Journal of Operational Risk **14**(4), 1–42
DOI: 10.21314/JOP.2019.232

Research Paper

Estimation of value-at-risk for conduct risk losses using pseudo-marginal Markov chain Monte Carlo

Peter Mitic[1,2,3] and Jiaqi Hu[4]

[1]Santander UK, 2 Triton Square, Regent's Place, London NW1 3AN, UK;
email: peter.mitic@santandergcb.com

[2]Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

[3]Laboratoire d'Excellence sur la Régulation Financière (LabEx ReFi), 12 Place du Panthéon, 75231 Paris Cedex 05, France

[4]Department of Statistics, University of Oxford, 24–29 St Giles', Oxford OX1 3LB, UK; email: 709839951@qq.com

(Received November 26, 2018; revised May 22, 2019; accepted July 22, 2019)

ABSTRACT

We propose a model for conduct risk losses, in which conduct risk losses are characterized by having a small number of extremely large losses (perhaps only one) with more numerous smaller losses. It is assumed that the largest loss is actually a provision from which payments to customers are made periodically as required. We use the pseudo-marginal (PM) Markov chain Monte Carlo method to decompose the largest loss into smaller partitions in order to estimate 99.9% value-at-risk. The partitioning is done in a way that makes no assumption about the size of the partitions. The advantages and problems of using this method are discussed. The PM procedures were run on several representative data sets. The results indicate that, in cases

Corresponding author: P. Mitic

Print ISSN 1744-6740 | Online ISSN 1755-2710

where using approaches such as calculating a Monte Carlo-derived loss distribution yields a result that is not consistent with the risk profile expressed by the data, using the PM method yields results that have the required consistency.

Keywords: pseudo-marginal (PM); Markov chain Monte Carlo (MCMC); importance sampling; value-at-risk (VaR); loss distribution; conduct risk.

1 INTRODUCTION

Modeling distributions for conduct risk has proved to be a challenge because the losses involved have particular properties, which will be noted in Section 1.3. Conduct risk forms part of Basel II risk category 4.2 ("Clients, Products & Business Practices" (CPBP)). The full Basel II risk class taxonomy may be found in Basel Committee on Banking Supervision (2006). The general guidance for modeling operational risk losses is given in Basel Committee on Banking Supervision (2011), although conduct risk is not treated as a special case. However, its characteristics set it apart from other risk classes. Basel Committee on Banking Supervision (2011) specifies that losses corresponding to any particular risk class should be modeled using fat-tailed severity distributions and frequency distributions that are appropriate for rare events. The fat-tailed category comprises distributions such as lognormal, Weibull and generalized Pareto. The class of frequency distributions centers on Poisson and negative binomial. All of these distributions have been routinely used in the insurance industry since the 1960s. A summary of usage in the insurance industry may be found in, for example, Alexander (2001) or Gay (2004). Modeling techniques used in the insurance industry have been widely adopted for operational risk modeling since 2000 (Cruz *et al* 2015).

The proposed introduction of the standardized measurement approach (SMA) for calculating operational risk capital effectively eliminates the need to do any modeling of operational risk (conduct risk included). However, operational risk modeling is still needed for prudent risk management. That is the purpose of the proposals in this paper. The accepted metric for both prudent risk management and regulatory capital assessment has been the value-at-risk at 99.9% (specified in Basel Committee on Banking Supervision (2011)), and that remains our target.

1.1 Use of the term "VaR"

In this paper, the term "VaR" should be taken to mean value-at-risk at 99.9%, unless some other percentage is indicated. The term "regulatory capital" is used interchangeably with value-at-risk at 99.9%. Given the onset of the SMA, the term

"prudential capital" might be more appropriate, since the purpose of the processes discussed in this paper is to manage risk capital in a prudential way.

1.2 Contribution of this paper

Conduct risk losses have become increasingly prominent in calculations of regulatory capital during the past decade. There is a specific reason for this: they can be extreme, and those extremities have a significant effect on the outcome of calculations. In many cases, these outcomes are not consistent with an established risk profile: the resulting capital may be too small to satisfy regulatory authorities, or unnecessarily large so that bank lending is impeded. Crude methodology adjustments and approximations can be made in order to address this problem, but it is difficult to justify them satisfactorily. The method presented in this paper – the "pseudo-marginal" (PM) Markov chain Monte Carlo (MCMC) approximation – is an attempt to calculate a satisfactory regulatory capital without making adjustments and approximations that cannot be justified. The PM technique has had few practical applications to date. Therefore, the results reported here should be seen as a significant step forward in conduct risk modeling and, more generally, in applied statistics.

1.3 The characteristics of conduct risk losses

The characteristics of conduct risk losses that set them apart from other risk classes are the following:

(1) severe aggregation of multiple small losses into single extremely large losses (often these aggregations take the form of provisions, which are discussed in Section 2.2);

(2) low annual frequency (sometimes less than fifty per year);

(3) a small number of large losses (often only one) that are significantly larger (two or more orders of magnitude) than their counterparts;

(4) in many cases the largest losses comprise subcomponents that cannot be used directly, either because they are not known precisely or because using them would invalidate Basel II regulations.

The above characteristics make conduct risk losses an extreme example of operational risk losses. Therefore, alternative ways to assess regulatory capital for them are needed. The aim of this paper is to present an alternative way to model conduct risk, using the PM Monte Carlo method. This is a relatively new statistical technique, which first appeared in 2009 and has not, to our knowledge, been applied in any financial context before. A detailed explanation will be given in Section 5.2. The

PM method is particularly well suited for modeling conduct risk because it specifically addresses the points listed above. In particular, it is capable of analyzing the effect on regulatory capital of a single loss that is significantly larger than all the others.

In principle, the characteristics listed above are not exclusive to conduct risk losses. We have, for example, observed external fraud data sets that have severe outliers. In nonfinancial contexts, the Fitch operational loss database includes exceptionally large losses that dwarf all others. The BP Gulf of Mexico oil leak is one of them. Therefore, although the references in this paper are to conduct risk, that category is generalizable to any data set that satisfies the above characteristics.

In this analysis, we seek to solve the problem of how to determine VaR for conduct risk losses that have the particular characteristics referred to above, such that the calculated capital is consistent with a bank's risk profile. The phrase "consistent with a bank's risk profile" means that the capital should not be so high that it inhibits lending significantly, and it should not be so low that it cannot cover unexpected further losses. Currently, a minimum capital value can be calculated using empirical losses only (the "empirical bootstrap"; see Mitic and Bloxham (2018, Algorithm 1)). If the calculated capital is less than the empirical bootstrap value, the latter would be accepted in preference to the calculated value. The empirical bootstrap value is used as a "reasonableness test" for VaR calculated by other means in Section 7.5. How to calculate an upper limit to 99.9% VaR is currently an open question. Despite a lack of clarity on a well-defined upper limit, we provide some guidance with the results in Section 7.

1.4 Structure of this paper

Section 2 gives an account of the main problems encountered when modeling conduct risk. Section 3 explains the data modification that the PM method uses to operate successfully. A literature review (Section 4) leads to a discussion of the PM method itself, which starts in Section 5 and continues in Section 6. There are many preliminaries that are components of the PM method, an account of which may be found in Section 5.2. The heavily statistical parts of that account are given in the online appendixes. In the results section (Section 7), we concentrate on the numerical results of the PM method when it is used with importance sampling rather than rejection sampling. The latter method is prohibitively slow, so we do not consider it in this analysis. The end result we seek is a single figure for 99.9% VaR for the data sets under consideration. These are presented in Section 7.3.1. Sensitivities to model parameters are discussed in Section 7.6. Finally, the advantages and disadvantages of the PM method used in the context of a VaR estimation for conduct risk are assessed in Section 8.

2 CONDUCT RISK MODELING

Modeling operational risk losses using severity and frequency distributions is problematic in some cases, and conduct risk losses are of particular concern. There is the general problem of which severity distribution to choose, given that several have passed a goodness-of-fit (GoF) test. The highest p-value for the calculated value of a GoF statistic may not always be the primary choice: other considerations could include the location of outliers or information-based statistics, such as the Akaike information criterion. More significantly for conduct risk, no distributions may pass the GoF test. In the latter case, a default distribution (often lognormal) must be used.

A further general problem concerns using a frequency distribution that is appropriate for rare event modeling, which may not be appropriate if events are not rare. For example, less than five events per year may be considered "rare", but not hundreds or thousands, as is the case with many operational risk data sets. Some conduct risk data sets we consider in this paper may be considered to satisfy this "rare" event criterion, but once additional (small) losses are included, "rare" degrades to "commonplace". Nevertheless, there are mitigating procedures. One is to use an empirical frequency distribution. Another is to use a normal frequency distribution. A Poisson distribution with parameter λ may be approximated by a normal distribution with mean and variance both equal to λ, provided that λ is sufficiently large. In practice, sufficiently large means greater than twenty, which applies for most operational risk data sets. We mention this general problem only because a Poisson frequency distribution is used extensively in operational risk VaR calculations, including those in this paper, even though a Poisson model might be a poor fit to data.

2.1 The "correct" value for VaR

We highlight a third problem when attempting to estimate operational risk VaR in a separate section, because it addresses the heart of the issue and is particularly important for conduct risk. However VaR is calculated, the resulting value may appear far too high or too low. The terms "too high" and "too low" are subjective, but a measure of them can be obtained indirectly. The regulator may decide that a bank's calculated capital is too low by comparing the calculated regulatory capital with that for other banks of similar size, and then demand additional reserves. Risk managers may consider the calculated capital to be too high because it is significantly more than a previously calculated value. Retention of excessive capital impairs the bank's ability to lend. Either way, we argue that the calculated VaR should be consistent with a bank's current risk profile. This is precisely the problem with the conduct risk losses considered for this paper: the VaR calculated in the same way as for other operational risk classes is judged to be "too high". The task is then to find a statistically sound

way to calculate a VaR for conduct risk losses that is consistent with a bank's current risk profile.

2.2 Provisions

So far, "loss" and "losses" have been used as general terms for the data under consideration. In the case of a large conduct risk loss, the term is more likely to be a provision: a fund reserved to cover future operational risk payments. Banks routinely provision anticipated losses, the practice being more prudent than having to react to actual losses in arrears. In recent years, banks have had to make extremely large provisions for payment protection insurance (PPI) compensation claims. Those provisions have been particularly high, and they are classified as a subcategory of conduct risk. The specific Basel II designation is "improper business or market practices" (see Basel Committee on Banking Supervision 2006, Annex 9, Paragraph 4.2). These large PPI provisions are exactly those referred to at the start of Section 1.3. They are extremely large and infrequent. Section 2.3 gives some examples. Following the establishment of a particular provision, payouts to individual customers will be made using that provision. In the case of PPI, the provision will likely accommodate thousands of such payouts, since individual payouts are rarely more than a few thousand pounds sterling.

The Basel Committee directives on data to be used for modeling severity distributions (Basel Committee on Banking Supervision 2011) clearly state that all losses resulting from the same operational risk event should be aggregated, given that they fall into the same risk category. These regulations also imply that any loss used for modeling should be listed on a company balance sheet. In this context, the provisions are the balance sheet items, not the individual payments to customers. Therefore, it is the provisions that should be modeled. In most cases, it is not possible to decompose any given provision into individual payments. Those payments are usually made against a current set of provisions, not against any particular member of the set. We refer to a provision's constituent payments as "missing data", because the number of them or their size is not necessarily known. The practice of aggregating multiple small losses by a single large provision is a particular characteristic of conduct risk losses. Other types of operational risk losses are provisioned, but CPBP provisions tend to be particularly large.

2.3 Examples of PPI provisions

In this section, we give an idea of the overall size of CPBP losses, in terms of both totals and individual provisions. Two publications from the UK Financial Conduct Authority (FCA) show how large the CPBP losses can be. Financial Conduct Authority (2018) reports that the mean redress paid was just under £2100 for each PPI

TABLE 1 Total payment protection insurance payouts, January 2011 to December 2018.

Year	PPI payouts
2011	2 137
2012	6 279
2013	5 220
2014	4 477
2015	4 476
2016	3 624
2017	3 360
2018	4 435
Total	34 008

Payout values are given in millions of pounds sterling. *Source:* Financial Conduct Authority (2019).

TABLE 2 Largest PPI provisions, 2012.

Bank	PPI provision
Lloyds	5 275
Barclays	2 176
RBS	1 735
HSBC	1 338
Santander	751
Bank of America (MBNA)	506
Others	1 219
Total	13 000

PPI provision values are given in millions of pounds sterling. *Source:* UK Parliament (2013).

complaint upheld in the second half of 2017, and just over £2500 in the first half of 2017. Table 1 shows yearly PPI refunds and compensations, sourced from Financial Conduct Authority (2019).

In addition to total PPI provisions, UK Parliament (2013) lists individual large PPI provisions for UK banks in 2012, shortly after PPI mis-selling was widely reported. The largest of these are shown in Table 2. By 2016, the total PPI provision had risen, and the *Guardian* published an article under the headline "Bill for PPI mis-selling scandal tops £40bn" (Treanor 2016). Note that the headline figure was somewhat exaggerated, especially as data for 2017 and 2018 was not available in 2016!

3 THE DISTRIBUTION OF CONDUCT RISK LOSSES

In the next subsection, we consider the effect of provisions on regulatory capital. The tail losses (especially the largest loss) have a significant effect on overall VaR. Provisions can form a major part of these tail losses, and we focus on the role of the largest loss as a provision. The effect of such low-frequency/high-value losses is to inflate VaR to an unrealistic level. Our aim is therefore to find a legitimate way to reduce the calculated VaR that is not wholly influenced by the largest loss. The general strategy is to partition the largest loss, the partitions being a proxy for the customer payments that it comprises. Two examples of "nonlegitimate" ways to reduce capital are ignoring large losses and partitioning large losses arbitrarily.

3.1 Partitions of the largest loss

Consider a set of N losses $S = \{x_1, x_2, \ldots, x_{N-1}, y\}$, where y is the largest loss (most likely a provision). Assume that all the x_i are drawn from a lognormal distribution. We select the lognormal distribution because it is a relatively simple fat-tailed distribution that is appropriate for many of the data sets we have considered. It is assumed that the largest loss is substantially larger than all the others. Theoretically, it is possible to decompose this provision into its constituent payments (the "missing data"), but in practice this is not practical (see the earlier discussion). Therefore, we aim to decompose the largest loss into n partitions, and we allow those partitions to represent the constituent payments. The value of n is treated as a parameter of the model developed in subsequent sections, and we require an estimate of it. Given such a decomposition of the largest loss into n partitions, we can then regard S as draws from a common lognormal distribution. Once the lognormal parameters have been determined, they can be used in a linear discriminant analysis (LDA) process (Frachot *et al* 2001) with a Poisson frequency distribution with parameter λ (the annual loss frequency) to derive VaR for a loss distribution.

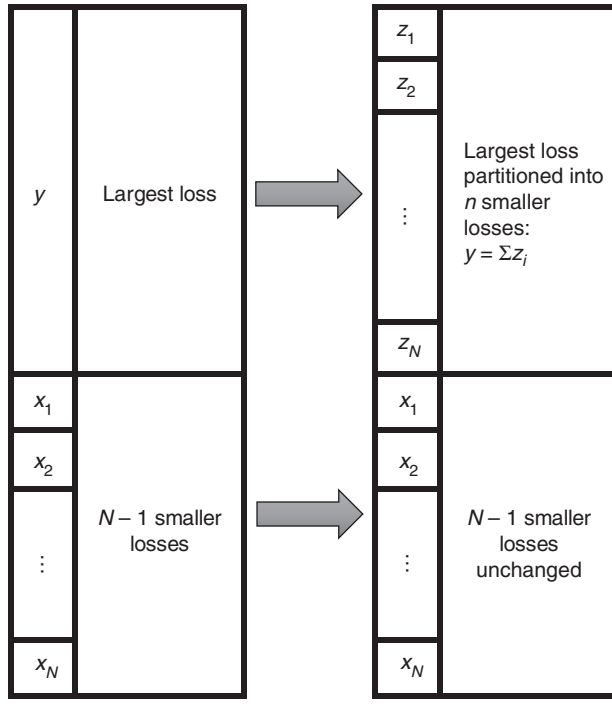
Let y be decomposed into n partitions $Z = \{z_1, z_2, \ldots, z_n\}$, each assumed to be lognormally distributed with parameters μ and σ^2:

$$\begin{aligned} y &= \sum_{i=1}^{n} z_i, \\ z_i &\sim \mathrm{LN}(\mu, \sigma^2), \quad i = 1, 2, \ldots, n, \\ n &\sim \mathrm{Po}(\lambda). \end{aligned} \tag{3.1}$$

The partitioning described above is illustrated in Figure 1.

In the LDA procedure, it is necessary to ensure that the sum of any draws from Z is y. We proceed using Bayes's theorem. The Bayes likelihood function can be written as the product of the densities of the x_i and of y. So, if the densities are

FIGURE 1 Partitioning of the largest loss y into n smaller losses z_i.



denoted by $f(x_i, \Theta)$, where $\Theta = \{\mu, \sigma, \lambda\}$ is a vector of parameters, $\boldsymbol{x}$ is a vector of the x_i and g is the density of y, then the likelihood function has the following form:

$$\mathcal{L}(\boldsymbol{x}; \Theta) = g(y \mid \Theta) \prod_{i=1}^{N-1} f(x_i \mid \Theta). \tag{3.2}$$

Evaluation of $g(y \mid \Theta)$ presents a problem, and it can be approximated by its expected value in the following way:

(1) generate n from a single draw from $\text{Po}(\lambda)$;

(2) generate the z_i from $\text{LN}(\mu, \sigma)$ subject to $y = \sum_{i=1}^{n} z_i$;

(3) calculate multiple instances of the likelihood function;

(4) calculate the mean likelihood.

3.2 The PM: general outline

We aim to fully define the probability distribution g in (3.2) so that we can take samples from it and calculate the 99.9% VaR. In advance of any calculations, we have to make use of what little we know about g. We can select what we think is a suitable distribution and choose parameters at random. With those parameters we test the likelihood given by (3.2). Progressively changing the parameter values enables us to see if the likelihood can be improved (ie, increased), and we aim to maximize the likelihood. In order to complete the likelihood calculations, we have to resort to Monte Carlo techniques (specifically MCMC) using the Metropolis–Hastings (MH) algorithm. The aim is to find a target stationary distribution for g by specifying a sequence of proposal distributions (with their parameters) and testing their likelihood. The problem lies in the MH calculation: it cannot always be done. The solution is to use the PM amendment of the MH algorithm. The PM method allows for an approximation to be made within the MH calculation. That approximation ensures that the calculation can proceed.[1]

4 LITERATURE REVIEW

This section contains references to and brief discussions of literature that is relevant to the theory of the PM method. Some alternatives are also included. Relevant terminology will be explained in detail from Section 5 onward.

4.1 PM: theory

A precursor to the development of the PM method appeared in Lin *et al* (2000) in the context of theoretical physics (specifically simulations of problems such as lattice quantum chromodynamics with dynamical fermions). In probability estimations using MCMC and the MH algorithm, it was necessary to ensure that probability estimates were not greater than 1 or less than 0. This was achieved by defining a two-stage loglikelihood ratio test, effectively a modification of the base MH loglikelihood ratio. This method indicated that advancements in accuracy and speed could be achieved by modifying the loglikelihood ratio. Further advances were achieved in the field of statistics, in addition to the 2003 breakthrough by Beaumont.

The PM method was first suggested in Beaumont (2003), although the term was not used at the time. The context was to investigate the genetic component of the threat to endangered species arising from low population size. The rate of species inbreeding is dependent on population size. The problem had previously been tackled

[1] The PM method is sometimes referred to in the literature by its full name: pseudo-marginal Markov chain Monte Carlo (PMMCMC).

by sampling gene frequency using serial samples taken from the same population and applying Bayesian analysis. Prior to 2003, the established method used for Bayesian analysis in which the form of a posterior distribution was not clear was the MCMC. In Beaumont's application, MCMC calculations proved to be too computationally intensive. As discussed in Section 5.1, we highlight the key amendment proposed in Beaumont (2003) to solve the computation problem. Finding proposal parameter candidates in step 2c of the MH algorithm that was used to implement the MCMC process proved to be intractable. Increasing sample sizes by grouped samples allowed for faster convergence to the target distribution; in Beaumont (2003), this "grouped sample" method was termed "grouped independence MH" (GIMH). In GIMH, sampling is from a distribution other than the (target) stationary distribution, and that is the key point of the PM algorithm.

PM is an adaptation of GIMH, and a precursor of it was suggested in Andrieu and Thoms (2008) in the context of adaptive MCMC methods. MCMC and the adaptive variations discussed in Andrieu and Thoms (2008) essentially rely on a homogeneity assumption for the proposal sequence that leads to the target stationary distribution. The idea that proposals should be replaced by approximations was intended to deal with nonhomogeneous cases.

A formal transition from GIMH to PM was provided by Andrieu and Roberts (2009), who coined the name by which the technique is now known. The first part of the name (pseudo) is due to the use of a "pseudo" distribution in place of a target stationary distribution. The second part of the name (marginal) is due to the calculation of a limiting marginal distribution using MCMC. A formal replacement of a proposal distribution by an estimator of it is described in Andrieu and Roberts (2009). The proof that such a substitution is valid under prescribed conditions is lengthy and complicated, so only a brief summary of the main points is given below.

- If a sequence of proposal distributions is irreducible and aperiodic, then the replacement of those distributions by estimators of them in the GIMH process can be proved to result in a convergent sequence (Andrieu and Roberts 2009, Theorem 1).

- The main result extends the above proof to provide a bound on the loss of efficiency of the approximating estimator compared with the ideal distribution. This loss can be made arbitrarily small by making the sequence of proposal distributions sufficiently long. The attached conditions are mild. First, the norm of the difference between the proposal distribution and its estimator must tend to zero as the sequence length increases. Second, weights associated with those estimators also tend to zero as the sequence length increases. The second point is important because convergence cannot be ensured if weights increase.

- Under more stringent assumptions, a GIMH sequence can be shown to converge uniformly.

Following the formal proof of the PM method, some extensions to and a discussion of it emerged.

Filippone and Girolami (2014) provide a general overview of the PM method in the context of predictive classification problems using Gaussian process Bayes priors. Comparisons are made with other classification methods: support vector machines (SVM) and the Gaussian process classifier (GPC). Filippone and Girolami (2014) include a discussion of sampling methods (including importance sampling) and consider a comprehensive analysis of the relative efficiency of classification algorithms. Using simulated data sets, they conclude that the PM achieved the best quantification of uncertainty compared with all other classifiers considered. SVM was the worst. Varying the number of importance samples showed that very few samples (no more than sixty-four) are needed to produce acceptable results.

The PM method may be viewed as a "noisy" version of the MCMC algorithm, because introducing an estimator of a stationary distribution results in some volatility of the parameter estimates. This point was addressed by Andrieu and Vihola (2015), who provided mild conditions under which the PM sequence of parameter estimates would converge. Further, they showed that the asymptotic variance of the PM algorithm converges to the asymptotic variance of the basic MCMC algorithm, provided that the accuracy of the estimators is increased. The impact of this result for our purposes is that we do not need to be particular over our choice of Bayes likelihood.

The issue of a "noisy" estimator was further discussed in Murray and Graham (2016). The MH algorithm incorporates an accept/reject step for the proposal distribution, or its estimator in the PM case. If the distribution of the estimator is heavy-tailed, the estimate often gets rejected and the sequence of parameter estimates can remain constant for long stretches. PM algorithms are often subject to this "sticking" behavior. In an improved PM algorithm, two parameter updates are used. The first uses a standard random number generator. The second uses random numbers that are evolved as part of a Markov chain on a joint auxiliary target distribution. The technique is known as "slicing". The result is that the parameter estimates advance provided that the functions in the loglikelihood part of MCMC are continuous almost everywhere.

It was observed that the size of the data sample used affects the efficiency of convergence of the PM sequence of proposal distributions. This problem was investigated by Doucet *et al* (2015), who showed that the number of samples should be chosen such that when the exact likelihood is efficient the variance of the loglikelihood approximation lies between 1 and 4. This result has a bearing on our choice

of sample size for a Bayes likelihood function (100 was used). See Section 6.1 for details.

Dahlin *et al* (2015) pointed out that the sequence of PM parameter estimates is often highly autocorrelated. They proposed a modification to the PM algorithm in which a Crank–Nicolson proposal function is used. The result was to introduce a positive correlation in the sequence, leading to improved computational cost.

Deligiannidis *et al* (2018) developed the correlated PM (CPM): a modification of the PM method that uses a likelihood ratio estimator calculated using two correlated likelihood estimators. It was found that the performance of the PM algorithm, as developed by Andrieu and Roberts (2009), was liable to degradation as the number of data points increased. The number of data points is not a problem for our study because the numbers used are relatively small. The CPM method uses a revised form of the ratio of estimators of the proposal parameters (see the algorithm detail in step 2c of Algorithm 5.1). The resulting ratio has a small relative variance, and therefore mimics the original MH better. Computation efficiency was found to be proportional to the number of data points. More significantly for our purposes, the variance of the loglikelihood ratio estimator should be between 1.0 and 2.0 in regions of high probability mass. Informally, Deligiannidis (personal communication) has informed us that a range of 1.0–3.0 is acceptable. Note that the result from Doucet *et al* (2015) is more generous because the conditions used for its derivation are less stringent. This point is mentioned in the discussion of the Bayes likelihood function in Section 6.1.

A variant of the CPM method (Tran *et al* 2017) is currently in preparation. Tran *et al* describe an approach known as "block PM", in which the random numbers that define samples are divided into blocks. The blocks are such that the likelihood estimates for the proposed and current values of the parameters to be estimated differ only by the random numbers in one block. The parameters for any proposal distribution are set to apply to one block only. The payoff is variance reduction, but at the expense of a more complicated loglikelihood ratio. See step 2c in Algorithm 5.1.

In recent studies by Quiroz *et al* (2018a,b), a PM framework was implemented, where at each iteration the loglikelihood from n observations was estimated from a random subset of those observations of size $m = O(\sqrt{n})$. Control variates are used to reduce the variability in the loglikelihood estimates, and a correlated PM scheme improved the acceptance probability in the MH algorithm. Two types of control variates are used: parameter expanded and data expanded. Parameter expansion uses a Taylor approximation of the loglikelihood function, and in data expansion, data are partitioned and the centroid of each partition is used in the loglikelihood function. The package of control variates with subsampling serves to reduce both the variance of the MCMC process and the computation time.

4.2 PM: applications

Few applications that are not associated with theoretical advances of the PM method exist, as the method is relatively new. The genetics application in Beaumont (2003) has already been discussed in Section 4.1, and that study should be seen as a major advance in the field of extensions of MCMC. A brief account of some others follows.

In Kronander *et al* (2015), light transport in heterogeneous media (smoke, clouds, fire, etc) was modeled by MCMC sampling of light paths. The aim was to synthesize visual renderings for films, games, etc. Basic MCMC did not yield a consistent estimator for a stationary path distribution. The PM method provided a means to better explore the path space, and to consequently reduce image noise.

The context in Peters *et al* (2017) is the allocation of the regulatory capital of insurance companies to business units. The task is similar to the problem tackled in this paper, as they both involve a partition of a total amount (in our case, it is a partition of the largest operational risk loss). The difference is that in Peters *et al* (2017) the number of partitions is known in advance, while in ours it is not. A further similarity is that the stationary distribution sought in Peters *et al* (2017) is lognormal, and the PM method is used to approximate it in the MCMC process.

4.3 Other numerical approximation methods

Some alternatives to the PM are discussed briefly here. They are mostly similar to the PM algorithm in that they modify an MCMC algorithm in some way. Of these, the particle MCMC (pMCMC) is probably the most viable alternative to PM, although it is more useful for higher-dimensional MCMC. It should be noted that PM directly addresses the issue of "missing data", the components of which have both unknown frequencies and unknown sizes.

4.3.1 pMCMC

pMCMC is an extension of particle filtering methods and is closely related to PM in that an unknown or intractable term is approximated by a known term. It was originally developed in the field of state-space time series analysis for filtering, smoothing and prediction. Although "particle" methods had been in use in some form or other since the 1950s, their statistical foundation was not formalized until 1996 (see Del Moral 1996). Significant advances to the technique were made by Andrieu *et al* (2010), who applied particle filtering to build efficient proposal distributions for MCMC. A less formal account of the underlying theory may be found in Gustavsson (2010).

The name "particle" in this context is a misnomer: "sample" would be more appropriate. Particle filtering proceeds by taking samples ("particles"), initially at random, from a proposal distribution. In the state space, at each time step, the state of each

sample is measured in order to calculate the corresponding space. Better fits are weighted more than less good fits. At the next time step, there is a resampling of particles with higher weights. As time progresses, the approximation should improve with increasing emphasis on the higher-weighted particles. Comments on the paper by Andrieu *et al* (2010) were made by Peters and Cornebise (2009), who noted the parallels with PM: an approximation to the proposal function can reduce the variance of the approximation.

The pMCMC method was extended in Peters *et al* (2013) by adding an adaptive stage. At each time step, an initial stage uses the pMCMC method to estimate a proposal function. In the second stage, variance reduction is achieved by using an optimal Kalman filter. Both stages use Rao–Blackwellizing: calculating marginal distributions in a way that uses the Rao–Blackwell formula. A similar technique was used by Doucet *et al* (2000) to improve sampling efficiency in high-dimensional spaces.

Septier and Peters (2015) describe the sequential MCMC (sMCMC) method, which is a further variant of particle filtering. Again, its purpose is to reduce variance (ie, "noise") due to weight degeneracy in high-dimensional systems. In sMCMC, the accept–reject step of the MH algorithm is implemented by an empirical approximation obtained from previous iterations of the algorithm. This recursive step allows sMCMC to sample more efficiently. Septier and Peters (2015) embed the sMCMC into a more general context that covers multiple MCMC variants.

4.3.2 Other miscellaneous methods

Further approximation methods exist for sums of lognormal random variables. In general, they provide improved accuracy at the expense of more extensive numerical computation and complexity. Two examples are Wu *et al* (2005) and Cobb *et al* (2012). Wu *et al* propose a method in which a Gauss–Hermite approximation is matched to the moment-generating function of the lognormal sum. In Cobb *et al* (2012), a lognormal distribution obtained using the Fenton–Wilkinson method is in turn approximated by a mixture of truncated exponentials. The parameters of the latter are polynomial functions of the lognormal scale parameter.

An earlier variant on the theme of replacement of an intractable MCMC term may be found in Peters and Sisson (2006). The context of this study is operational risk, and in particular how MCMC may be used with the Tukey "g-and-h" distribution, which has no closed form. As a result, the value of a proposal function is estimated by a metric that calculates the distance between a vector of summary statistics acting on both the simulated and observed data. The simulated data is used if the value of the metric is within an acceptable tolerance and is rejected otherwise. This method is also discussed in Marjoram *et al* (2004).

The "zigzag" algorithm, discussed in Bierkens *et al* (2016) and Bierkens and Roberts (2017) is a different type of MCMC amendment. It proceeds by replacing the exact gradient of the log of the Bayes posterior with an unbiased estimator, obtained by subsampling. Bierkens and his coworkers show that the stationary distribution resulting from this approximation tends to that of the Bayes posterior (Bierkens and Roberts 2017; Bierkens *et al* 2016). Further, with variance reduction using control variates of the unbiased estimate, very efficient scalable results can be obtained in a big data context.

5 THE PSEUDO-MARGINAL METHOD: BACKGROUND AND TECHNICALITIES

In this section, we summarize the elements of Markov chain and Bayesian methods with an emphasis on how they relate to the PM method as applied to conduct risk losses. Their main impact is a modification of the MH algorithm.

MCMC can be used as an alternative to calculating the expected value of a set of observations of a distribution, which is not applicable if such samples are dependent. So, for a random variable X with probability distribution $p(x)$, $x \in [a, b]$, let $\{\theta_1, \theta_2, \ldots, \theta_n\}$ be a sequence of parameters corresponding to random dependent observations $\{x_1, x_2, \ldots, x_n\}$ of X. The dependency is governed by the Markov chain transition matrix $\boldsymbol{T}$, which specifies that the probability that each observation θ_t is equal to some value s_t depends only on the previous observation θ_{t-1}. The goal of the MCMC process – and its means of propagation, the MH algorithm – is to determine a stationary distribution $p_T(\theta)$, defined as follows:

$$\boldsymbol{T}\, p_T(\theta) = p_T(\theta). \tag{5.1}$$

MCMC is used to estimate $p_T(\theta)$ numerically in cases where its functional form is not readily apparent. It is usual to reject the first part (the "burn-in") of the Markov sequence $\theta_1, \theta_2, \ldots$ due to instability compared with the latter part of the sequence. We typically used the first 30% of the sequence as the burn-in. The way in which samples are obtained is a significant problem in the implementation of our analysis. Rejection sampling is commonly used: instead of sampling from the stationary distribution, samples are drawn from a different "proposal" distribution $q(\theta)$. The sample from $q(\theta)$ is accepted with probability $p_T(\theta)/Wq(\theta)$, where W must satisfy $p_T(\theta) \leqslant Wq(\theta)$ for all values of θ. It is not unusual for most samples to be rejected, implying that the sampling process is very lengthy. This was the case in our implementation, and we used importance sampling instead. With importance sampling (see, for example, Liu 2004), a more accurate approximation can be obtained by sampling from an alternative distribution. Values of θ that have a more significant effect on maximum likelihood calculations can be obtained.

5.1 The MH algorithm

The MH algorithm is crucial in determining the stationary distribution in (5.1) in cases where no analytical form for it is apparent. Good explanations may be found in Chib and Greenberg (1995) and Robert (2016). Given the importance of the MH algorithm for the PM method, we give a summary of the MH algorithm here.

Starting with an initial parameter value θ_1, the MH algorithm generates a sequence $\theta_1, \theta_2, \theta_3, \ldots$, from which an approximation of the stationary distribution $p_T(\theta)$ may be made. In each of M cycles, a new value of θ, θ^*, is proposed, having been drawn from a proposal distribution $q(\theta^* \mid \theta)$. That value is either accepted or rejected according to a random draw from a $U[0, 1]$ distribution. If it is rejected, the current value of θ is retained. The MH algorithm is summarized in Algorithm 5.1.

ALGORITHM 5.1 (MH algorithm)

(1) Initialize an arbitrary initial value $\theta_1 > 0$.

(2) For $t = 2, 3, \ldots, M$, repeat the following steps for each value θ_t:

   (a) propose a new candidate θ^*, drawn from the proposal distribution $q(\theta^* \mid \theta)$;

   (b) draw a single random number u from $U[0, 1]$;

   (c) calculate the acceptance ratio

$$r = \min\left(1, \frac{p_T(\theta^*)q(\theta_t \mid \theta^*)}{p_T(\theta_t)q(\theta^* \mid \theta_t)}\right);$$

   (d) if $u < r$, $\theta_{t+1} = \theta^*$; otherwise, $\theta_{t+1} = \theta_t$.

(3) Discard an initial burn-in percentage (often 20–30%) of the θ_t. The remaining values mirror the target stationary distribution $p_T(\theta)$.

5.2 From MH to PM

See step 2c. The only difference between the MH and PM algorithms is in the ratio of the p_T terms in step 2c of Algorithm 5.1. In that step, p_T is replaced by a non-negative unbiased estimator of p_T, denoted by $\hat{p}_T$. Andrieu and Roberts proved that is it sufficient to use an unbiased estimator of the target distribution in the MCMC process for convergence to the target distribution. Therefore, the PM algorithm is identical to the MH algorithm, but with step 2c replaced by the steps in (5.2). The latter equation needs the calculated estimators $\hat{p}_T(\theta^*)$ and $\hat{p}_T(\theta_t)$ of $p_T(\theta^*)$ and $p_T(\theta_t)$, respectively.

ALGORITHM 5.2 (PM algorithm: PM, step 2c)

$$r = \min\left(1, \frac{\hat{p}_T(\theta^*)q(\theta_t \mid \theta^*)}{\hat{p}_T(\theta_t)q(\theta^* \mid \theta_t)}\right). \tag{5.2}$$

The replacement of $\hat{p}_T(\theta_t)$ by $\hat{p}_T(\theta^*)$ and $p_T(\theta_t)$ by $p_T(\theta^*)$ looks like a small change, but it is a very significant one. With suitable estimators, it allows a tractable calculation in the MH algorithm. We can therefore give a one-sentence summary of PM as follows:

> PM = MH with sampling from an unbiased estimator of the MCMC stationary distribution.

The rigorous proof in Andrieu and Roberts (2009) that the stationary distribution admitted by the Markov chain of the PM algorithm (using the unbiased estimator $\hat{p}_T$) is the same as the stationary distribution admitted by the Markov chain of the MH algorithm (using p_T) is complicated. Consequently, a simplified outline of the proof is given in Appendix A online. This proof is based on the outline of Picchini (2018), with notation amended to correspond to the notation used in this paper. Another outline proof may be found in Zheng (2016).

The key point in the proof is the substitution of the estimator $\hat{p}_T(\cdot)$ for the stationary distribution $p_T(\cdot)$ admitted by the Markov chain, which has a transition matrix $\boldsymbol{T}$ such that the exact relationship $\boldsymbol{T}\,p_T = p_T$ applies. A rigorous justification for this step is given in Andrieu and Roberts (2009, Theorem 1). There are underlying assumptions that the Markov chain associated with $\boldsymbol{T}$ must be irreducible and aperiodic. The next subsection gives a brief overview of the principal features of the online appendixes.

5.2.1 PM: key points

Appendix A online sets out the basis of the approximation of a general distribution $q(\theta)$ by the mean of random samples drawn from the conditional distribution $q(\theta \mid s)$. This conditional distribution is a marginal of a joint distribution of θ and a "disturbance" parameter, s. Denoting the mean value of the samples by $\hat{q}(\theta)$, the final result in Appendix A online is that the expected value of $\hat{q}(\theta)$ is $q(\theta)$:

$$\mathbb{E}_s(\hat{q}(\theta)) = q(\theta). \tag{5.3}$$

The result in (5.3) is extended to a Bayesian context in Appendix B online. The underlying theory is given in Section 5.2.2.

As far as operational risk VaR is concerned, the aim of using the PM approximation is to reduce the calculated VaR to a level consistent with an established risk profile, but only when it is needed. The conditions under which it should be used

were discussed in Section 1.3. The key points are the presence of one or more very large outlier losses, leading to a calculated VaR that is considered (by risk managers) to be inconsistent with the risk profile of a financial institution. Further comments on this point may be found in Sections 7.2.1 and 7.8, where the results are presented.

5.2.2 PM: key points in a Bayesian context

When the stationary distribution p_T is used in a Bayesian context, the likelihood is denoted by $p_T(x \mid \theta)$, and the prior by $f(\theta)$. In both cases, θ is a parameter vector. The posterior is then given by $p_T(\theta \mid x)$ in (5.4), which incorporates the "evidence", $p_T(x)$:

$$p_T(\theta \mid x) = \frac{p_T(x \mid \theta) f(\theta)}{p_T(x)},$$
$$p_T(x) = \int p_T(x \mid \theta) f(\theta)\, \mathrm{d}\theta. \tag{5.4}$$

Appendix B online gives an account of how the PM method is applied when the target distribution is a Bayesian posterior.

Extending the ideas in Section 5.2.1, the general distribution $q(\theta)$ is replaced by a Bayesian posterior distribution. The starting point is the relationship between a Bayesian posterior $q(\theta^*|\theta)$, a prior $q(\theta^*)$ and a likelihood expressed in terms of the Markov chain stationary distribution $p_T(\theta \mid \theta^*)$. That relationship also has a normalization factor $p_T(\theta)$:

$$q(\theta^*|\theta) = \frac{p_T(\theta \mid \theta^*) q(\theta^*)}{p_T(\theta)}. \tag{5.5}$$

The next stage is threefold, and it is necessarily included because it is assumed that the distributions $q(\cdot)$ and $p_T(\cdot)$ are intractable. They are replaced by estimators, which are approximations but can be computed. First, the augmentation introduced in Appendix A online is applied to functions q and p_T. Second, the functions themselves are replaced by estimators, denoted by "hat" symbols. Lastly, the joint density of θ^* and s is given as a product of their respective densities, $q(\theta^*)$ and $\bar{p}(s)$:

$$\hat{q}(\theta^*, s \mid \theta) = \frac{\hat{p}_T(\theta \mid s, \theta^*) q(\theta^*) \bar{p}(s)}{\hat{p}_T(\theta)}. \tag{5.6}$$

The argument then proceeds to show that the expected value of the estimator $\hat{p}_T(\theta)$ is the exact function $p_T(\theta)$:

$$\mathbb{E}_s(\hat{p}_T(\theta)) = p_T(\theta). \tag{5.7}$$

The interchangeability of a function with its estimator is the key point of the PM method.

Once $\hat{p}_T(\theta)$ has been replaced by $p_T(\theta)$, and the exact Bayesian relationship

$$\frac{q(\theta^*|\theta)}{q(\theta^*)} = \frac{p_T(\theta \mid \theta^*)}{p_T(\theta)}$$

has been applied, (5.6) reads

$$\hat{q}(\theta^*, s \mid \theta) = \frac{\hat{p}_T(\theta \mid s, \theta^*) q(\theta^* \mid \theta) \bar{p}(s)}{p_T(\theta \mid \theta^*)}. \tag{5.8}$$

The final stage is to show that the target posterior distribution can be obtained by calculating the marginal distribution of $\hat{q}(\theta^*, s \mid \theta)$, ie,

$$\int \hat{q}(\theta^*, s \mid \theta)\, \mathrm{d}s = q(\theta^* \mid \theta). \tag{5.9}$$

All the terms on the right-hand side of (5.8) can be calculated once they are replaced by their estimators. Then, the integral in (5.9) can be approximated by Monte Carlo sampling.

5.2.3 Determination of VaR

It should be remembered that the point of using the PM method is to determine a single VaR value for any given data set. Further, that value should be consistent with the risk profile expressed by the data: neither too low nor too high. The task is then to determine that single value by selecting an appropriate number of partitions for the largest loss.

Before carrying out any numerical trials, it was not clear how many partitions (the number n in Section 3.1) to use. Experience shows that a reduction in VaR results if the largest loss (y in Section 3.1) is partitioned, but it is not possible to tell in advance how large that reduction will be. There is an additional problem that if too many partitions are used, the increased frequency will lead to an increase in VaR. Some values for the number of partitions make sense intuitively. For example, given data spanning Y years, it makes sense to subdivide the largest loss into Y partitions (not necessarily equal). It also makes sense to subdivide them into $2Y$ partitions, each of which nominally represents a six-month period. Continuing in this way, $12Y$ partitions represent a nominal split by month. Further partitioning is clearly possible but may not correspond to any well-defined time period.

A general strategy is to repeat the complete PM calculation for a varying number of partitions and to defer a decision until all results have been reviewed. A possible outcome is that a limiting value for VaR exists as the number of partitions increases. It is not clear in advance of calculations whether or not such a limit exists. The results in Section 7.3 are not totally clear on this point, but once obtained, a strategy to determine VaR can be advanced. Details are reported in Section 7.3.

6 APPLICATION OF THE PSEUDO-MARGINAL METHOD TO CONDUCT RISK LOSSES

We seek to partition the largest loss into a series of smaller losses, not necessarily all the same size. The smaller losses replace the largest loss, and lognormal parameters for the result are determined using the PM method outlined above. A much weaker alternative is to simply replace the largest loss by a "sensible" partitioned form of it, but that approach suffers from certain deficiencies. First, the way the partitioning should be done is arbitrary. Second, simple partitioning is not justified by the Basel regulations (Basel Committee on Banking Supervision 2011) unless there are particular partitions that could be associated with provisions for particular conduct risk events. Third, the results (Section 7) show that an arbitrarily small VaR value can be achieved by using a sufficient number of partitions.

6.1 Dirichlet likelihood

The Dirichlet likelihood function models the precise situation described in Section 3.1: the largest loss is partitioned, but neither the number of partitions nor the amount allocated to each partition is known in advance. The likelihood function represents sums of draws from a lognormal distribution derived by partitioning the largest loss. We regard the largest loss (treated as a provision) as containing a number of "unknown" smaller losses, which are the payments covered by the provision. The partitions of the largest loss act as proxies for these payments. We cannot use payments directly, so we simulate them in a statistically rigorous way. A Monte Carlo technique based on the Dirichlet distribution is used to draw samples for this Monte Carlo process. Using a Dirichlet distribution for the Bayesian likelihood means that no assumptions are made about the number of elements in each partition. In general, (6.1) gives the n-fold Dirichlet density $f(x,\alpha)$ in which $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is a parameter vector and the support of the random variables $x = \{x_1, x_2, \ldots, x_n\}$ is $(0,1)$, with $\sum_{i=1}^{n} x_i = n$:

$$f(x,\alpha) = \frac{\Gamma(\sum_{i=1}^{N} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} x_i^{\alpha_i}. \tag{6.1}$$

In (6.1), the variable n is the number of partitions of the largest loss, defined in (3.1). For simplicity, we use $\alpha_i = 1$, $i = 1, \ldots, n$. In this case, (6.1) reduces to a Gamma function:

$$f(x,\alpha) = \Gamma(n) \quad \text{for all } x_i \in x. \tag{6.2}$$

The Dirichlet distribution is used to estimate a value for the likelihood estimator in (3.2). Dirichlet samples of size 100 were sufficient to achieve a stable result without

compromising the time taken to do the whole calculation. In the PM algorithm, there is a trade-off between convergence of the lognormal parameters μ and σ and the computing time. Using a large sample size will typically result in MH averages with low asymptotic variance, but the computing time required to construct the likelihood estimator $g(y \mid \Theta)$ increases. A sample size of 100 is consistent with the variance conditions specified in Doucet *et al* (2015) and Deligiannidis *et al* (2018).

6.2 Gamma priors

We now assign prior distributions for the parameters of the assumed lognormal distributions for the proxy losses z from Section 6.1. There are two lognormal parameters to consider: μ and σ. There is much uncertainty about the distribution of μ, and an uninformative prior in the form of a Gamma distribution with a large expected value is appropriate. There is more certainty about σ. VaR calculated from a lognormal distribution is very sensitive to the value of σ, and if this is too high, VaR is likely extremely large. Therefore, we set the prior on σ such that its expected value is unlikely to result in a very large VaR. So, a maximum value of $\sigma = 3$ is preferable. A Gamma distribution $\Gamma(a, b)$ has expected value a/b. Therefore, the following Gamma distributions are appropriate for the priors:

$$\mu \sim \Gamma\left(\frac{1}{1000}, \frac{1}{1000}\right), \qquad \sigma \sim \Gamma\left(\frac{1}{3}, \frac{1}{3}\right). \tag{6.3}$$

Both distributions in (6.3) have an expected value of 1, but the range of values generated in a random sample for the μ prior is much greater than that generated in a random sample for the σ prior. This is consistent with the uninformative choice of prior for μ and the informative choice of prior for σ. The probability that $\sigma > 3$ is approximately 0.1, which is small enough for the purpose of finding proposal parameter values in the PM algorithm.

6.3 Posterior

The required stationary distribution p_T is the limiting posterior distribution obtained from a sequence of parameter vectors $\{\theta_1, \theta_2, \ldots, \theta_n\}$. The result is very sensitive to the value of the initial parameter, θ_1. Therefore, care has to be taken to choose suitable initial values, which are reported in Section 7.

In the MCMC process, the parameters of p_T, μ and σ are updated according to the following schema:

$$\left.\begin{aligned} \mu_{t+1} &= \mu_t + \varepsilon, \\ \sigma_{t+1} &= \sigma_t + \varepsilon, \\ \varepsilon &\sim N(0, \eta^2). \end{aligned}\right\} \tag{6.4}$$

The subscripts refer to successive steps in the Markov chain. Both μ and σ are perturbed by a stochastic term ε, which is normally distributed with mean zero and constant variance η^2. In practice, $\eta = \frac{1}{10}$ was used.

6.4 The Fenton–Wilkinson approximation

Equation (3.1) assumes a partitioning of the largest loss, y, into n sublosses, each having a $\mathrm{LN}(\mu, \sigma^2)$ distribution, where n is a Poisson random variable. Therefore, any random sample drawn from the set of sublosses is a sum of lognormally distributed random variables. Informal trials show that the sum of a large number of lognormally distributed random variables also appears to be lognormally distributed, but only under certain conditions, as detailed below. The convergence of the sum of a large number of lognormal random variables to a single lognormal random variable is formalized in the Fenton–Wilkinson (Fenton 1960) approximation (see (6.5)). This approximation is considered here simply because a sum of lognormal random variables occurs naturally in the foregoing analyses, and it therefore appears to be a viable alternative to the PM method.

If $y = \sum_{i=1}^{n} z_i$, where $z_i \sim \mathrm{LN}(\mu, \sigma^2)$, then $y \sim \mathrm{LN}(\mu_y, \sigma_y^2)$, where

$$\left.\begin{aligned} \sigma_y^2 &= \log\left(\frac{\mathrm{e}^{\sigma^2} - 1}{n+1}\right), \\ \mu_y &= \log(n\mathrm{e}^{\mu} + \tfrac{1}{2}(\sigma^2 - \sigma_y^2)). \end{aligned}\right\} \tag{6.5}$$

The equations in (6.5) are based on the approximate relationships $\mathbb{E}(y) = n\mathbb{E}(z_i)$ and $\mathrm{VaR}(y) = n\mathrm{VaR}(z_i)$ and are derived by matching moments. They are used in conjunction with an MH MCMC algorithm to produce a sequence of successive values for the parameters of a target distribution. It was hoped, but not guaranteed, that such a sequence would converge. The Fenton–Wilkinson approximation is known to be a good approximation to some distribution tails, but not, in general, to distribution bodies. The approximation can be improved by matching the second and third moments, or even the third and fourth moments, if upper tail or extreme upper tail probabilities are the main interest. Details of these and similar amendments may be found in Rook and Kerman (2015).

7 RESULTS

In this section, we report the results of using the PM method to decompose the largest loss into proxy payments, and we attempt to settle on a VaR value for the data considered. The results are compared with a much simpler method in which the largest loss is partitioned uniformly. VaR is determined by fitting a distribution to the result

of that partitioning, then using the LDA procedure. In addition, we report on the stability and sensitivity of the parameter values obtained.

7.1 Data and implementation

The customizations described in Section 6 are applied to six CPBP data sets. The first four are internal and are labeled CPBP29, CPBP9, CPBP1879 and CPBP2032. The numerical part of the names gives the number of losses in each case. The date range for CPBP29 and CPBP9 is January 2011 to December 2015 inclusive, and the date range for CPBP1879 is January 2013 to December 2017 inclusive. CPBP2032 has a 6.5-year data window of January 2012 to June 2018. The fifth data set, CPBP-ORX, is external, and originates from the Operational Risk data eXchange (ORX).[2] The data set CPBP-ORX is significantly larger than the internal data sets. It has nearly 7600 losses, the largest of which is £6553 million, which is 40.5% of the total loss in this data set. The effect of a large number of smaller losses is also significant. CPBP9 is an extreme example of conduct risk losses. With only nine losses, distribution fitting is unreliable. The largest loss is £1220 million, which is 75% of the total loss. The data sets CPBP29, CPBP9, CPBP1879, CPBP2032 and CPBP-ORX are representative of data that we have seen in the past few years and are consistent with the characteristics of the conduct risk losses discussed in Section 1.3. They contain large losses, but CPBP1879 and CPBP2032 have about 1800 "smaller" losses (ie, less than £10 000), totaling approximately £5 million. That total is small compared with the mean annual loss, but the increased annual frequency has a significant effect on the numerical results reported in Section 7.3.

The data set CPBP-RAN is a synthetic data set, generated as a random sample of size 500 from a LN$(12, 2)$ distribution. In addition, it has one large loss of £1.1 billion, which is approximately twice the sum of all the other 500 losses, making a total of £1.628 billion for all 501 losses. CPBP-RAN nominally spans the five years 2014–18. Although CPBP-RAN is labeled as "conduct risk", it resembles other risk classes we have observed in that there is one clear outlier.

All the results reported were calculated using the R statistical language, run on a Microsoft Windows computer with an i7 processor and 16 GB of RAM. Timings are given in the relevant sections.

7.2 Results without partitions

As a preliminary to reporting the results using the PM method, we first give the 99.9% VaR for the data in Section 7.1 with the unmodified data (ie, with no partitions). These are the "base" results that are obtained by using the LDA procedure

[2] See https://managingrisktogether.orx.org/.

TABLE 3 CPBP data: 99.9% VaR, no partitions.

Data set	Base VaR	VaR without largest loss	Largest loss (%)	TNA	Empirical bootstrap VaR
CPBP29	8 317	4 776	35.4	0.045	2 165
CPBP9	65 424	10 727	75.0	0.065	3 649
CPBP1879	8.7	8.1	36.3	0.066	2 293
CPBP2032	5.1	4.8	35.6	0.064	2 073
CPBP-ORX	313	213	40.5	0.053	20 597
CPBP-RAN	747	632	67.2	0.008	3 363

TNA, transformed normal GoF measure for the lognormal severity fit. All VaR values are in millions of pounds sterling.

(Frachot *et al* 2001), having fitted a lognormal distribution. In the brief discussion after Table 3, we indicate why these "base" results are unsatisfactory. The table also shows the VaR value if the largest loss is removed so that we may see its effect. The "Largest loss (%)" column shows the largest loss as a percentage of the sum of all losses.

Table 3 also shows the 99.9% VaR results obtained using the empirical bootstrap. These are indicators of minimum VaR since, in a random sample, it is impossible to draw a sample member that is greater than the largest empirical loss. The TNA column gives the "transformed normal" GoF measure for the lognormal severity fit. This measure was specifically developed for the context of operational risk losses and is discussed in detail in Mitic (2015). Note that, since distribution parameters have to be estimated from data, established GoF tests (such as Kolmogorov–Smirnov) are not strictly applicable. The critical two-tail 95% significance value of the TNA statistic is 0.068. Any measured TNA value less than that can be regarded as an indicator of a suitable data fit at 5% significance. All VaR values are in millions of pounds sterling.

We consider the "base" VaR for CPBP29 too high and unduly affected by the largest loss. The largest loss is due to PPI compensations, which are nonrecurrent. However, removing it is unsafe, as it is possible that it could be replaced in future years by an equally large loss due to something else. We have found that, over the past six years, the "base" result is considerably greater than the sum of all other risk classes, and it is consequently not credible.

The "base" VaRs for CPBP1879 and CPBP2032 are too low, and are due to the effect of higher-frequency, low-value losses. They do not reflect the presence of the largest losses, which are of the same order as that for CPBP29. The huge disparities between the empirical bootstrap values and the (lognormal) fitted values are a startling illustration of the way the largest loss affects VaR. The disparities are due to the difference between pointwise sampling for the empirical bootstrap and

localized sampling in the case of the fitted distribution. In the former case there is an inflated probability of drawing several large losses in a random sample based on the mean annual loss frequency, since each loss has the same probability of being drawn. Drawing a very large loss is unlikely in the latter case.

CPBP-ORX appears to be an intermediate case that has an exceptionally large loss, with a long tail of smaller losses to mitigate the largest losses. Prior to doing the calculations, it was thought that the largest loss of CPBP-RAN would have a dramatic effect on VaR. After calculations, it was seen that the effect was there, but it was not as dramatic as originally thought.

Although CPBP-RAN passes the TNA GoF test for a lognormal distribution, it passes the same GoF for a generalized Pareto distribution, but with a near-to-zero TNA test statistic of 0.0076. The implication is only that CPBP-RAN could have been generated as a random sample from a generalized Pareto distribution.

The "base" VaR results for CPBP9 show in a dramatic way that a Monte Carlo-based "loss distribution" modeling approach is inappropriate. It is clearly incorrect, and is comparable with the gross domestic product (GDP) of many European countries. For example, the GDP of Spain in 2017 was US$38 103 million.[3]

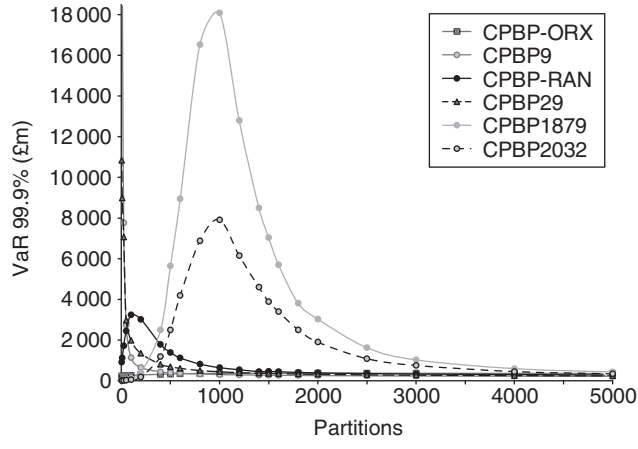
7.2.1 Inappropriate application of the PM method

The PM method is intended for use where the largest loss (or losses) are much larger than the others. These are typified by the data sets in Table 3. They are all extreme cases, and the huge VaR values that emerge from alternative calculations clearly indicate that an alternative is required. The PM method is not intended for use when it is not needed. We consider that the distinction between when it is and when it is not needed should be a subjective decision in the same way that a choice of modeling distribution or threshold is. Therefore, at this stage, we do not wish to be overprescriptive and define rigid criteria for judging when the PM method should be used. Ultimately, the choice reduces to whether or not the capital calculated using other means is too large to be contemplated by the organization concerned. Both risk managers and risk modelers have a good feeling for when this happens. A more objective alternative would be to compare the calculated capital of an organization with that of similar organizations. That, too, is subjective, because all organizations are different.

Having said that, it is instructive to consider what might happen if the PM method is applied to a data set when it need not be. Therefore, we applied it to the data set CPBP-RAN after removing the largest loss. The resulting data set is termed CPBP-RAN-LN. Without that largest loss, CPBP-RAN-LN reduces to the original random sample from a lognormal distribution that formed the basis of CPBP-RAN. Since

[3] https://data.oecd.org/spain.htm.

FIGURE 2 PM 99.9% VaR results.



For CPBP-ORX, the actual VaR values (in millions of pounds) have been divided by ten for clarity.

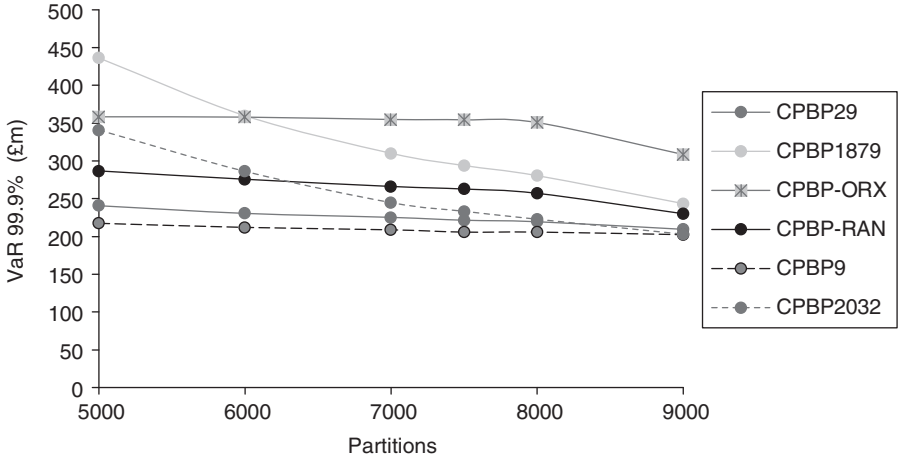
CPBP-RAN-LN is the result of a random sample from a lognormal distribution, a simple lognormal fit suffices for the "correct" VaR calculation. The results for the "correct" calculation and its PM equivalent are reported in Section 7.8.

7.3 PM with importance sampling

The PM method using importance sampling was applied to the data sets in Section 7.1 using partition sizes ranging from 5 to 8000. The variation of VaR with the number of partitions is illustrated in Figure 2. In each case, 25 000 Monte Carlo iterations were used. The sensitivity of VaR to the number of Monte Carlo iterations is discussed in Section 7.6. The underlying data for these figures is shown in Appendix C.1 online. Note that, in order to fit all cases in the same figure, the profile for CPBP-ORX shows the actual values (in millions of pounds sterling) divided by ten (ie, the actual VaR values are the displayed values multiplied by ten). That way, the shape of the CPBP-ORX profile can be seen in the context of all the others.

The shapes of all but one of the profiles in Figure 2 are essentially the same: an initial increase in VaR followed by a relaxation to what appears to be a limiting value. The initial rise for CPBP29 is not apparent in Figure 2. It is only expressed in the "base" case (ie, with no partitioning), when VaR $= 8317$ (as in Table 3). The profile for CPBP9 is similar to the profile for CPBP29 but starts its relaxation stage (at five partitions) at a much higher VaR value. The initial points are not shown since they

FIGURE 3 PM 99.9% VaR results: 5000–9000 partitions.



For CPBP-ORX, the actual VaR values (in millions of pounds) have been divided by ten for clarity.

would mask the shape of the others. They are (5, 241 289) and (10, 61 259). CPBP29 and CPBP9 are almost coincident.

Figure 3 is an extension of Figure 2 for partitions in the range 5000–9000. It shows the distinction between the profiles of the data sets more clearly. Note that for CPBP-ORX the actual VaR values (in millions of pounds) have been divided by ten for clarity (as for Figure 2).

We note the following points.

- Inconsistent results appeared for 8000 or more, but less than 10 000, partitions. The calculated VaR could be either significantly larger or significantly smaller than the VaR values obtained for partitions in the range 5000–7500. Further, very different VaR values could be obtained for repeated runs using either 9000 or 10 000 partitions. We attribute this to the nonconvergence of the MCMC process, even for up to 100 000 Monte Carlo iterations.

- Floating point operations failed for more than 10 000 partitions. The Dirichlet likelihood part of the PM calculation has an exponential term that cannot be evaluated within the precision limits of the R interpreter (see Appendix D online).

- The limiting value referred to above is not clear cut. It is discussed below in Section 7.3.1.

TABLE 4 99.9% limiting VaR.

Data set	Limiting VaR
CPBP29	409
CPBP1879	436
CPBP2032	286
CPBP9	375
CPBP-ORX	270
CPBP-RAN	313

Limiting VaR values are given in millions of pounds sterling.

The PMMCMC process can be lengthy. For 25 000 iterations it took fifteen minutes for the smallest number of partitions and the smallest data sets, rising linearly to approximately six hours for the largest number of partitions with the largest data sets. The LDA part of the PM calculation takes one to two minutes, depending on the annual frequency. As such, it is a very small component of the total time taken for the most complete calculations.

7.3.1 PM with importance sampling: VaR estimation

The discussion in Section 5.2.3 refers to a procedure for estimating VaR that was to be dependent on the results derived in Section 7.3. The profiles in Figure 2 appear to be trending toward a limit (but not necessarily the same limit for all cases). In fact, a small downward drift is apparent between 5000 and 8000 partitions (see Figure 3).

Nevertheless, we assume that a limiting VaR value exists and model the relaxation (downward-sloping) portions of Figure 2 by an exponential distribution of the form

$$v = v_m + (L - v_m)(1 - \mathrm{e}^{\lambda(p - p_m)}). \tag{7.1}$$

This model depends on identifying the number of partitions that corresponds to the maximum VaR (ie, the local maximum of each profile in Figure 2). In (7.1), p and v denote the number of partitions and VaR, respectively; the local maximum is denoted by the point (p_m, v_m), the limit point is denoted by L and the exponential relaxation factor is denoted by λ.

Table 4 shows the results of the above procedure. The limiting VaR values are correct to the nearest million pounds and were calculated using the VaR values obtained for 8000 partitions or less. For more than 8000 partitions, we noted instability in the results of the PM VaR calculation. We discuss this point further below.

Note that the limiting VaR for CPBP-ORX is much larger than the other limiting VaR values. This reflects the larger size of the data set, and the extremely large losses

within it. Care should be taken when reviewing the results in Table 6 in Appendix C.1 online. A limiting value has been assumed in order to be able to derive a single VaR value per data set. It is possible that if many more partitions are used, the calculated VaR values will fall far below the limiting values indicated. That calculation is not possible since floating point operations become intractable for more than 10 000 partitions, and the Markov chains become unstable at that level. Possibly, then, any further decline in the calculated VaR in the region of 8000–10 000 partitions is due partly to that instability.

Some suggestions for alternative strategies for determining the required single VaR values are indicated below:

- find the VaR corresponding to the number of partitions for which the magnitude of the gradient of a (p, v) profile becomes sufficiently small (say 1%);
- calculate the mean value of VaRs for a high-value range of partitions (5000–8000).

These tend to produce lower values than those in Table 6 in Appendix C.1 online, which is less acceptable from a prudential viewpoint.

The R code to support the PM method and the VaR calculation is given in Appendix C.1 online.

7.3.2 Combination of internal and external data

The VaR values in Table 4 are listed separately, but in practice they are usually used in "internal plus external" data combinations. In this section, we therefore present the results of such a combination in which the limiting VaR calculated using the external data set, CPBP-ORX, is combined with each of the other internal data sets in turn. A common method of doing this is to use a linear combination in which a weight, w, is calculated using the Bühlmann–Straub credibility method. Consequently, the weight is known as a credibility weight. The Bühlmann–Straub method is described in Bühlmann and Gisler (2005). It is reiterated in a simpler form in Mitic and Bloxham (2018), where the two methods of calculating variance (one is the variance of all the data, the other is the variance of means of segments of the data) are clarified. So, if V_{int} and V_{ext} are the VaRs for an internal and external data set, respectively, the combined VaR, V, is given by

$$V = wV_{\text{int}} + (1 - w)V_{\text{ext}}. \tag{7.2}$$

This linear combination is a very quick way to combine two VaR values, given that they have already been calculated separately. The credibility weight is more

TABLE 5 Internal–external 99.9% limiting VaR combination.

Internal data set	w	V
CPBP29	0.9979	409
CPBP1879	0.9219	423
CPBP2032	0.9391	285
CPBP9	0.9980	375
CPBP-RAN	0.9653	312

V values are given in millions of pounds sterling.

often used to weight internal and external data within a maximum likelihood calculation. A combined "internal plus external" distribution is sought, and its parameters are estimated by a sequence of likelihood calculations. At the ith such calculation, distinct likelihood values, $L_{i,\text{int}}$ and $L_{i,\text{ext}}$, respectively, are calculated using the internal and the external data. The quantity $L_i = wL_{i,\text{int}} + (1-w)L_{i,\text{ext}}$ is the likelihood value returned. The maximum likelihood is then $\max(L_i)$, and the corresponding "internal plus external" distribution parameters are recorded. VaR is calculated using those parameters by any appropriate means. We usually find that the results obtained using the approximation in (7.2) are close to those obtained using maximum likelihood. Table 5 shows the VaR results when CPBP-ORX is combined with each of the other data sets using the numerical results in Table 4.

The values of w shown in Table 5 are all near 1, indicating, as is often the case, that there is a significant bias in favor of internal data. The purpose of the external data is to augment the internal data with higher-value losses that have not been realized internally, but could be in the future. Therefore, the combined "internal plus external" VaRs should be much closer to the internal values in Table 4. The date ranges for some internal/external combinations in Table 5 are not strictly comparable, but nevertheless, the Bühlmann–Straub method is sufficiently robust. Usually, external losses are far in excess of internal losses, and VaRs obtained for internal/external combinations are consequently significantly greater than those using internal data only. Table 5 shows that adding external data increases VaR in all cases except CPBP1879 and CPBP2032. The PM method has the effect of damping the significance of huge losses (which are outliers) such that the VaRs for internal and external data sets are of comparable size.

7.4 Use in practice

The profiles of the data sets considered, as illustrated in Figure 2, are essentially similar. There is an initial rise (which may be very rapid) to a peak value, followed

by a relaxation to a region of "deemed convergence", from which a final VaR figure may be extracted. In order to do this, much of the profile must be present. We recommend running PM calculations covering the entire range of 5–8000 partitions with a much reduced number of MCMC iterations. In practice, only 2000 are necessary to get a good idea of what the VaR profile will look like. That process takes about thirty minutes. Afterward, it is possible to focus on the relaxation portion of the VaR profile and to do more accurate calculations. The total time requirement for accurate calculations (at least 10 000 MCMC iterations, and preferably more than 25 000) is in the region of twenty-four to thirty-six hours. Given that the calculation need not be done frequently, that time is not too much of a burden. However, it is strongly suggested that multiple runs are done in the region near 8000 partitions to establish the stability of the result. Further, a trimmed mean of all results for multiple repeated runs should be used to determine VaR.

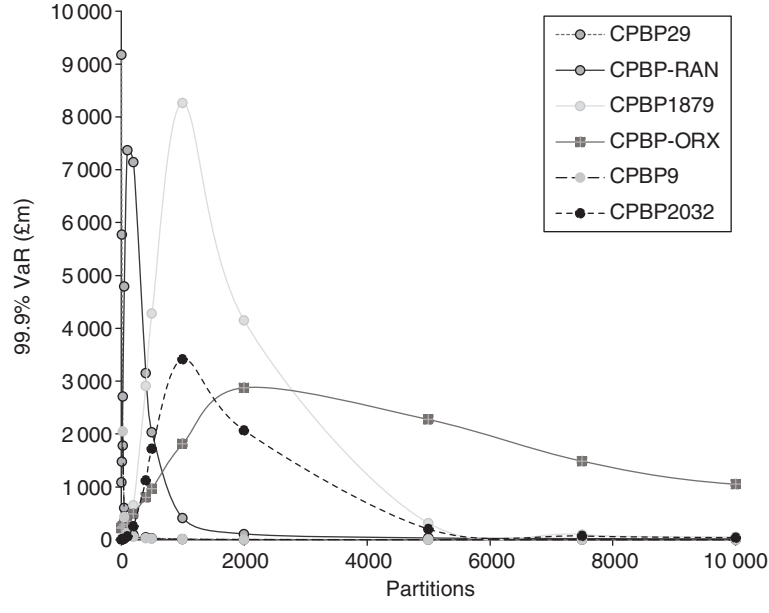
7.5 Uniform partition results

The PM method proceeds by partitioning the largest loss in a statistically sound way and calculating VaR using the partitioned data. In this section, we consider, for comparison, a simple partitioning that is not acceptable statistically but is nevertheless intuitively appealing. We call this the "uniform partition" method. In uniform partitioning, the largest loss is simply split into a prescribed number of equally sized sublosses that replace the largest loss. Although this partitioning method is easy, it is inconsistent with Basel regulations (Basel Committee on Banking Supervision 2011), which require that all losses used for modeling should originate from distinct operational risk events. A case can be made for subdividing the largest loss into y partitions, where y is the number of years spanned by the data. The reason for this would be to amortize the largest loss over a number of accounting years. However, the split into equally sized sublosses is arbitrary. Figure 4 shows the results. The data used for Figure 4 is given in Table 7 of Appendix C online. Note that the profiles in Figure 4 for CPBP29 and CPBP9 appear to be almost coincident.

The advantage of uniform partitioning is that it is fast. The time required to generate each point in Figure 4 was approximately 1.5 minutes using one million Monte Carlo cycles in the LDA process.

The profiles in Figure 4 mask a fundamental problem with uniform partitioning. With a sufficiently large number of partitions, it is possible to obtain a VaR value that is arbitrarily close to zero. Given the vertical scale in Figure 4, it is not clear whether a limiting value is zero or slightly greater than zero. However, Table 7 in Appendix C shows that VaR values do, indeed, tend to zero. In some cases, many iterations in the LDA procedure are needed. For example, CPBP-ORX needs approximately 100 000 partitions to see a near-zero limit. Further, partitioning by the number of years (or

FIGURE 4 Uniform partitioning of the largest loss y into n smaller losses z_i.



even the number of months) spanned by the data has little effect on "base" VaR. Both of these properties are clearly undesirable, so uniform partitioning is not a safe procedure.

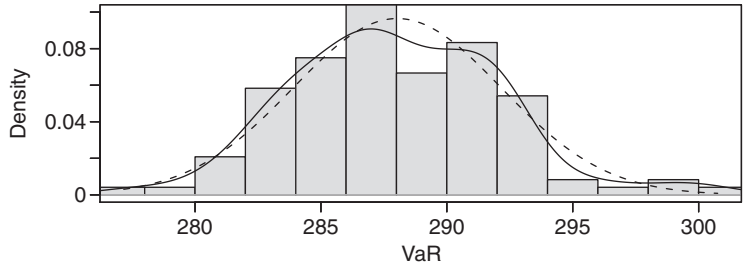
7.6 PM with importance sampling: sensitivities

Two sensitivities are discussed in this section. First, we consider the stability of the calculated VaR by repeating the entire VaR process multiple times. Then we consider the stability, in a single VaR calculation, of the calculated lognormal parameter values to the number of iterations in the MCMC process.

7.6.1 MCMC stability

To get some idea of the response of the PMMCMC process to its input parameters, the same procedure was run 120 times with the same set of parameters for 10 000 MCMC iterations, and the VaR values obtained were recorded. The data set CPBP2032 was chosen because it is typical of current internal data sets. The number of partitions used was 5600, which corresponds closely to the number required to generate the limiting VaR value given in Table 4 for CPBP2032 (£286 million).

FIGURE 5 MCMC stability (10 000 iterations).



120 VaR calculations for CPBP2032, 5600 partitions.

Figure 5 shows a histogram of the results obtained, with a superimposed density plot and fitted normal distribution.

The data for Figure 5 gave the following summary statistics, with a symmetric 95% confidence interval $(\mu_{\mathrm{L}}, \mu_{\mathrm{U}})$ based on a normal distribution:

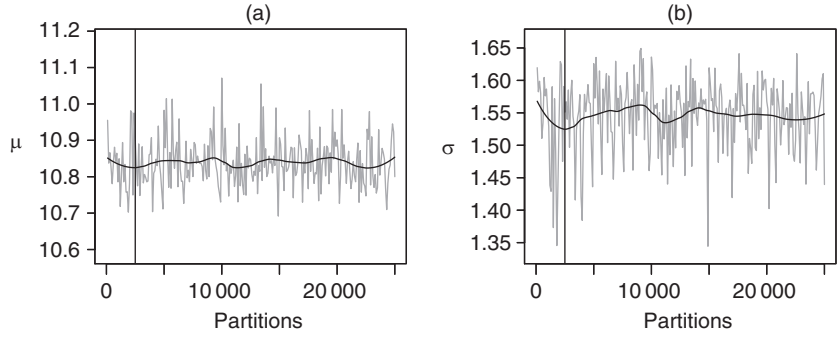
$$\left.\begin{aligned} \mu &= 288.0, \\ \sigma &= 4.14, \\ (\mu_{\mathrm{L}}, \mu_{\mathrm{U}}) &= (279.9, 296.1). \end{aligned}\right\} \tag{7.3}$$

The form of the profile in Figure 5 is approximately normal, and passes a TNA test for normality at 98.4% confidence. The dashed line shows a normal density based on 1000 samples, from a $N(\mu, \sigma)$ distribution (with parameters as in (7.3)). The time taken for each complete VaR calculation was approximately fifty-five minutes. Therefore, the time taken to generate the data for Figure 5 was approximately $55 \times 120 = 6600$ minutes ($\sim$110 hours).

7.6.2 Stability of lognormal parameters

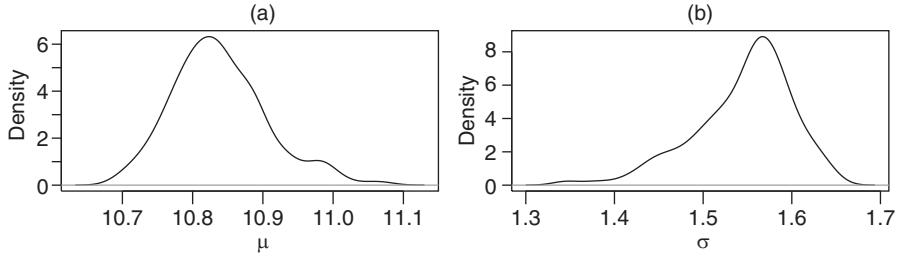
To study the effect of the number of MCMC iterations on the stability of the calculated lognormal parameters, the values of the lognormal μ and σ parameters were recorded during the course of one instance of the MCMC process with 10 000 iterations. The results are shown in Figures 6 and 7. In Figure 6, the gray traces show the actual parameter values recorded, each with a Loess-smoothed trace in black. The smoothing parameter in each case (ie, Loess parameter "span") was 0.3, which indicates local trend without excessive volatility. The vertical lines indicate the cutoff

FIGURE 6 PM importance sampling MCMC progression for lognormal parameters (a) μ and (b) σ.



CPBP2032 data, 5600 partitions.

FIGURE 7 Density plot for lognormal parameters (a) μ and (b) σ.

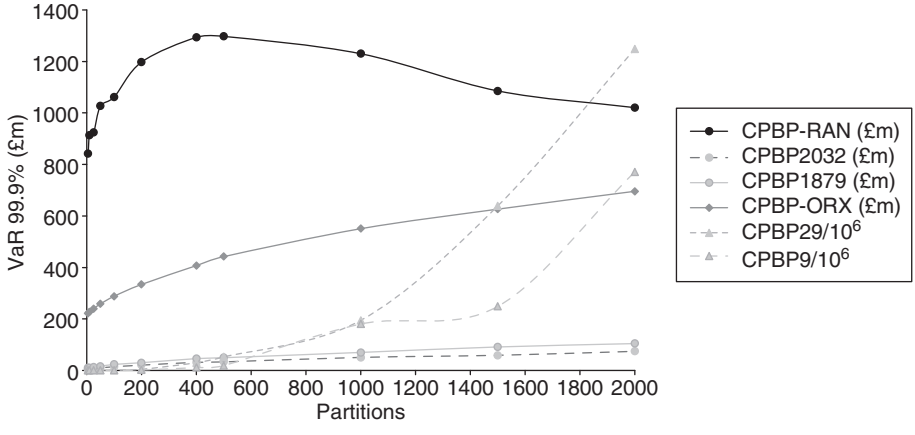


CPBP2032 data, 5600 partitions.

points for the MCMC burn-in (2500), which were used to derive the results in Figure 2. The corresponding density plots for μ and σ are shown in Figure 7. Both densities include only the post-burn-in values.

Although the parameter traces in Figure 6 vary by significant amounts in some cases, there is no overall trend. The trace of smoothed values is effectively horizontal. If the Loess "span" parameter is increased to 0.9 (ie, 90% smoothing), the μ smoothed trace is almost horizontal, and the σ smoothed trace has only a very minor concave appearance. These traces indicate that the mean values of the post-burn-in μ and σ values can be safely used without needing excessive MCMC iterations and that a short burn-in (eg, 25%) is also safe.

FIGURE 8 Fenton–Wilkinson approximation for partitioning the largest loss y into n smaller losses z_i.



The mean values for μ and σ for the post-burn-in period shown in Figures 6 and 7 were 10.84 and 1.55, respectively. The corresponding VaR was £292 million, which is well within the confidence interval in (7.3).

7.7 Fenton–Wilkinson approximation results

Figure 8 shows the results of approximating the sum of n partitions of the largest loss, each modeled by a lognormal distribution, by a single lognormal distribution that has parameters given by (6.5). The R code for the Fenton–Wilkinson approximation and the associated MH MCMC procedure are given in Appendix D online.

Two of the profiles in Figure 8 have been scaled in order to fit them all on one graph for comparison (for CPBP29, CPBP9, the actual VaR is the VaR indicated multiplied by 1 000 000). The points to note are the shape of the profiles and any apparent limiting behavior.

The actual values from which the profiles in Figure 8 were derived are given, without scale factors, in Appendix C.3 online.

The profiles in Figure 8 indicate that the Fenton–Wilkinson approximation is unsuitable for use in our context. All data sets except CPBP-RAN show increasing VaR as the number of partitions increases. The increase appears to be very slight for CPBP2032, CPBP1879 and CPBP29, but the values in Appendix C.3 confirm that it is significant compared with the magnitude of the VaR values observed. Some of the values listed in that appendix are extraordinarily large. For example, the largest VaR values obtained are in the region of £100 000 billion. Such values are even larger

than huge sovereign economic figures like the 2018 Q2 gross national product of the United States (US$15 057 billion)![4] Values that big are a warning that the method is not appropriate. The existence of a limiting VaR is doubtful. Arguably, CPBP-RAN does approach a limit near 2000 partitions, but it is difficult to tell, as floating point operations fail for more than 2000 partitions.

The advantage of using the Fenton–Wilkinson approximation is that the time taken to generate results is very short. Using 25 000 MCMC iterations, the slowest profile to generate (CPBP-ORX) in Figure 8 took less than five minutes, and the fastest (CPBP9) took less than one minute.

7.8 Inappropriate use of the PM method

In Section 7.2.1, the idea of using the PM method when it is not needed was considered. Specifically, the data set CPBP-RAN-LN was derived from the data set CPBP-RAN by removing its largest loss. As a result, CPBP-RAN-LN resembles the lognormal distribution from which it was generated as a random sample. Fitting a lognormal distribution to CPBP-RAN-LN resulted in lognormal parameters $\mu = 12.023$ and $\sigma = 1.888$. The LDA calculation using those parameter values yielded a "correct" 99.9% VaR of £640 million. That VaR value is appropriate given the type of distribution and its parameter values.

However, applying the PM method yielded a much lower limiting VaR value of £92 million. That value is clearly not appropriate since CPBP-RAN-LN does not satisfy the (somewhat loose) conditions set out in Section 1.3 that both the largest loss and the calculated VaR should be extreme values. The largest loss for CPBP-RAN-LN is only 16.6% of the sum of all losses, whereas the largest losses for other data sets considered in this paper are at least 35% of the sum of all losses. A small number of partitions of the largest loss for CPBP-RAN-LN result in approximately "correct" VaR values. Figure 9 shows how VaR varies with the number of partitions for CPBP-RAN-LN. The profile is similar to the profiles in Figure 3.

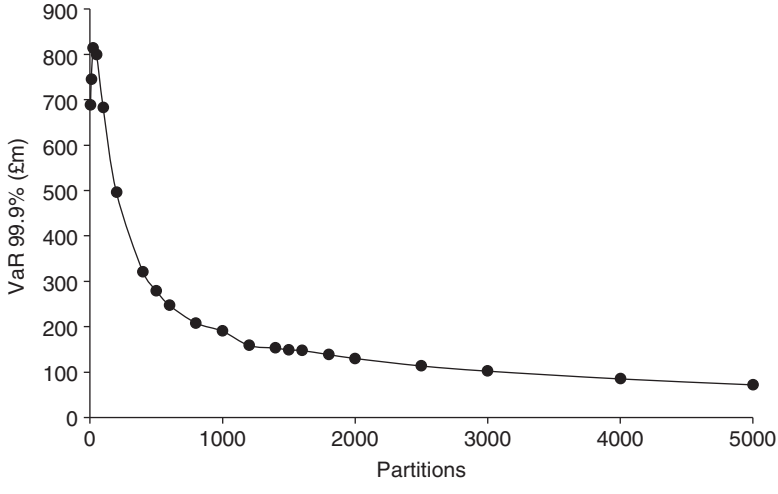
We stress again that discretion is needed when applying the PM method. It should be used only when there is a clear indication that other methods are inadequate. We consider that this decision should be qualitative, based on the judgment of risk managers.

8 CONCLUSION AND FURTHER WORK

The aim of this study was to determine a value for the 99.9% VaR for conduct risk losses such that the value obtained is consistent with the risk profile expressed by

[4] Converted from US$20 310 billion (https://fred.stlouisfed.org/series/GNP) at an average rate of 1.3489 for 2018 Q2 (https://www.bankofengland.co.uk/boeapps/database).

FIGURE 9 Use of the PM method in inappropriate cases.



those losses. The phrase "consistent with the risk profile" is not well defined, but practitioners have a very good sense of its meaning. The VaR value should not be so high that it actively prevents a bank from lending because of retention of the VaR amount as a reserve. There is no objective standard for assessing a maximum value for VaR given a set of losses other than a general idea of what the reserve should be. The VaR value should not be so low that any reserve cannot cover unexpected losses. The empirical bootstrap can be used to determine a minimum value, but the examples in Table 3 show that this metric is unreliable in the case of conduct risk losses. The imbalance between the largest loss (taken here to be a provision that comprises an unknown number of small payments to customers) and smaller losses can easily lead to gross distortions in a simple curve-fitting calculation. We have proposed the PM method as a way to estimate a VaR that avoids calculating an unrealistic value. The largest loss (assumed to be a provision comprising multiple unknown payments to customers) is partitioned in a statistically sound way in order to simulate the unknown components of the largest loss. The results show an apparent limit for VaR as the number of partitions increases, but an additional assumption – that a limiting VaR can be calculated using an exponential distribution – is still needed to be able to quote a "single" figure for VaR. A surprising consequence of applying the PM method to several data sets is that the final VaR values obtained are similar (apart from CPBP-ORX), despite the extreme dissimilarities of the original data sets. The internal data sets considered all represent the same risk profile. They differ because

of factors such as improved data collection, reallocation of losses to alternative risk units and error corrections.

8.1 Further work

The following are suggestions for continued work.

(1) Replace the lognormal distribution used in the PM method by an alternative such as a lognormal mixture or a generalized Pareto distribution.

(2) Partitioning as described in this paper is applied to the largest loss only. It could be applied to further large losses, although partitioning should be restricted to losses that are actually provisions. Extending partitioning in this way poses a combinatorial problem with respect to the number of cases that could be considered.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

ACKNOWLEDGEMENTS

The authors are grateful for the help of Professor George Deligiannidis at the Department of Statistics, University of Oxford, in the preparation of this paper.

REFERENCES

Alexander, C. (2001). *Market Models: A Guide to Financial Data Analysis*. Wiley.

Andrieu, C., and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics* **37**(2), 697–725 (https://doi.org/10.1214/07-AOS574).

Andrieu, C., and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistical Computing* **18**, 343–373 (https://doi.org/10.1007/s11222-008-9110-y).

Andrieu, C., and Vihola, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Annals of Applied Probability* **25**(2), 1030–1077 (https://doi.org/10.1214/14-AAP1022).

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* B **72**(3), 269–342 (https://doi.org/10.1111/j.1467-9868.2009.00736.x).

Basel Committee on Banking Supervision (2006). International convergence of capital measurement and capital standards. Revised Framework, June, Bank for International Settlements. URL: https://www.bis.org/publ/bcbs128.pdf.

Basel Committee on Banking Supervision (2011). Operational risk: supervisory guidelines for the Advanced Measurement Approaches. Report, June, Bank for International Settlements. URL: https://www.bis.org/publ/bcbs196.pdf.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

Bierkens, J., and Roberts, G. (2017). A piecewise deterministic scaling limit of lifted Metropolis–Hastings in the Curie–Weiss model. *Annals of Applied Probability* **27**(2), 846–882 (https://doi.org/10.1214/16-AAP1217).

Bierkens, J., Fearnhead, P., and Roberts, G. (2016). The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Annals of Statistics* **47**(3), 1288–1320 (https://doi.org/10.1214/18-AOS1715).

Bühlmann, H., and Gisler, A. (2005). *A Course in Credibility Theory and Its Applications*. Springer.

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *American Statistician* **49**(4), 327–335 (https://doi.org/dc9n).

Cobb, B., Rumí, R., and Salmerón, A.(2012). Approximating the distribution of a sum of log-normal random variables. In *Proceedings of the 6th European Workshop on Probabilistic Graphical Models, PGM 2012, Granada, Spain*. Department of Computer Science and Artificial Intelligence, Granada University. URL: http://leo.ugr.es/pgm2012/proceedings/eproceedings/cobb_approximating.pdf.

Cruz, M., Peters, G., and Shevchenko, P. (2015). *Fundamental Aspects of Operational Risk and Insurance Analytics: A Handbook*. Wiley (https://doi.org/10.1002/9781118573013).

Dahlin, J., Lindsten, F., Kronander, J., and Schön, T. B. (2015). Accelerating Monte Carlo methods for Bayesian inference in dynamical models. Preprint (arXiv:1511.05483) (https://doi.org/10.3384/diss.diva-125992).

Deligiannidis, G., Doucet, A., and Pitt, M. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society* **80**(5), 839–870 (https://doi.org/10.1111/rssb.12280).

Del Moral, P. (1996). Nonlinear filtering: interacting particle solution. *Markov Processes and Related Fields* **2**(4), 555–580.

Doucet, A., de Freitas, N., Murphy, K., and Russen, S. (2000). Rao–Blackwellised particle filtering for dynamic Bayesian networks. In *Proceedings of the 16th Uncertainty in Artificial Intelligence Conference (UAI-2000), Stanford, CA*, Boutilier. C., and Goldszmidt, M. (eds), pp. 176–183. Morgan Kaufmann, San Francisco, CA.

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**(2), 295–313 (https://doi.org/10.1093/biomet/asu075).

Fenton, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems* **8**(1), 57–67 (https://doi.org/10.1109/TCOM.1960.1097606).

Filippone, M., and Girolami, M. (2014). Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(11), 2214–2226 (https://doi.org/10.1109/TPAMI.2014.2316530).

Financial Conduct Authority (2018). Complaints data analysis: 2017-H2. Report, April, Financial Conduct Authority, London. URL: https://www.fca.org.uk/publication/data/complaints-data-analysis-2017-h2.pdf.

Financial Conduct Authority (2019). Monthly PPI refunds and compensation. Report, July, Financial Conduct Authority, London. URL: https://www.fca.org.uk/data/monthly-ppi-refunds-and-compensation.

Frachot, A., Georges, P., and Roncalli, T. (2001). Loss distribution approach for operational risk. Working Paper, Social Science Research Network (https://doi.org/10.2139/ssrn.1032523).

Gay, R. (2004). Pricing risk when distributions are fat tailed. *Journal of Applied Probability* **41**, 157–175 (https://doi.org/10.1239/jap/1082552197).

Gustavsson, F. (2010). Particle filtering theory and practice with positioning applications. *IEEE Aerospace and Electronic Systems* **25**(7), 53–81 (https://doi.org/10.1109/MAES.2010.5546308).

Kronander, J., Schon, T., and Unger, J. (2015). Pseudo-marginal Metropolis light transport. In *SIGGRAPH Asia 2015 Technical Briefs*, Article 13. ACM, New York (https://doi.org/10.1145/2820903.2820922).

Lin, L., Liu, K., and Sloan, J. (2000). A noisy Monte Carlo algorithm. *Physical Review* D **61**(7), 074505 (https://doi.org/10.1103/PhysRevD.61.074505).

Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing*, pp. 31–48. Springer Series in Statistics. Springer (https://doi.org/10.1007/978-0-387-76371-2).

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2004). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100**(26), 15324–15328 (https://doi.org/10.1073/pnas.0306899100).

Mitic, P. (2015). Improved goodness-of-fit measures. *The Journal of Operational Risk* **10**(1), 77–126 (https://doi.org/10.21314/JOP.2015.159).

Mitic, P., and Bloxham, N. (2018). A central limit theorem formulation for empirical bootstrap value-at-risk. *The Journal of Model Risk Validation* **12**(1), 49–83 (https://doi.org/10.21314/JRMV.2018.182).

Murray, I., and Graham, M. (2016). Pseudo-marginal slice sampling. *Proceedings of Machine Learning Research* **51**, 911–919.

Peters, G., and Cornebise, J. (2009). Comments on "Particle Markov Chain Monte Carlo" by C. Andrieu, A. Doucet and R. Hollenstein. Preprint (arXiv:0911.3866).

Peters, G., and Sisson, S. A. (2006). Bayesian inference, Monte Carlo sampling and operational risk. *The Journal of Operational Risk* **1**(3), 27–50 (https://doi.org/10.21314/JOP.2006.014).

Peters, G., Briers, M., Shevchenko, P., and Doucet, A. (2013). Calibration and filtering for multi factor commodity models with seasonality: incorporating panel data from futures contracts. *Methodology and Computing in Applied Probability* **15**(4), 841–874 (https://doi.org/10.1007/s11009-012-9286-7).

Peters, G., Targino, R. S., and Wüthrich, M. (2017). Bayesian modelling, Monte Carlo sampling and capital allocation of insurance risks. *Risks* **5**(53), 1–51 (https://doi.org/10.3390/risks5040053).

Picchini, U. (2018). Why and how pseudo-marginal MCMC works for exact Bayesian inference. Article, March 26, Umberto Picchini's Research Blog. URL: https://bit.ly/32U1Is3.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2018a). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association* **114**(526), 831–843 (https://doi.org/10.1080/01621459.2018.1448827).

Quiroz, M., Villani, M., Kohn, R., Tran, M.-N., and Dang, K.-D. (2018b). Subsampling MCMC: an introduction for the survey statistician. *Sankhya A* **80**(Supplement 1), 33–69. https://doi.org/10.1007/s13171-018-0153-7.

Robert, C. P. (2016). The Metropolis–Hastings algorithm. Preprint (arXiv:1504.01896v3 [stat.CO]).

Rook, C., and Kerman, M. (2015). Approximating the sum of correlated lognormals: an implementation. Preprint (arXiv:1508.07582 [q-fin.GN]).

Septier, F., and Peters, G. (2015). Langevin and Hamiltonian based sequential MCMC for efficient Bayesian filtering in high-dimensional spaces. *IEEE Journal of Selected Topics in Signal Processing* **10**(2), 312–327 (https://doi.org/10.1109/JSTSP.2015.2497211).

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2017). The block pseudo-marginal sampler. Preprint (arXiv:1603.02485v5 [stat.ME]).

Treanor, J. (2016). Bill for PPI mis-selling scandal tops £40bn. *Guardian*, October 27. URL: https://bit.ly/2B1Kc92.

UK Parliament (2013). Panel on mis-selling and cross-selling: written evidence from Which? UK Parliamentary Publications and Records SJ 015, UK Parliament. https://bit.ly/2ooCaEm.

Wu, J., Mehta, N., and Zhang, J. (2005). Flexible lognormal sum approximation method. In *IEEE Global Telecommunications Conference, GLOBECOM '05*, pp. 3413–3417. IEEE Press, Piscataway, NJ (https://doi.org/10.1109/GLOCOM.2005.1578407).

Zheng, X. (2016). Pseudo-marginal Metropolis–Hastings approach and its application to Bayesian copula model. MSc Thesis, School of Mathematics and Statistics, UNSW Sydney, Australia. https://web.maths.unsw.edu.au/~josefdick/preprints/Zheng_thesis.pdf.