead consider using simulations to spot events that rarely happen using variance reduction techniques; and finally, spatial processes are difficult to approximate because of the lack of independence between the simulations themselves, and a convergence is only achievable through an enormous number of simulations.

4.2 Big data forecasting

Many completely new methods from one hand, and techniques borrowed from other disciplines from the other, are nowadays available to researchers and practitioners in order to predict future outcomes. In line with Eklund and Kapetanios (2008), who provided a good review of all these methods, we adopted their classification of forecasting methods into four groups: single equation models that use the whole datasets for estimation purposes; models that use only a subset of the whole database, even though a more complete set is provided; models that use partial datasets to estimate multiple forecasts averaged later on in a conclusive result; and finally, multivariate models that use the whole datasets with the aim of estimating a set of variables.

The first group is quite wide, and includes common techniques used differently, such as ordinary least square (OLS) regression or Bayesian regression, as well as new advancements in the field, as in the case of factor models. In the OLS model, when the time series dimension exceeds the number of observations, the generalized inverse has

to be used in order to estimate the parameters. Bayesian regression (De Mol, Giannone, and Reichlin, 2008) starts instead from a *prior* probability, and updates this likelihood through an incremental amount of observations to finally obtain a *posterior* probability. Of particular interest in this type of regression is the ability to use the whole dataset reducing though the magnitude of parameters, and in this way providing a lower variance with respect to standard estimators. Finally, a part of this first group is the factor models (Stock and Watson, 2002). These models are able to select ex-ante which data contains the greatest predictive power, and use then only few variables to form the forecasting equation. Famous models of this kind are the principal component analysis (PCA), its dynamic version, and the subspace methods. The PCA technique estimates the matrix of the linear combinations of all factors through their eigenvectors, and then considers only the first k ones with the greatest weight (*loading vector*). If in place of the covariance matrix the spectral density one with different frequencies is used, the technique is called dynamic PCA. The subspace method starts instead from a parametric state space model: a simple multivariate OLS model estimates the coefficient, while the factors are then obtained through a reduced rank approximation (Kapetanios and Marcellino, 2003).

In the second group, we expect the most of the data to be noisy and not really meaningful, thus we use features or variables selection models to identify ex-ante the more significant predictors. This class of methods provides then a stratagem to avoid, if wanted, the problem to deal with large datasets. The dimensionality reduction allows indeed the use of the simple linear forecasting model on the most appropriate subset of

variables, which are skimmed down through information criteria. Boosting, LASSO, Least Angle Regression (LAR), are some of the most common techniques belonging to this class. Boosting entails running univariate regressions for each predictor, and selecting the model that minimizes a chosen loss function. The residuals thus obtained are then explained through running a second round of regressions using the remaining predictors, and selecting again the one that minimizes the same loss function as before. The process is repeated until a certain information criterion is met. On the other hand, the LASSO and LAR apply penalty functions to the regressions. LASSO expects the norm of estimated vector to be less than a specified shrinkage threshold; the LAR works similarly to the boosting, even though at each step is not included a new variable, but the relative coefficient is incremented by the amount that makes it not to minimize the loss function anymore.

Many other procedures exist, such as stepwise regression, ridge regression, or still the more exotic genetic algorithm and simulated annealing (Kapetanios, 2007). The stepwise regression estimates the model starting from no variable (forward regression) or from all-variable model (backward), adding and subtracting at each step the variable that improve the model the most, and keeping repeating the procedure until the process stops. The ridge regression instead includes a quadratic term, which penalizes the size of the regression coefficients. The genetic algorithm (Dorsey and Mayer, 1995) is an optimization method that iterates towards a solution selecting, in the same fashion of natural selection, only the best fitting features. Finally, simulated

annealing creates a nonhomogeneous Markov chain using the objective function that converges to a maximum/minimum of the function itself.

On the other hand, the third group is the averaging model one, which embeds mainly the Bayesian model averaging (BMA) and frequentist model averaging. While the frequentist approach entails the creation of model confidence sets from which model likelihood can be derived, the BMA methodology provides a Bayesian framework for the combination models. Different combinations of relationship between predictors and dependent variables are estimated, and weighted altogether in order to obtain a better forecasting model, with weights corresponding to the posterior probabilities.

Finally, the last group uses the whole datasets to estimate a set of variables (Carriero, Kapetanios, and Marcellino, 2011). Reduced rank regression, Bayesian VAR, and Multivariate Boosting, as already proposed in Eklund and Kapetanios (2008), are only few of the models belonging to this class. Reduced rank regression works similarly to a set of classic Vector Autoregressive models, but as soon as underlying dataset is large, those models become quite noisy and rich of insignificant coefficients. Therefore, a rank reduction can be imposed, constraining the matrix of the coefficients in the VAR model to be much smaller than the number of predictors. Differently from this model, which compressed the greatest informative value into few predictors, the Bayesian counterpart of the VAR focuses instead on constraining the data – imposing the restrictions as priors –, although it maintains a dependency between data and coefficient determination. On the other hand, the multivariate boosting is quite similar

to its simple version explained beforehand. The main difference lies in measuring at each step a multivariate model (instead of a single equation as in simple boosting), starting from a zero coefficient matrix and recursively setting the single coefficients that explained better the dependent variables to be non-zero.

Many models have been reviewed in the previous two sections, but the list is not exhaustive. Indeed, according to Varian (2014), classification and regression trees (CART), bagging, bootstrapping, random forests, and spike and slab regression are other models commonly used in machine learning applications. Decision trees are used when it is necessary to describe a sequence of decisions/outcomes in a discrete way, and they work quite well in nonlinear cases and with missing data issue. On the other side, bootstrapping estimates the sampling distribution of some statistics through repeated sampling with replacement, while bagging averages different models obtained with manifold bootstrapping of different sizes. The random forests approach is a machine learning technique that grows a tree starting from a bootstrap sample, and selects at each node a random sample of predictors. This process is repeated several times, and a majority vote is applied for classification purposes. Finally, spike and slab regression is another Bayesian hierarchical model that specifies a (Bernoulli) prior distribution of the probability of a set of variables to be included in the model – this is called *spike*, i.e., the probability of a coefficient to be non-zero. Afterwards, a second prior is selected, i.e. a prior on the regression coefficient deriving from the previous choice of certain variables (the *slab*). Combining the priors and repeating the process

multiple times, it is possible to obtain two posterior distributions for both the priors, as well as the prediction for the overall model.

4.3 Other considerations on big data modeling

Varian (2014) suggests few additional points that should be considered as the dataset becomes larger: causal inference (Angrist and Pischke, 2009) and model uncertainty.

Causation and correlation are two far distinct concepts, and the issue with large noisy datasets is that it is much easier to spot out spurious correlations that do not have real meaning or practical validation. Random-control groups may be necessary in those cases, but sometimes a good predictive model can work better (Varian, 2014).

The second feature pointed out is that, when it comes to large datasets, averaging models may be more effective than choosing a single one. Usually a simple unique representation is analyzed because of data scarcity, but nothing prevents the decision maker from using multiple models as soon as more data allow it.

5. Conclusions

The huge volume and velocity of data coming from a variety of different sources has forced financial industry to react fast, optimizing current practices, establishing new standards, and exploiting new technologies to achieve the computational power needed to design and deal with new instruments. The management of market risk fosters regulators and financial institutions to adopt new compliance processes: the

data storage step became more structured and wide, the analysis deeper and more accurate, the simulations more extensive and significant, and cross validation more meaningful and tested.

A set of efficient forecasting and simulation tools become therefore essential in handling correctly the increasing complexity in financial markets. In this paper a review of most common concepts and techniques in risk management has been provided, and then an extensive list of models useful in a big data setting has been made available. Although is not possible to identify an algorithm that works the best in every situation (a kind of *fit-them-all* procedure), a combination of different models can be used efficiently to increase the model accuracy, reliability, and practical utility.

Big data are clearly not a panacea for all the business and markets issues, but they offer for sure a new view on things, enlarging the spectrum of relationship we are able to analyze, providing new horizons and directions for professional to act in a more efficient, ethic, and compliant way, and increasing our knowledge of the markets themselves.

References

1. Ahmadi-Javid, A. (2011). “Entropic Value-at-Risk: A New Coherent Risk Measure”. Journal of Optimization Theory and Applications 155(3): 1105-1123.
2. Albanese, C., Bellaj, T., Gimonet, G., Pietronero, G., (2011). “Coherent Global Market Simulations and Securitization Measures for Counterparty Credit Risk”. Quantitative Finance 11 (1): 1-20.
3. Angrist, J. D., Pischke, J. S. (2009). “Mostly Harmless Econometrics”. Princeton University Press.
4. Bollen, J., Mao, H., Zeng, X. (2011). "Twitter mood predicts the stock market". Journal of Computational Science Volume 2 (1): 1-8.
5. Carriero, A., Kapetanios, G., Marcellino, M. (2011). “Forecasting Large Datasets with Bayesian Reduced Rank Multivariate Models”. Journal of Applied Econometrics 26 (5): 735-761.
6. Corea, F., Cervellati, E. M. (2015). "The Power of Micro-Blogging: How to Use Twitter for Predicting the Stock Market". Eurasian Journal of Economics and Finance, 3 (4): 1-7.
7. Culotta, A. (2010). "Towards detecting influenza epidemics by analysing Twitter messages". Proceedings of the First Workshop on Social Media Analytics: 115-122.
8. De Mol, C., D. Giannone, and L. Reichlin (2008): “Forecasting using a large number of predictors: is Bayesian regression a valid alternative to principal components?”. Journal of Econometrics 146: 318-328.
9. Dhar, V., Chang, E. A. (2009). "Does Chatter Matter? The Impact of User-Generated Content on Music Sales". Journal of Interactive Marketing Volume 23 (4): 300-307.
10. Dorsey, R. E., Mayer, W. J. (1995). “Genetic Algorithms for Estimation Prob- lems with Multiple Optima, Nondifferentiability and Other Irregular Features”. Journal of Business and Economic Statistics, 13 (1): 53-66.
11. Eklund, J., Kapetanios, G. (2008). “A Review of Forecasting Techniques for Large Data Sets”. Queen Mary working paper no. 625: 1-18.
12. Einav, L., Levin, J. D. (2013). “The Data Revolution and Economic Analysis”. Working Paper No. 19035, National Bureau of Economic Research.
13. Green, A. (2015). “XVA: Credit, Funding and Capital Valuation Adjustments”. Wiley, 1st edition.
14. Hassani, H., Silva, E. S. (2015). “Forecasting with Big Data: A Review”. Annals of Data Science 2 (1): 5-19.
15. Hull, J. (2015). “Risk Management and Financial Institutions”. Wiley, 4th edition.

16. Hull, J., White, A. (2014). “Valuing Derivatives: Funding Value Adjustments and Fair Value”. Financial Analysts Journal, Vol. 70 (3): 46-56.
17. Jorion, P. (2006). “Value at Risk: The New Benchmark for Managing Financial Risk”. (3rd ed.). McGraw-Hill.
18. Kapetanios, G. (2007). “Variable Selection in Regression Models using Non-Standard Optimisation of Information Criteria”. Computational Statistics & Data Analysis 52: 4-15.
19. Kapetanios, G., and M. Marcellino (2003): “A Comparison of Estimation Methods for Dynamic Factor Models of Large Dimensions.” Queen Mary, University of London Working Paper No. 489.
20. Kroese, D. P., Brereton, T., Taimre, T,, Botev, Z. I. (2014). “Why the Monte Carlo method is so important today”. WIREs Computational Statatistics 6: 386-392.
21. Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E., McCulloch, R. (2013). “Bayes and big data: The consensus Monte Carlo algorithm”. EFaBBayes 250 conference 16.
22. Silver, N. (2013). “The Signal and the Noise: The Art and Science of Prediction”. Penguin Books, Australia.
23. Smith, D. J. (2015). “Understanding CVA, DVA, and FVA: Examples of Interest Rate Swap Valuation”. Available at SSRN: http://ssrn.com/abstract=2510970.
24. Stock, J. H., Watson, M. W. (2002). “Macroeconomic Forecasting Using Diffusion Indices.” Journal of Business and Economic Statistics, 20, 147–162.
25. Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M. (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media: 178-185.
26. Varian, H. R. (2014). "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives, 28 (2): 3-28.
27. Veldhoen, A., De Prins, S. (2014). “Applying Big Data to Risk Management”. Avantage Reply Report.