Dissecting Initial Margin Forecasts: Models, Limitations and Backtesting

Vladimir Chorniy[*], Sergii Arkhypov[**]

Initial version: 28 July 2023

Current version: 24 August 2024

Abstract: Initial Margin (IM) has significant impact on counterparty exposure, pricing, capital and limits. This makes modelling of future IM a critical part of counterparty risk management and pricing. All current approaches, which we review in this paper, consider IM as Value-at-Risk (VaR) and therefore develop methods for forecasting such forward VaR and some move to propose backtesting of the VaR forecasts for model verification. We show that this approach is limited and that it also biased the industry in terms of backtesting and model verification. This article attempts to correct this. Firstly, we discuss that IM is not VaR, VaR being only its approximation. Then we show that even with 'IM is VaR' assumption, the forecast of IM is *forecast of a forecast*, which is principally different from 'just' forward VaR forecasting. We demonstrate fundamental limitations which follows from this, and after reviewing existing literature on IM forecasting, we propose a generic backtesting and verification framework accommodating both forecasting limitations and existing models. For model verification we consider two approaches: direct backtesting/monitoring and elicitability related approach. Our analysis also includes the special case of bank's model of IM/VaR being near perfect replica of an exchange's IM.

Key words: Counterparty risk, Backtesting, IMM, Initial Margin, SIMM, CCP.

JEL classification: G13; G15; G17.

1. Introduction

Posting of Initial margin (IM) was always an important part of Central Clearing Counterparty (CCP) risk management practice, but after publication of requirements to exchange IM for

[*] BNP Paribas, Risk Analytics and Modelling, London. *E-mail*: vchorniy@yahoo.com.

[**] BNP Paribas, Risk Analytics and Modelling, London. Corresponding author. *E-mail*: sarkhypov@gmail.com.

majority of non-cleared OTC[1] derivatives (Basel Committee on Banking Supervision, 2015), it became mandatory to post IM across most[2] of counterparty risk perimeter. IM has significant impact on counterparty exposure and consequently on capital and limits. By affecting exposure it also impacted CVA/DVA/KVA[3] and led to creation of dedicated valuation adjustment - Margin Value Adjustment (MVA). This results in increasing interest towards IM modelling.

The industry is still adjusting to this new margining regime and as yet final consensus has not been reached on the way to forecast future initial margin requirements as part of counterparty credit risk (CCR) modelling. Replicating every CCP/exchange methodology and the Standard Initial Margin Methodology (SIMM) is very costly and probably unfeasible. Therefore a modelling relying on some form of proxy and leveraging on institution's own CCR calculations is attracting attention. Such way of thinking gives a rise to different flavours of so called Dynamic Initial Margin (DIM) models, see review in (Caspers, et al., 2017) and later in this paper.

In this article, we discuss similarities between DIM models which allow us to propose a backtesting approach applicable to all of them. The article is organised as follows. First in Section 2, we discuss differences between Value-at-Risk (VaR) and IM, which appear to be neglected in current models. This informs the discussion in Sections 2 and 3 about expected quality of future predictions for IM. Then in Sections 4 and 5 we review available Dynamic IM models and specifically analyse their common underlying features. This analysis allows the construction of single general backtesting and monitoring approach for all models discussed in Sections 6 and 7. In the Section 8, we outline alternative IM verification approach broadly related to the concept of elicitability and comparative backtesting. Section 9 concludes.

2. VaR and Initial Margin, the same, only different

The main purpose of IM is to provide cover for a change in the (netted) portfolio value over the margin period of risk (MPOR). This change closely relates to the definition of Value-at-Risk, as it represents Counterparty VaR at 1% quartile instead of typical 99%. This created a tendency within the industry to model IM as if it was VaR[4]. There are two major problems with this approach.

[1] OTC: Over- the-Counter.

[2] The requirement to post IM for OTC was introduced in stages. The first stage which required IM for largest counterparties went live from September 2016. The last 6th stage for counterparties with an average aggregate notional amount of more than €8 billion went live in September 2022.

[3] CVA: Credit Value Adjustment, DVA: Debt Value Adjustment, KVA: Capital Value Adjustment

[4] Examples for IM forecasting based on 'IM is VaR' assumption will be given later in this paper, but this assumption has wider use, see, for example, a paper discussing the assessment of IM methodologies (Murphy, 2023).

Firstly, in most cases live IM models are rather removed from 'pure' VaR and often they are much more than VaR. Whereas in the case of non-cleared OTC standard industry approach to calculate IM, SIMM[5], is approximately a sensitivity based nested var-covar model with limited diversification, which at least has VaR as an inspiration, in case of CCPs, IM are much further diverged from VaR. For example, Eurex IM takes account of Filtered Historical VaR, Stressed VaR, Event Risk, Model Adjustments and Liquidity add-on[6] (Eurex Clearing, 2018), while CME uses well known SPAN, which relies on 16 pre-defined scenarios and add-ons (CME Group, 2019). These two examples are representative of the large population of CCP's. In short, *IM is not VaR*, at best VaR is only an approximation for IM, and in some circumstances a poor one. Therefore ability to forecast VaR, on which, as we will discuss later in this paper, the industry research has concentrated, is not the same as forecasting IM, there are important methodological differences.

Secondly, even if IM was equal to VaR, i.e. '*IM is VAR'*, in case of CCPs a bank would need to forecast IM, i.e. VaR, calculated by some other external entity. Let us assume as an illustration that this entity's VaR model is faulty and the bank's one model is perfect. In this case, good VaR forecast by bank's internal VaR model will not match faulty CCP's one, and as a forecast of CCP's IM it is a failure. Of course, in practice given efforts from regulators and systemic importance of CCPs, it is reasonable to assume that CCP or SIMM IM models are fit-for-purpose models. 'Fit-for-purpose' means capable of meeting internal and supervisory requirements[7]. However, even assuming that CCP's IM is a good VaR forecast, the task for the bank's IM model is not VaR forecast, but a forecast of somebody else's VaR forecast. As we will show below the difference on any given day between two (different) *fit-for-purpose* VaR models can be significant, so one fit-for-purpose VaR model could fail (at particular point in time) to forecast VaR from another fit-for-purpose model.

Now we will illustrate how industry standard VaR model types can differ from each other on day-to-day basis, and thus demonstrate the *fundamental limits* in IM forecasting using VaR forecasting approach. These fundamental limits to forecasting could be examined through the limits of the replication of one VaR model with another VaR model. (Although the replication error of today's VaR could be removed through scaling coefficient, it will resurface in the forecast. The replication error and the forecast error driven by it are fundamentally the same. This will be clearer once such scaling is discussed in Sections 5 and 6.) We will compare fit-

[5] Standard Initial Margin Model is a standardised methodology developed by financial industry and published by industry body, International Swaps and Derivatives Association, (International Swaps and Derivatives Association, 2018).

[6] One could argue that all such add-ons enhance the ability to predict an extreme percentile and thus is part of broadly defined VaR, but we use VaR in narrower, commonly accepted sense where portfolio VAR is driven by the co-evolution of risk factors to fixed horizon assuming liquidity. This use reflect industry practice including CCPs', hence the discussion of VaR vs. add-ons in CCPs' public materials.

[7] 'Fit-for-purpose' is the term which is used frequently in this paper and warrants a brief discussion. A fit-for-purpose model meets requirements of all relevant stakeholders. The stakeholders could be internal and cover risk management, IT and trading, and external, the supervisors. The requirements could be precisely defined, for example, via backtesting benchmarks, or could be more approximate in terms of the model reactivity and procyclicality; IT may require compatibility with existing infrastructure. Note that fit-for-purpose definition does not demand for the model to be the best, it carries no view on the relative merits of one model vs. another.

for-purpose[8] VaR models in three steps with each step retaining the fit-for-purpose quality of the models, but separating them further: 1) ignore the models' methodological differences and consider the only numerical noise/uncertainty; 2) models are identical, but their calibrations (specifically data) are different, with each calibration fit for purpose; 3) models are different, but both fit for purpose and industry standards, and use the same data for calibration. Our illustrations are not exhaustive, there are a lot more reasons for the divergence (e.g., market data source, etc.). This only strengthens our point regarding the limits of VaR targeting approach.

1) For this first step illustration we consider numerical noise in Monte Carlo VaR and historical VaR. Majority of internal banks' CCR models (IMM and xVA[9]) are relying on a Monte Carlo (MC) approach and thus are subject to uncertainty of this method. Level of MC error depends on the number of simulations, standard deviation of MC estimation of 99% percentile are approximately 5% with 5000 simulations and 12% with 1000 simulations (assuming normal distribution). Let us assume that MC VaR (based on internal CCR engine) will be used to forecast VaR (i.e., IM) calculation by CCP which is likely to be related to historic VaR (HVAR). For HVaR, again assuming normal distribution, the corresponding error is 14% for 750 simulations (3 years of history), reaching 24 % for 250 simulations (1 year). MC numerical noise is likely to be visible daily if random generator seeds change daily, whilst HVaR numerical error is hidden. However from the point of view of numerical error HVaR is identical to MC VaR with 'black box' methodology with only 250 simulations. Thus, assuming 5000 simulations for MC VaR and one year based HVaR, VaR targeting approach is trying to forecast one VaR forecast with 24% of uncertainty with another forecast with 5% of uncertainty. Note, this is before any difference between methodologies is taken into account.

2) In this second illustration we consider MC VaR only. If a bank's internal MC based CCR/xVA model to be used as CCP's VaR predictor, the potential differences within MC CCR models (viewed as VaR) are relevant. Basel regulations require minimum 3 year time series for calibration. Thus, for example, calibrating the same model using either 3 or 5 year time series is still fit-for-purpose. So if we consider the simplest case of VaR on a single stock portfolio (gross simplification of real portfolios, indeed, but still relevant as counterparty portfolios, especially for corporates, tend to be directional and often driven by limited number of risk factors), then VaR will be proportional to stock volatility. We compared the relative difference between 3 year and 5 year volatilities for 1 day returns for each individual stock in the Euro Stoxx 50 index from 2004 to 2019. For example, volatility of 2015-2018 interval will be compared per stock to volatility of 2013-2018 interval respectively. For the entire period the

[8] We do not formally prove that actual simple models we use are fit-for-purpose, however all of them are model types used in the industry, and as such recognised as fit-for-purpose by practitioners and regulators. Thus the lack of formal validation of our 'toy' models does not reduce the generality of our argument.

[9] Internal Modelling Method: CCR methodology used both for regulatory reporting and (with possible small differences) for bank's internal risk management. The regulatory requirements were introduced in Basel 2, further specified in Basel 3 regulations. xVA is a generic designation of various valuation adjustments, such as CVA or KVA introduced earlier.

average of 90% percentile of relative error for any given day is 20.6%, corresponding standard deviation is 9.3%, whilst the largest error is 45.2%.

3) For final illustration we select three different model types, all are fit-for-purpose and representing existing industry standards, using the same data for calibration: 3 years. We compare HVaR, Filtered HVaR (FHVaR) and MC VaR of equity portfolios. HVaR is based on unweighted relative returns and uses rolling daily calibration; in FHVaR returns are normalised rolling 6 months volatility[10], and also uses rolling daily calibration; and MC VaR is based on lognormal distribution (i.e. GBM, geometric Brownian motion, is assumed) and uses annual recalibration based on preceding 3 years of observations.

We apply these VaR models to long-short portfolio constructed by assigning random weights between -1 and 1 to stocks of the Euro Stoxx 50 index and then rolling all three VaR estimates for the period from July 2007 to July 2019. Figure 1 below shows the results for one such portfolio as well as portfolio's daily PnL (profit and loss, i.e., portfolio MtM changes).

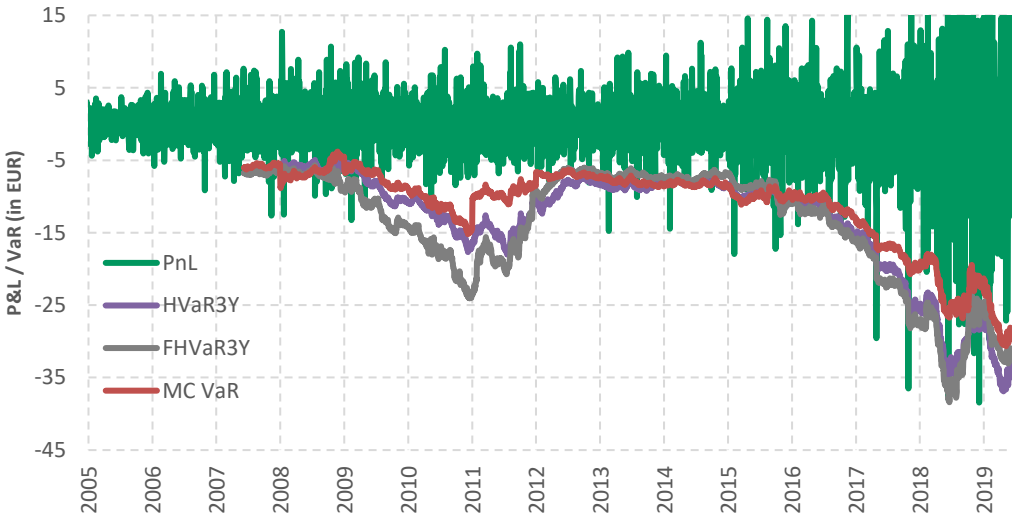

Figure 1: Comparison of different VaR methodologies for static long-short portfolio Euro Stoxx 50. 1-day VaR is shown

The relative difference between methodology pairs is presented on the Figure 2. The "A-to-B" relative difference is calculated as $(VaR\ A - VaR\ B)/(VaR\ B)$.

[10] The calculation of FHVaR is based on idea of filtered historical simulation, which requires normalisation by an estimate of instantaneous or spot volatility. For an overview of filtered historical simulation see (Gurrola-Perez & Murphy, 2015) and references therein. The use of standard deviation of six months of daily returns (annualised) as spot volatility estimate is a simplification compared to the usual use of more advanced (and more dynamic) approaches such as GARCH (Generalized Autoregressive Conditional Heteroscedasticity) or EWMA (Exponential Weighted Moving Average), but for our purposes does not reduce the generality of the results.

Figure 2: Relative differences between the VaR methodologies

Figure 2 clearly shows that all three VaR methodologies differ significantly from each other nearly through entire length of the observations. Even under benign market conditions (between 2012 and 2019 marked by red dotted outline) relative difference between them could vary within the range of 30%, with difference variously realised on time scales from a day-to-day to longer periods, – it can persist days, weeks, months or even years. Figure 2 is a representative[11] of the relative VaR behaviour on such portfolios. To verify that the effect is persistent we repeated this numerical experiment 100 times using 100 randomly weighed long-short portfolio from the Euro Stoxx 50 index. Mean of the relative error distribution is approximately zero as expected if all three models are fit-for-purpose models, coefficient of variation[12] (CV) of relative error is highest for FHVaR to MCVaR ratio and is approximately 15% for calm markets (2013-2019) and 30% for FHVaR over entire period, which includes major financial crisis. Assuming normal distribution this means that relative error could reach more than 30% in 5% cases even in calm markets. We also conducted the same analysis on portfolios with non-static weight (replicating constant sum invested) to exclude the growth of absolute value of VaR and the observed relative errors have the similar pattern (results not presented).

At this point we note that the results above are done for 1-day VaR, which is mainstay of majority of VaR modelling. However IM is usually designed to cover risk over longer period: from, for example, few days for some CCPs to 10 days for SIMMS. To verify that our conclusion stands we repeated our analysis for 5-day and 10-day VaRs resulting in similar,

[11] We will further discuss this example in Section 6.2.

[12] Also called relative standard deviation.

slightly higher CV's[13]. Figure 3 illustrate relative differences between the VaR methodologies for 10-day VaR for same portfolio as Figure 2.

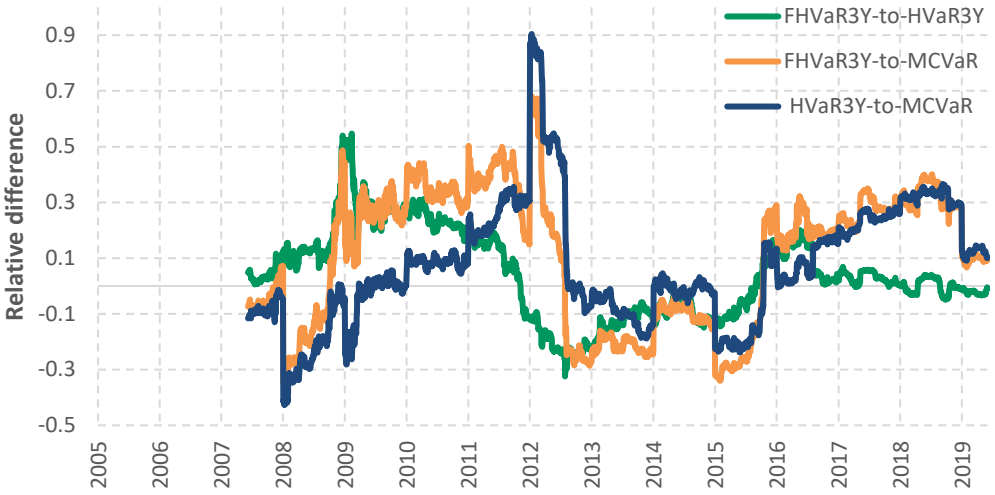

Figure 3: Relative difference of different VaR methodologies, 10 days

In practice an attempt to approximate an external fit-for-purpose VaR measure with an internal fit-for-purpose measure has even more sources of the diversions. Apart of methodological differences, discrepancy could be caused among others by the following areas: market data and data providers, the timing of the market data updates, and calibration frequency (both internal and external systems could have different recalibration schedules and frequencies).

At this stage it is important to note two points. Firstly, we have shown, that even if IM is VaR, and strictly speaking it is not, forecasting one IM/VaR with another VaR carries irreducible uncertainty on day-to-day basis even if both models are good. Obviously, the choice of the target accuracy is the choice of an individual bank, however if viewed from the angle of uncertainty our results above suggest that requesting an error to stay below roughly 30% 'most of the time' is a reasonable and relatively strict requirement, which even may have to be relaxed at the times of market turmoil. We shall return to the discussion of acceptable error and the methods of monitoring and backtesting of IM models in the later sections.

Secondly, as we will show in the review of existing models for IM forecasting in Section 4, practically all industry approaches (except for the full replication) are based on ability to forecast VaR in the future. The assumption is that once a good model to forecast forward VaR

[13] In this footnote for completeness we provide more details of mean and CV of relative error. Although mean of relative error is close to zero, there a few small persistent biases which reflect well know properties of three fit-for-purpose models and support our earlier assertion that our 'toy' models are representative of each type (cf. footnote 8). GBM based MC VaR lacks fats tails and on average underestimates 99th percentile approximately by 5-10% in calm period and 15-20% for entire period (higher for 1-day, lower for longer periods). The relative behaviour of filtered vs. 'plain' historic VaR also follows a familiar pattern, with FHVAR being higher in crisis period and lower in quiet period.

The CVs for 1-day VaR for FHVaR/MCVaR, HVaR/MCVaR , FHVaR/HVaR are 14%, 11% , 13% for calm period and 32%, 16%, 29% for full period and for 10-VaR are 20%, 15%, 15% and 33%, 18%, 27% correspondingly.

is built, the job of forecasting of IM is complete. We have shown that this is clearly not true. However, our criticism is not saying that the industry effort is misguided, we are saying it is one-sided. A practical approach to forecast IM *is* likely to involve ability to forecast forward VaR, – it is a necessary, but not a sufficient requirement for successful IM modelling, as at the very least it needs to be coupled with the requirement to understand and monitor model risk involved. We will discuss backtesting and monitoring approach for all models in Sections 6 and 7.

Having discussed the fundamental limitation to the accuracy of IM forecasts, the next step will be to examine whether there are any regulatory indication of tolerance to IM forecasting errors. The following section will address this matter.

3. Potential regulatory view of tolerated uncertainty

In TRIM CCR[14] (European Central Bank, 2019) European regulators introduced boundaries to EEPE[15] uncertainty with the view of tolerated capital accuracy. In effect, although perhaps it is not regulator's intention, it provides the general level of tolerated uncertainty of IMM CCR model. However as we discuss in the next section, which reviews various approaches of IM modelling, within IMM CCR model IM is expected to be forecasted with the same model: it is used to generate bank's internal forward VaR forecasts which then are converted into IM forecasts. Then the EEPE error tolerance could be also used to establish boundaries for quality of IM forecast. The regulatory prescribed boundaries are related to the levels of numerical error only. More precisely (European Central Bank, 2019) and (European Central Bank, 2017) suggest boundaries for EEPE Monte Carlo Error at 5% and 10%. As forward VaR is effectively a collateralised PFE, the task, in terms of CCR methodology, is to convert the EEPE error into PFE error.

Intuitively, this error conversion is driven by two factors. One factor is the linking of the uncertainty of expectation at given time t into uncertainty of percentile at the same time point, i.e. $EE(t)$ to $PFE(t)$, a well-known exercise. This depends on the shape of the distribution and therefore on a process(es) driving the exposure. The second factor is the fact that EEPE is the average of non-decreasing function of *EE(t)* over time. Thus EEPE stability depends on the shape of EE function. For example, the upward sloping EE function translates into the same EEE shape and assuming relative error of $EE(t)$ as independent of time, the EEPE numerical noise will be driven by the high end of EE/EEE curve. The upward sloping EE profile reduces the effect of averaging across time compared to the flat EE profile.

[14] TRIM CCR: Targeted Review of Internal Models Counterparty Credit Risk.

[15] EEPE stands for expected effective positive exposure, EE: expected exposure, EEE: effective expected exposure, PFE: potential future exposure, IMM: internal model method.

Therefore generally speaking the conversion of the error from EEPE to PFE/VaR is both model and portfolio dependent. We have explored different EE functions based on different processes and, indeed, the conversion coefficient varies, however we believe that a simple case of geometric Brownian motion (GBM) process is a good overall representation for CCR purposes, see (Chorniy & Arkhypov, 2020) for details.

By varying seed and number of MC paths we could obtain distribution of EEPE and corresponding set of PFE profiles. ECB[16] in (European Central Bank, 2019) defines MC EEPE uncertainty as normalised standard deviation multiplied by 1.96. For consistency PFE uncertainty of each horizon is calculated similarly. Overall PFE uncertainty is an average across all horizons. Figure 4 below provides EEPE to PFE MC error conversion for collateralised position on an equity contract for difference which depends on one risk factor diffused with GBM, 1% drift and 20% volatility. MC simulation has 24 steps over 1 year (corresponding to the PFE margin period of risk of two weeks or 10 day VAR) and gold standard collateral agreement is assumed (zero minimum transfer amounts and thresholds), so collateral equals to contract value at previous time point. Each point on the plot correspond to one MC run with unique seed. The points are in groups as some MC runs had equal number of simulations.

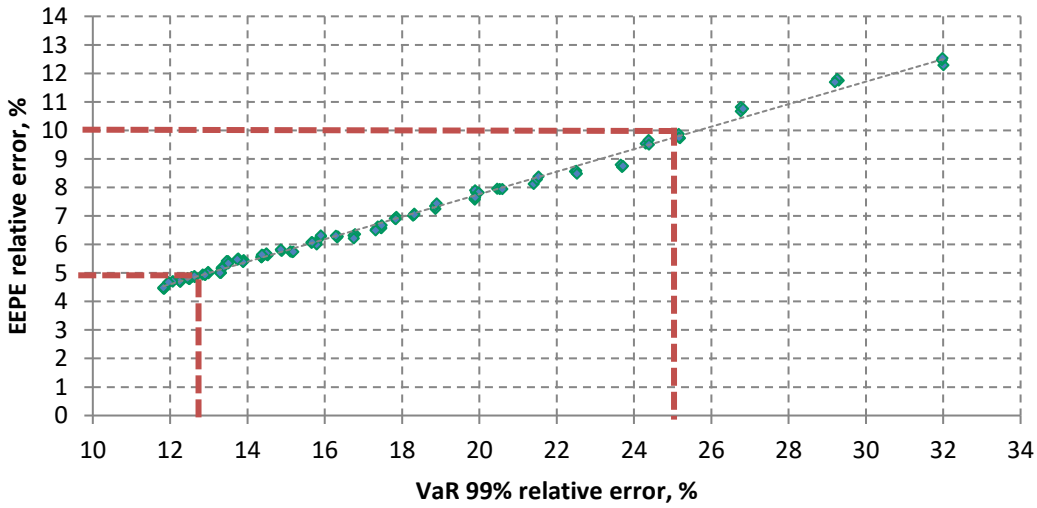

Figure 4: Relation between EEPE relative error and PFE error

This result demonstrates that regulatory tolerated level of EEPE uncertainty could be approximately translated into initial margin uncertainty with the coefficient of 2.5. This means 5% and 10% EEPE uncertainty translates into in an approximately 13% and 25% for IM uncertainty. It is a pure coincidence, but perhaps a lucky one, that 25% is very close to 30% tolerance which our discussion of fundamental uncertainties implies.

[16] European Central Bank.

4. Review of IM forecasting models. Looking where the light is

In this section we review existing industry approaches to model future IM requirements. Our aim is to understand the fundamental limitations of these models as well as their common features, so in the later sections we can develop a standard method for their assessment and backtesting.

All models to forecast initial margin could be split into three different classes: Non-simulating models; Approximate models; and Replicating models. Formally speaking latter two classes could be referred to as Dynamic IM (DIM) approach, however the established practice is to only refer to Approximate models as DIM, we will also use this convention.

There are two different types of Non-simulating IM models. First type, widely known as 'Flat Initial Margin', assumes that future margin level is equal to the IM held currently and flat out to the longest maturity. Second one, improves on the first by allocating part of total IM to each trade and then by assuming that each allocated amount stays constant through the life of the trade. This captures roll off effect. We suggest that non-simulating models may be too crude to satisfy strategic IMM requirements in a major bank or even more so the requirements of an advanced xVA desk.

At the opposite side, Replicating models aim at a full replication of IM calculation algorithm at a future point in time while simulating only the evolution of input variables. Within this class of models, first possible option is a brute force approach when future initial margin is calculated directly from MC evolved market data. The use of the brute force is very computationally demanding, with additional complexity that nearly each CCP requires separate calculation methodology[17] and sometimes not all details of the calculation are readily available even to CCP members. Second option is to approximate brute force by applying specific numerical techniques for calculation of path-wise future value of greeks. Such models were introduced in (Antonov, et al., 2018) and (Chan, et al., 2018) and are specifically relevant to SIMM case as they aim to achieve results comparable with brute force approach under less computational burden. Thus we list them as Replicating models for SIMM. Since SIMM is viewed as an approximation of VaR, these models can be also applied to cleared derivatives: SIMM approximates forward VaR using scenarios of the bank's CCR internal engine, and this approximate VaR is in turn used to approximate CCP's VaR (Antonov, et al., 2019). However, in this case the model becomes an approximate one and we will classify it as such.

Remaining models fall into a broad approximate class, for which an emerging common term is Dynamic Initial Margin (DIM) approach. Most of these models aim to satisfy IMM requirements by accounting for contractual details and forward variability of netting sets. The

[17] CCPs licensing its methodology, for example, SPAN methodology, may be covered by the same replica.

models which calculate IM directly and provide for practical implementation achieve this by using already existing CCR engine to estimate initial margin requirements, thus reducing the complexity and the expense of implementation. (Caspers, et al., 2017) and (Chan, et al., 2018) provide an early overview of Dynamic Initial Margin models, whilst (Ganesan & Hientzsch, 2021) overview adds more modelling approaches including the use of Johnson distributions introduced by (McWalter, et al., 2022), as well as considers using nested MC with limited number of nested paths.

1. The regression DIM

This is the first and the widest type introduced by (Green & Kenyon, 2015), (Andersen, et al., 2017) and (Anfuso, et al., 2017); further discussion can be found in (Caspers, et al., 2017). IM is assumed to be proportional or equal to VaR, which in turn could be assumed to be proportional to variance of PnL. Then variance is scaled to VaR with a constant, generally based on an assumption of normality. Also one can create sub-classes on the specifics of regression: the conditional variance is estimated via regression of simulated MtM changes or risk factors.

The Gaussian assumption can be avoided by using, for example, Johnson distribution (McWalter, et al., 2022). For detailed discussion see (Ganesan & Hientzsch, 2021) and references therein. If nested MC with limited number of nested paths is used to improve regression accuracy and does not aim for full replication, it still falls under this type in an Approximate category.

In this model type IM forecast is MC path dependent (MC paths generated by CCR MC engine).

2. Simple (scaled VaR) DIM

This type of model considers, for a given future point in time, the portfolio MtM moves over the margin period of risk across all scenarios, select the 99% quantile move (VaR), and use this as scenario-independent IM estimate as presented in (Wilkens & Moran, 2017). Unlike some of regression DIM models, this method does not rely on the assumption of normal distribution, but IM is only a function of time and is not dependent on individual CCR MC scenarios.

3. Sensitivity based IM

We already discussed that modelling of SIMM (replicating or based on numerical approximation) once treated as VaR becomes a VaR approximation driven by portfolios sensitivities. (Caspers, et al., 2017) and (Ganesan & Hientzsch, 2021) overview also other approximations, e.g. delta-gamma VaR. This model type produce IM forecast which is MC path dependent.

We will refer to path-dependent models as stochastic.

For completeness we should state that there is another set of approximate models which could be considered as IM models: Exogenous models. This is a broad class which covers approaches which do not directly model variables (e.g. risk factors, MtM etc.) driving IM and some do not provide value for IM at all, but model some other quantity (e.g. EE or MVA) which incorporates the impact of IM. In case of EE it is based on scaling EE approach: IM is implied indirectly within exposure itself via ratio between EE with and without initial margin ("efficiency ratio"), see (Andersen, et al., 2017), but not actually calculated. (Lou, 2016) addresses MVA pricing directly again only implying IM under redundant reserve assumption, whilst using sensitivity based type if direct modelling of IM is needed (delta-approximated VaR)[18]. Such exogenous models may satisfy capital calculations, but do not provide PFE projections, which is needed for day-to-day risk management, nor could be easy subject (if at all) of backtesting.

In all methods if IM calculated as VaR does not match actually observed IM (i.e. CCP's IM or SIMM) additional scaling is proposed. Such scaling is explored in (Anfuso, et al., 2017) and (Caspers, et al., 2017), but, for example, is only briefly mentioned in (Ganesan & Hientzsch, 2021). Such scaling is an important part of IM model and we will dicuss it in the next two sections.

At this point we note that all DIM models are "looking only where the light is": firstly, they all assume IM to be VaR and secondly, only partially address via scaling the fact that target IM/VaR is an output of another external model. That said, this methodological homogeneity allows us to observe structural similarities between all these models, which will be discussed in the next section. This in turn will allow us to propose a generalised model monitoring and backtesting approach later in the paper.

5. Analysis of structural components and verification of IM models

In this section we start by considering structural commonalities across DIM models. This will allow us to propose a general and largely model independent verification framework.

Hallmark of all models of this class is an attempt to approximate initial margin calculated by external model (e.g. SIMM, CCP's) with an output from existing internal model, f, by inclusion of scaling function or constant, α. As such we could present all DIM models forecasts in generalised form, as:

[18] Any classification is, of course, an approximation. (Trillos, et al., 2015) models IM directly as a process (not clear how generalise this method across large multi-asset portfolios and/or with non- linear instruments), so it could be viewed as an exogenous model, but it also look similar to simple DIM type, but without use of internal CCR engine, with IM process playing the role of MC CCR engine scenarios and valuation functions.

$$IM(t,h) = \alpha(t,h) \times f\left(PV_k^{[j]}(t,h), \Delta PV_{n,k}^{[j]}(t,h), \ldots\right) \qquad \text{Equation 1}$$

where $\alpha(t,h)$ is a scaling function, t is the valuation time and h is the time horizon of the forecast (as seen from t) of external IM, $IM(t,h)$. In a simplest form, $\alpha(t,h) = \alpha(t,0)$ is a constant to align a known from an external source $IM(t,0)$, i.e., today's IM calculated by CCP or SIMM, with today's estimate derived by internal model, $f\left(PV_k^{[j]}(t,0), \Delta PV_{n,k}^{[j]}(t,0), \ldots\right) = f(t,0)$, which in some models is precisely $VaR(t,0)$

$$\alpha(t,0) = \frac{IM(t,0)}{f(t,0)} \qquad \text{Equation 2}$$

It is assumed that $\alpha(t,h)$ does not depend on *simulated* market state, hence from implementation perspective it is MC path-independent; $f\left(PV_k^{[j]}(t,h), \Delta PV_{n,k}^{[j]}(t,h), \ldots\right)$ represents an internal model component of the future initial margin state, which, as we discussed in the previous section, is an internal forecast of VaR or VaR-like measure at horizon h; f could be MC path-dependent and can depend on portfolio value, $PV_k^{[j]}(t)$, or/and its change over MPOR, $\Delta PV_{n,k}^{[j]}(t)$, or other variables such as underlying risk factors; j is a path in case of path-dependent model, n is a number of days in MPoR, k is a chosen percentile of the exposure distribution. Recall that we will refer to the path-dependent models as stochastic. For example, regression DIM and sensitivity DIM models discussed in the previous section are stochastic, whilst simple DIM model is not. In any given model only some of the underlying variables: $PV_k^{[j]}(t)$, $\Delta PV_{n,k}^{[j]}(t)$, etc., might be used.

Thus there are two important assumptions behind all DIM models: first, future initial margin could be decomposed into two different components; and second, one of the components is market independent, while the other is market dependent.

The first component is scaling part $\alpha(t,h)$ in Equation 1, the second is internal model part $\left(PV_k^{[j]}(t,h), \Delta PV_{n,k}^{[j]}(t,h), \ldots\right)$ in the same equation. The first component is used to align forecasted IM with known external at initial date and in theory should be market independent. It is a key model assumption. In a simpler case of a constant, this component should also be independent of the forecast horizon. Note that we omitted word 'scenario' when describing $\alpha(t,h)$ as 'market independent'. This is intentional. We will return to the market (in)dependence of $\alpha(t,h)$ in the next section.

The second component, internal model part f is clearly linked to the portfolio specifics and evolution, hence depends on market state (MC scenario).

The quality of the forecast of IM could be impacted by either component: scaling function or internal model. Ideally model assessment should consider verification of both components separately and their product, as it will make transparent sources of IM forecast error. The first component, a scaling function (or a constant) is assumed to be market independent. Then its

stability under different market conditions becomes a key model assumption. Another way to formulate this requirement is to observe that if this component is not impacted by the change of the market, it restricts how and why it could change with time. Thus verification or monitoring of the stability assumption could constitute important part of the model assessment. It will be discussed in the next sections.

The second component is already subject to some or even complete assessment. In case of simple DIM model (scaled VaR) (Wilkens & Moran, 2017), this component is equal to collateralised PFE at specific percentile and should be already backtested as part of standard backtesting procedure under the current regulations (see for example (Basel Committee on Banking Supervision, 2010)). Therefore no additional verification is required. Such backtesting also provides some verification for stochastic DIM, but it is not the full verification if the internal model part is path-dependent.

Backtesting of path-dependent forward VaR predictions is possible, for example using PIT (probability integral transform) based test along the lines described by (Anfuso, et al., 2017), however if scaling part is verified it may be sufficient and cost effective to conduct full backtest only at a cumulative level, i.e. verify the performance of the entire DIM model. Thus we propose in this article that to verify quality of all Dynamic IM models minimum of two steps are required: first, verification of key model assumption about stability of scaling part and second, backtesting of the forecasted initial margin against realised margin. The potential methodologies for these will be discussed in Section 6 and 7 correspondingly.

6. Scaling function analysis

6.1. Spot scaling behaviour and scaling as function

As discussed above, stability of scaling is a key assumption for all DIM models and thus requires monitoring. The simplest monitoring set up is for the case of constant scaling, while for scaling function the analysis of the behaviour of initial value $\alpha(t, 0)$ appears to be a minimal requirement (i.e. necessary, but not sufficient), however as we will explain below in practice it is likely provide most of the required information.

Let us discuss what types of the variability of scaling one could expect. We start with the most idealised case and move to more realistic ones. First case, the perfect one, is when internal model is identical to external model (e.g., SIMM or CPP's). In this case there is no need in any model adjustment: scaling function is constant and equal to one. Of course, as discussed in the previous sections this is very unlikely to happen. Second case is when internal model is able to capture all dynamics of external model but requires some scaling which is constant. Again, it is largely a theoretical case. On a first thought, it could reflect the case when internal model is able to capture distribution of external system but uses different percentile. In such case the

ratio will be stable as long as shape of the distribution is stable and this means, on second thought, this is also unrealistic. In more practical, but still a good model case, scaling could be constant (at least approximately, i.e., within some tolerance) for a constant maturity portfolio, with no trades allowed to be added or removed. In this case the scaling function will be portfolio dependant, but market independent, and it is probably the best case which could be at least approximated in practice. Let us assume that this is the case for IM forecast model and consider its implication for the model validation.

In such model $\alpha(t,0)$ may change gradually as maturity of portfolio changes, but will change abruptly if deals expire due to run-off thus *changing* the composition of portfolio. To make crude example, lets assume internal bank's CCR/VaR engine does not model implied volatility as a risk factor, whilst CCP does. We also assume that missing risk factor causes understatement of VaR. Then for the portfolio of futures and options, where futures dominate short term, whilst long term portfolio residual are options, scaling will jump once futures expire and only options remain. For the portfolios with larger discrepancies in the risk factor modelling between internal and external model one also would expect noisier $\alpha(t,0)$ due stronger market dependence, so in real-world use scaling stability will pass for some portfolios and fail for others.

In practice the difference in modelled risk factors between external and internal model could be more subtle, but nevertheless cause this type of step wise dependence of scaling on portfolio composition. The simple way to reflect portfolio run-off is with a scaling function $\alpha(t,h)$ as a piecewise constant at set of future horizons h. If the assumption of market independence holds then this piecewise function can be computed without moving portfolio forward, starting with full portfolio and then reducing it to sub-set of deals still present in the portfolio after selected horizon h. If run-off is gradual, this approach could be modified to include interpolation between different scaling levels.

Note that formally for full monitoring of scaling stability we should monitor stability of scaling both at initial point, $(t,0)$, and scaling at future times (t,h). However if the scaling is only portfolio dependent, the effect of adding $\alpha(t,h)$ to $\alpha(t,0)$ stability monitoring only adds different portfolios under different market conditions. So in practice, if the selection of portfolios is comprehensive enough (for example, covering usual composition changes due to run-off) and monitoring is done over sufficiently long time, monitoring of only $\alpha(t,0)$ should provide the same information as a full monitoring. In our basic example of mixed futures and options, this means creating additional artificial portfolios consisting of options only and running them for a sufficiently long time while monitoring $\alpha(t,0)$.

In this section we always assumed $\alpha(t,h)$ being market independent, but allowed a caveat in the previous section. We shall address it now. In Section 5 we separated DIM forecast into $f\left(PV_k^{[j]}(t)\,,\Delta PV_{n,k}^{[j]}(t)\right)$ which models VaR/IM evolution internally and takes account of market evolution and scaling $\alpha(t,h)$ which then adjust this internal VaR to external VaR/IM forecast. Any 'knowledge' of internal model about market evolution is *by definition* inside f.

However (Anfuso, et al., 2017) and (Caspers, et al., 2017) introduce additional scaling to $\alpha(t,h)$ to compensate for known shortcomings of internal model f. For example, to reflect term structure of risk factor volatilities or mean reverting from high/low volatility regime or to introduce conservatism for longer term projections. This is achieved by a multiplier linking spot scaling to long term scaling, $\alpha(t,h) = \alpha(t,0)b(t,h)$, with

$$b(t,h) = \left(\frac{a^{\infty}}{\alpha(t,0)} + \left(1 - \frac{a^{\infty}}{\alpha(t,0)}\right)e^{-\beta(h)\times h}\right) \qquad \text{Equation 3}$$

where a^{∞} stands for a conservative long term scaling value, $\alpha(t,0)$ is a scaling at initial point t, and function $\beta(h)$ determines transition over time. Within our framework this $b(t,h)$ should be part of internal model f. However (Anfuso, et al., 2017) and (Caspers, et al., 2017) incorporate it into $\alpha(t,h)$. Generally speaking, any type of correction which introduce market dependence, make stability of $\alpha(t,0)$ at time zero a weaker proof of stability at h. Thus there is a clear argument to make $b(t,h)$ part of f. However if internal model already exists and is used as CCR engine, and if CCR engine is already backtested as part of Basel requirements, then keeping $b(t,h)$ out allows bank to re-use existing backtesting as part of DIM verification. Of course, the argument could be reversed as this would mean that bank has a methodology to have better CCR forecasts via $b(t,h)$ correction but is not using it for non-IM CCR forecasts. We would thus argue improving f and then backtesting with $b(t,h)$. In practice, the argument is academic as we are not aware of any reports of this type of correction being used in the industry. For completeness we should also mention the possibility of introducing an additional scaling to compensate not the known shortcomings of internal model f, but methodological differences between internal model and external one (with both as per earlier discussion fit-for-purpose and overall on par in their predictive quality). In this case, leaving existing backtesting framework of f unaffected, it may be easier to introduce such scaling into scaling $\alpha(t,h)$, and if the correction is market dependent, then alpha will be market dependent. However, again, we are not aware of any discussion of such corrections.

If a correction to compensate known methodological differences between internal and external model is impractical, there is still a way to reflect the impact of two models' interaction, if $\alpha(t,h)$, dependence on time is introduced. Even if scaling (for a given portfolio) is fully market independent, it is still affected by numerical and other 'noise': we have shown in Section 2 that $\alpha(t,0)$, could vary significantly over different time scales. The impact of this 'noise' theoretically could be reduced by setting scaling function to converge to long time mean a^{∞} with $\alpha(t,h)$, calculated re-using Equation 3 for this new purpose. Of course, such correction would introduce an extra requirement to define methodology and then verify accuracy of long-term scaling parameters: α^{∞} and $\beta(h)$. Given the difference between VaR projections by different models could be driven by a variety of causes as we already discussed, the reliable correction of such type may not be easy to achieve, although if a monitoring procedure is established along the lines discussed in the next section a possible α^{∞} estimation will be its by-product. Finally we note that if $\alpha^{\infty}$is derived, by further scaling it down (for received IM)

or up (for posted IM) a desired degree of conservatism could be injected into longer term IM forecast.

6.2. Initial scaling assessment. Impact of the quality of IM model replication

In this section we propose a practical way of monitoring the initial scaling, ratio $\alpha(t,0)$. First we briefly recap previous discussions: Sections 4 and 5 show that validity of DIM modelling require stable, ideally constant ratio per portfolio. Sections 2 demonstrated that this is not achievable, and thus the actual requirement for $\alpha(t,0)$, is to stay within a certain range. The acceptable range is for the bank's model owners and risk management to determine, however based on results presented in Sections 2 and 3, we suggest 30% as a reasonably strict requirement (the requirements may be relaxed at the time of extreme market turmoil), so we should require that 'most of the time' the ratio should not deviate more than 30% over sufficient, but practical of length observation, say, one year. (There is an important exception to 30% being "reasonably strict", which we will discuss at the end of this section.) In line with common industry practice it is useful to set up traffic light zones for variability of the ratio. The key zones are green zone (model passes) and red zone (model fails), each defined by a corresponding boundary. The outcomes below green zone boundary suggest that model passes and outcomes above red zone boundary suggest that model fails. In this case 30% would define the red zone boundary[19]. For brevity, we will refer to red zone boundary as simply red boundary and correspondingly green zone boundary will be referred to as green boundary.

To monitor this type of requirement quants have a variety of tools. We suggest one of the simplest. We can assume that daily observations of spot scaling ratio $\alpha(t,0)$ for a given portfolio is daily observation of a stochastic process. In this case Figure 5, Section 6.3 plots this process' evolution. The 30% requirement translates into a coefficient of variation (CV) of this process (or a relative standard deviation) ν, $\nu(\alpha) = \sigma(\alpha)/\mu(\alpha)$, not exceeding 30%, where σ and μ are its standard deviation and mean. If 30% is the red boundary, then it is convenient to set up 10% for the green boundary[20]. Assuming the system is classified as green in its performance, the probability to observe initial scaling above red boundary, i.e., above 30% error is equivalent to probability to observe values outside 3 standard deviation range. In case of normal distribution it equals to 0.27%, and even in the most general case for any distribution

[19] Of three illustrations in Section 2, second and third ones are assets and instrument specific (equity and linear instruments) and only the first one is true for any asset; the discussion in Sections 3 is relevant for any portfolio. Other practitioners may wish to expand coverage of our equity specific study, however given approximations of IM as VaR, and expected differences for IM model vs. IM predicting model we do not see the requirement of 30% as red boundary tightened further, even if, say, third illustration alone is to suggest lower bound for another asset class or instrument.

[20] It is up to each institution's preference whether there should be three zones in total (green, yellow, red) or more. If four zones are introduced with yellow and amber zones in between green and red ones, then 20% could be set as the boundary between yellow and amber zones providing intuitive set up with equally spaced boundaries: 10%, 20%, 30%.

it is less than 11% (according to the Chebyshev's inequality). Additionally the results from monitoring of initial scaling ratio for live portfolios could be used for day-to-day performance assessment to keep track on production environment changes and technical issues (missed feeds etc.). We will further discuss this in the next section.

Counterparty risk management tends to be concerned with both average risk (via EEPE and regulatory capital provision) and more extreme risks (via PFE and limit management). Thus separate monitoring of very large errors (which means days when forecast clearly failed) may be desirable and treating $\alpha(t,0)$ as a process presents quants with variety of tools for this purpose as well. For example, additional measure could be added by looking how 'fat' are tails of $\alpha(t,0)$ distribution are (for example, compared to the normal distribution) and setting up an alert correspondingly[21].

Treating $\alpha(t,0)$ as a process allows us to make two further observations. Firstly, we note it makes filtration of the jumps a fairly standard quantitative task, see for example (Chorniy & Greenberg, 2015) for discussion of jump filtration. Why such filtration may be desirable is discussed in the next section. Secondly, if we monitor changes of $\alpha(t,0)$ by setting limit to the ratio of volatility and mean of the observations, the reasonable question to ask is whether the mean of these observations has a meaning (no pun intended). Since we can view stochastic process behaviour as a sum, $\alpha(t,0) = \mathrm{c} + \varepsilon$, of deterministic part, c, and a random noise, ε, then c is 'true', long term value, which in Equation 3 was denoted as α^{∞} at the end of discussion in Section 6.1. In this case the mean is an estimate of α^{∞}. Of course, the use of this interpretation and the use of Equation 3 would mean that the influence of run-off (cf. Section 6.1) and the other effects of $\alpha(t,0)$ dependence on portfolio composition is negligible. Run-off effects could be accounted for by combining a set of long term values with piecewise function for $\alpha(t,h)$ as discussed in Section 6.1.

Now for the rest of this section we return to discuss the case when 30% as CV red boundary is not "reasonably strict". The 30% or a similar number is predicated on the assumption that one fit-for-purpose VaR model forecasts the value of another fit-for-purpose VaR model, with both model sufficiently different, as full replication of IM methodology is not practical. However if the bank's IM forecast uses a very close, but not identical replication of actual IM model (SIMM or a particular CCP), for example, perfect replication under most conditions and very close numbers under some other conditions, then 30% difference between the IM forecast and realisation and correspondently 30% CV of $\alpha(t,0)$ will point to a model or infrastructure failure. In this case the CV bounds should be tighter. It is possible to derive an estimate of these bounds by using a different interpretation of $\alpha(t,0)$, this is done in Appendix.

In summary, Appendix introduces a 'correlation' between models as a measure reflecting the degree of replication, and then uses it to develop a formalism from which a practical rule of

[21] Of course these are illustrations and the boundaries should be calibrated by each institution to account for inherent specifics embedded into their models and risk management processes.

thumb could be defined as follows. Classify models into three wide groups: 'generic', 'close replication', and 'near perfect replication', choose red boundary for the 'generic' based on main discussion in this paper (for example, 30%), and then convert it using 0.9 and 0.99 correlations for 'close replication' and 'near perfect replication' models. Table 1 shows such conversion of 30% 'generic' red boundary and 10% 'generic' green boundary. Also we note that the same formalism allows alternative treatment of uncertainty from Sections 2 and 3, but still results in similar recommendation of about 30% as a red boundary for generic case.

	Generic	Close replication models	Near perfect replication models
Green	10.0%	5%	1.5%
Red	30.0%	15.0%	4.5%

Table 1: Green and red boundary depending on the degree of model replication

The results of Appendix also could be viewed from an opposite angle. Instead of asking when 30% as CV red boundary is not "reasonably strict", bank risk management or trading may postulate what boundaries it is willing to tolerate and from that derive degree of replication required. For example, a bank may accept 30% boundary for most CCPs/clients, but require tighter boundary in some cases, or choose one level of modelling precision for PFE and capital calculation in risk management (IMM) engine and another, possibly tighter, for pricing, i.e., CVA/MVA calculation in pricing engine.

6.3. Initial scaling monitoring

Although we recommend monitoring of scaling stability as part of IM model verification, the ultimate test of IM forecasting model is backtesting of IM discussed in Section 7. This means that $\alpha(t, 0)$ monitoring can be supplementary, narrower in its focus as we will discuss in this section. We also note, to be discussed shortly, that treating $\alpha(t, 0)$ as a stochastic process, which is one of many possible approaches (compare for example with backtesting of IM, Section 7), is well suited for this narrower task.

To understand performance of IM model it is useful to know how well it performs as a forecast of another model's VaR in a narrow sense: excluding the impact of re-calibration. The sudden change of ratio could be driven by (regular) re-calibration banks' CCR engine. If $\alpha(t, 0)$ is viewed a stochastic process this will manifest itself as a jump, which can filtered with a standard tool kit. The jumps also could be caused by changes in the portfolio composition, e.g. trade unwinds. However the unwind caused jumps may not be the result of the methodological

difference between internal and external VaR model, as we previously discussed. They also can be triggered by the other part of CCP IM methodology, for example, by change of VaR and Stress VaR relations after deal(s) expiry (recall that CCP IM include other components in addition to VaR estimate). Thus depending on the aim of monitoring, the analysis of $\alpha(t,0)$ dynamic may be done with raw data or with jumps filtered out, thus for example, (approximately) removing impact of internal model recalibration, trade cancellations or non-VaR components in CCP IM methodology. We will further discuss such filtering in Section 7 in connection with backtesting of IM forecasts.

The detailed discussion of actual behaviour of spot scaling $\alpha(t,0)$ lies outside of this article. The generic (unfiltered) results may look like Figure 5 (these were obtained by forecasting IM for fixed income and equity vanilla instrument for using MC VaR model as per Section 2 vs. SIMM and Eurex for the period of 2018), both displaying gradual change, noise and jumps; the latter can be filtered out in stability analysis as discussed. As we noted in the previous section monitoring of $\alpha(t,0)$ for live portfolios could be used for day-to-day performance assessment to keep track on production environment changes and technical issues, for example, missed feeds, etc. These are likely to manifest themselves as jumps.

Interestingly, in the example of Figure 5, the plot also shows scaling levels as typically above one (in 'normal' market conditions as one currently views 2018); this will provide an illustration for the next section.

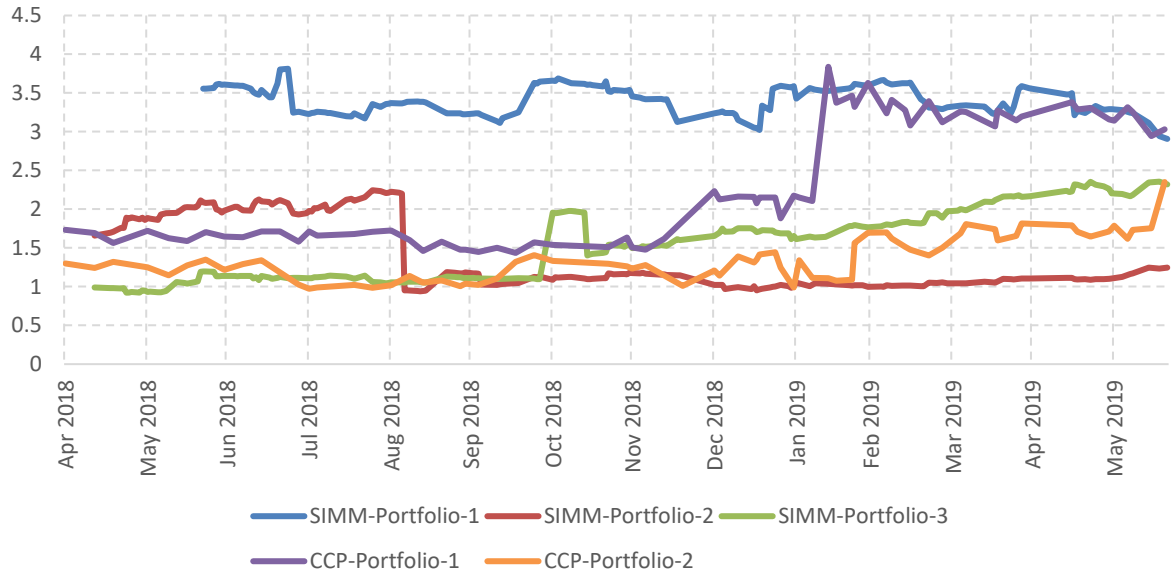

Figure 5: Typical examples of initial scaling variability

6.4. Implication of scaling monitoring on CCR backtesting framework

Once the monitoring of scaling is set up as discussed above and long enough observation are obtained, the results are likely to have direct implication on the regulatory backtesting, specifically whether IM backtesting is a regulatory requirement.

ECB guide to internal models (European Central Bank, 2019) Section 9 (77) states that "if direct back-testing of the exposure of margined netting sets is not feasible, institutions should have a separate validation of the margining process, of collateral value changes and of netting set market value changes over the relevant time horizons."

Once full set of $\alpha(t,0)$ monitoring results is available, it may lead to the clear conclusion whether netting set backtesting of exposure with IM is statistically feasible. If typical scaling factor is above one[22], and f, an internal forecast of VaR, is done at 99th percentile (which is industry standard), then PFE of portfolio is non-zero only if calculated at above 99th percentile. For example, majority of ratios' shown on Figure 5 is above 1.25, which correspond to 99.8th percentile (assuming normal distribution). The number of observations over a few years typical for banks' CCR backtesting framework is not sufficient to backtest such percentile prediction even over the shortest meaningful horizon (margin period of risk of 5 days). In case of market risk IRC[23] forecasts (VaR-like measure at 99.9th percentile) backtesting is not required by regulators. If backtesting of exposure at netting set is not possible, backtesting of components: exposure without IM and IM itself, becomes a regulatory requirement. Of course, even if backtesting of total exposure was possible, it is always a good practice to backtest components to be able quickly identify the source of model weakness.

7. Backtesting of IM forecasts

As we concluded in Section 5 to verify the performance of DIM models two steps are recommended: 1) monitoring of scaling stability and 2) backtesting of the initial margin forecast against the realised margin. Previous section has discussed the stability, this section addresses the backtesting.

Firstly, we need to clarify basic definitions. In Section 2 we discussed that IM forecast is not a forecast of forward VaR. It can be treated as such only approximately, as external IM requirement by CCP or SIMM could be approximated (and is assumed by most IM forecasting models) as VaR estimator, which itself is a PnL forecast. Then we discussed consequences of the forecast of forward VaR by one model predicting another model's spot VaR estimator, with

[22] Given that CCPs IM is 'more than' VaR and SIMM also have significant degree of conservatism, the ratio larger than one is not surprising.

[23] IRC, incremental risk charge, is VaR-like measure for credit instruments in market risk regulatory framework.

two models possibly unrelated. At this stage we need to step away from our discussion how external IM is generated. From banks' point of view the external IM requirement is given, it is a spot value. Then bank's IM forecast is the forecast of this spot value in the future. This makes IM model forecast a *point* forecast. General discussion of point forecasts could be found in (Gneiting, 2011). A relevant comparison from everyday life is a temperature forecast. The temperature reflects a complex phenomenon in a single number and thus is forecasted as a single number. The temperature parallel will also become useful later in this section where we will discuss that in general form IM models produce *conditional* point forecasts.

Thus to verify IM model we need to develop a backtesting methodology for a point forecast. This places IM backtesting outside of existing banks' backtesting framework, which was built to verify market and CCR models. This point is significant and appears to be underappreciated by the industry. To clarify the last statement, let us look at the broader picture. In general, forecasts fall under one of the three types: point forecasts, interval forecasts and probability density forecasts (probability forecasts are sometime viewed as a separate, fourth, type). The interval forecasts include sub-type specifically relevant to risk management: tail forecasts (one-sided interval). The interval (tail) forecasts and probability density forecasts are the types produced by market and CCR models: specifically, VAR (market risk) and PFE (counterparty risk) are tail forecasts. The EE and related measures (e.g., EEPE) reflect the whole distribution and therefore are directly related (function of) to the density forecasts. ES (expected shortfall) reflects the density of the tail of the distribution, hence it has links to both tail and density forecasts. The existing banks' backtesting framework for all these risk measures rely on their nature, so the existing framework is not always directly transferable to point forecasts[24].

Of course, ES and all other single value risk measures above could also be viewed as point forecasts. However this angle had little practical influence on industry backtesting with perhaps one exception. The review (Gneiting, 2011), which we already mentioned, pointed out that ES is not an elicitable risk measure (whilst VaR is). This led to an intense, but relatively short-lived debate centred around the concern that non-elicitable risk measure cannot be backtested. However, firstly, it was reasonably quickly realised that from the view of the practical and regulatory task of backtesting it is largely irrelevant, or at least misleading, as backtesting is tasked to determine whether a given model is fit-to-purpose and not to compare (or rank) a few competing models. Secondly, from more theoretical angle it was realised that concepts of elicitability and backtestability has certain flexibility and could be extended to other risk measures, specifically ES. Please see (Chorniy, TBA) for an overview of the elicitability debate and for the detailed discussion of market and CCR backtesting framework focusing on interval forecasts and probability density forecasts. However we note that later, in Section 8,

[24] It should be noted that in practice point forecasts are plentiful and a fundamental part of any internal CCR engine, which usually involves Monte Carlo method and stochastic processes. The parameters of such processes, e.g. drift and volatility, estimated by some calibration methodology are point forecasts of their future realisations. The accuracy of those forecasts (calibration parameters) are verified within the banks' frameworks, however usually not via backtesting approaches employed in market risk and CCR, which, as we discussed, target stochastic measures.

we will return to the elicitability related topic in the context of multiple VaR models used by banks and CCPs, and based on that discussion we will propose a novel verification approach.

Let us go back to the task of defining a methodology for the backtesting of DIM models. It is useful to start our discussion with a basic case of simple (scaled VaR) DIM (type 2 in Section 4). In addition to providing convenient illustration to the proposed backtesting approach, it seems to reflect a model type expected to dominate CCR modelling, judging by informal discussions at industry forums. xVA calculation may require more advanced path dependent models (i.e. types 1 or 3, in Section 4). In the future a preference for path dependent models might be also informed by the needs of liquidity modelling. Some advanced CCR and XVA engines can model all individual cash flows, not only MtMs. This means engine can stochastically model all settlement and pre-settlement cash flows, which is then can be fed to bank-wide stochastic liquidity model. If this capability is deployed IM posting and receiving is another set of cash flows which need to be modelled per scenario.

7.1. Backtesting of simple DIM

Simple DIM produces a single IM forecast per time horizon, which as we have already noted is a point forecast. Generally speaking for a good model the forecast values should be 'close enough' to realisation, so to ascertain forecast quality we need to select a scoring function (as for example listed in (Gneiting, 2011) and define the acceptability bounds. The actual choice is guided by the risk management priorities of the institution. For example, at a given forecast horizon an approach used in Section 6.2 for initial scaling assessment could be recycled, however such scoring function treats over and understatement of IM equally[25]. This may be preferred for general assessment of the model or for xVA purposes, however for the risk management main concerns are directional: underestimating exposure is the danger to a bank. In case of OTC bilateral trading or CCP client leg it is an overestimation of IM, in case of CCP leg (assuming posted IM is not bankruptcy remote), it is an underestimation of IM. As the error of concern is always one sided, the sum of daily relative errors, which has a sign, provides an example of a simple and transparent scoring function[26]: $Err(t,h) = \frac{IM_{For}(t,h) - IM_{Re}(t+h)}{IM_{Re}(t+h)}$, where $Err(t,h)$ is relative error and $IM_{For}(t,h)$ is forecasted IM for a particular portfolio at time t and forecast horizon h, and $IM_{Re}(t+h)$ is realised IM corresponding to the forecast at time t for horizon h, i.e. realisation at $t + h$.

The use of relative error as scoring function allows the same choice of test bounds as for scaling monitoring: based on the results presented in Sections 2 and 3, we suggest 30% as a reasonably

[25] An exception count, a main staple of backtesting, if used for such point forecast, will result in even simpler test simply detecting the presence of bias, but not the size of the error.

[26] A bank may choose another, one sided measure, and this choice may depend both on whether additional measures are chosen as we discuss in the next paragraph and on the degree of sophistication of the alpha monitoring.

strict requirement as a red boundary for average of relative error over a reasonable period of observation (say, one year, to match typical regulatory and economic capital horizons). This measure also retains the link to regulatory guidance as discussed in Section 3. In line with discussion in Section 6.2, in a case of close IM replication, risk management may wish to adjust the bounds accordingly.

As we already discussed in Section 6.2 a good CCR model should perform well for expectation and middle of the exposure distribution, as well as for high percentiles. Usually banks' CCR backtesting framework aims to verify both, specifics depending on banks framework, as it could be done through common or separate tests, see (Chorniy, TBA) for a general review of the approaches. If the average of the relative error is chosen, which reflects predominantly the centre of error distribution, a bank may choose to have a separate measure to target large (extreme) errors of IM forecasts or the shape of error distribution. This could be set to complement a measure monitoring very large errors of $\alpha(t, 0)$ distribution, if one was chosen. For example, if scaling monitoring concentrates on frequency (fat tails), the backtesting of IM could include size, if one measure is two-sided, another can concentrate on the underestimation of exposure, and so on. Measures to evaluate the quality of the IM forecast and to monitor the scaling stability could be designed as complimentary in the other respects as well. For example, if scaling monitoring incorporates filtering of jumps, as discussed in Section 6.3, the primary result of backtesting of IM margin should be unfiltered.

7.2. Backtesting of stochastic DIM

Now we turn to the case of stochastic DIM. Stochastic DIM models produce an individual forecast for each MC path, hence entire forecast is a distribution. This allows two interpretations and correspondingly two backtesting approaches.

First approach is to view stochastic DIM model as producing a *conditional* point forecast, i.e. the point forecast *conditional* on the realisation of particular future market. The example of a temperature forecast as a point forecast provides a useful illustration. If a weather model predicts future temperature in a single location, then, obviously, its forecasts should be cross checked against the temperature measured at this location. However if the model generates temperature forecast for a random location each time, then to verify the model, the temperature will have to be measured at the location corresponding to the specific forecast. In case of stochastic DIM the realisation should be compared to the forecast corresponding to the observed market state, i.e. stochastic DIM is a point forecast conditional on the market realisation.

In theory this means the need to choose MC scenario identical to the market universe observed[27], which given MC is the set of discrete scenarios, generally is not possible. In

[27] For path-dependent portfolios this also includes matching the universe's history.

practice, we can only hope to find a scenario 'suitably' close. Again, returning to our example of temperature, it may be possible to roughly verify forecasts for London with an observer located in its suburb, but not with the one located in Paris. Thus for backtesting of conditional point forecast to be used in practice a method needs to be devised to take into account the difference between predicted and realised markets. As far as we are aware such backtesting methods do not exist, so if backtesting conditional point forecast to be used, new research is required[28].

The second approach treats forecasted IM as a synthetic distribution forecast: it is a joint forecast of DIM *and* market evolution together. By definition, in path-dependant models IM is known at each moment in time on each MC path. Therefore we could build a distribution of the forecasted IM for different horizons. Once such distribution is obtained it becomes a standard task for backtesting. For example, exception counting or PIT could be used, see (Chorniy, TBA) for general review, with specific example given in (Anfuso, et al., 2017). Such backtesting examines *jointly* CCR model's risk factor generation and accuracy of IM forecast. This approach is already widely used by the industry for backtesting of other objects.

The regulators request backtesting of IMM CCR engine components, see for example (European Central Bank, 2019) and typical framework will have at least three layers. First layer backtests risk factor generation. Second and third layers backtest individual pricers[29] and portfolio level exposure correspondingly. Both involve synthetic distribution forecast: it is a joint forecast of market evolution and either pricing valuation or portfolio exposure. Backtesting IM by the same method will mean simply replacing a pricer or a portfolio with IM.

This approach is well known both to practitioners and regulators, and ready to be used, whilst approach based on conditional point forecast still requires further research, so at the first glance the choice between two backtesting approaches is clear: second approach should be used. Then the recommendation will be to use monitoring of scaling both for simple and stochastic DIM methodologies, but for simple DIM to use point forecast backtesting (e.g., relative error measure), and use synthetic distribution forecast for stochastic DIM. If stochastic DIM backtesting to be implemented immediately, then banks probably do not have a choice. However in the long run we believe a closer look at both approaches is warranted.

Let us examine them closer. We start with the first approach, conditional point forecast backtesting. It has benefit that it tests IM model only, without impact of the other components of CCR model, so it has certain mathematical purity, however so far this did not attract industry to it. More importantly is that with this approach the acceptable error and backtesting pass/fail boundaries could be expressed in terms of relative error or its equivalent. This is very

[28] One possible way to proceed would be first to define a measure of the distance between the observed and any given forecasted market universe (MC scenario), and then to define the backtesting method taking this distance into account, in simplest case, by defining maximum acceptable distance between chosen forecast scenario and the realisation. We suggest that signatures which currently attract attention within the industry may prove a fruitful tool to deploy.

[29] Pricer: pricing function which derives the value of an instrument for a given market.

transparent measure, which is easy to communicate to risk managers and senior management. In the current climate, where model risk management (MRM) takes key role, it is an important feature.

The second approach has pass/fail calibration based on setting (low) probability of type I error (correct model erroneously rejected), recall Basel or FRTB mandated VaR backtesting colour zones (Basel Committee on Banking Supervision, 1996), which in turn causes the backtesting to reject incorrect model with reasonable probability. The backtesting results based on this definition is in practice harder to translate into MRM message to risk practitioners and senior management. That said, we notice that so far this approach remains to be the only one used by the industry. However in case of IM, we believe, there are additional considerations.

To illustrate them it will help to compare IM and pricer backtesting. Pricers used for trading are required to be precise (at least for regular traded instruments), validated by a model validation team and usually are not subject to backtesting verification, while pricers in a IMM CCR engine can be more approximate and are usually monitored by backtesting, which uses the second approach. The pass/fail criteria (boundaries) for pricers' backtesting do not depends on the portfolio or counterparty. However in case of IM, as we already discussed, banks may choose different degree of precision and correspondingly degree of replication for different CCPs. This means different backtesting boundaries depending on a counterparty. Also IM forecasts used in CVA/MVA calculation may require much higher precision, then IM forecasts built into IMM CCR engine. If IM backtesting is to be used to verify front office engine it will create yet another set of backtesting boundaries.

In this case to retain the second approach banks will need to translate these different relative error requirements into different probabilities of type 1 error to be used for calibration. As it stands there is no methodology for this. Thus if banks want to have flexibility of the varied level of precision, they should invest in research to develop the conversion of relative error boundary into type 1 probability-based boundary. However, the alternative is to invest in research into the first approach, backtesting of conditional point forecasts. We believe the investment in conditional point forecast method may be warranted as in a long run it may be more compatible with transparent model risk management.

Thus in our view both conditional and synthetic backtesting have a certain difficulty in deriving an answer for stochastic DIM. An alternative way to think about it, is that the difficulty lies not with the answer, but with the question which should be rephrased and this we will address in the next section.

8. Comparative approach to verification of stochastic DIM models. Elicitability.

Let us step back. If we want to model the degree of protection IM provides over period of trading activity, not the actual value of the exposure on a selected day (MC time horizon in CCR engine), we need to know how well IM model of CCP or SIMM is suited to its 'task'. The 'task' could be fully known or possibly known only approximately from CCP public information (for the rest of this section we will refer to CCP IM only and omit SIMM for brevity). The CCP IM's 'task' might be to provide a good VaR model at 99th or some higher percentile, or use VaR as only a floor in a larger model offering protection better than VaR alone. In practice it is the *realised* quality of CCP IM which is important: its degree of accuracy or conservatism. If internal, CCR engine based, IM model is built to the same quality, i.e. to the same accuracy or conservatism, then this internal model used in CCR engine should show the level of risk the bank is running through the time as if external model was exactly replicated. The question how accurate one model predicting another is replaced with a new question: is one model is as 'good' as another. In simple words, this new approach to initial margin assessment is to demand from both models to be either equally 'good' or 'bad'. We promised in Section 7 to return to the concept of elicitability, and, indeed, the relationship of this comparative approach to elicitability and a broader topic of comparative backtesting is evident.

To answer this new question analytically one needs to define what is 'good' and then to define scoring function which reflects how 'good' the model is, and finally to require that the scoring value of both IM models to be very close. Obviously the definition what constitutes a good[30] model (and correspondingly level of risk we attempt to measure) is not unique and thus the definition of scoring function should be a subject of discussion. Even the method of computing a chosen scoring is a subject of discussion. For example, let us assume that bank internal IM methodology achieved a high degree of external IM replication or generally designed in such a way that no scaling is necessary, i.e. $\alpha(t,h)$ is expected to be one for all portfolios. In this case the comparison could be done between immediate forecast of internal IM and CCP IM. However if scaling is used then at time zero both forecasts are matched by definition, thus the score of CCP IM treated as time zero forecast need to be compared with score of forward forecast of internal IM model. We assume, just for argument's sake, that we can create a score that can handle both point forecast of CCP IM and a distribution forecast of stochastic DIM on equal footing. Later we will present a possible example. In this case the score for forward forecast is in the terms of the previous section synthetic: it includes the quality of the MC simulation of the risk factors. In theory this may affect the score. As an extreme example, if the risk factor simulation is poor representation of the real world and none of MC scenarios even resemble real world realisation, then IM model fully replicating CCP's IM model may not match CCP IM score. In practice bank's model validation and backtesting ensure that MC

[30] For example, VAR models may be valued for accuracy, but IM models may be more valued for conservativeness.

simulations are reasonable representation of real world and such argument is largely theoretical, with possible exception of markets during extreme crisis.

Thus if industry were to move to comparative approach there should be wider discussion both of theoretical aspects and, we should not forget, of practical, infrastructure aspects. Hence we can only introduce this topic, full investigation clearly lies outside of this article. However, we believe it may be useful to present a simple example of a potential approach.

Let us consider IM as a tail measure of exposure distribution, which could be a percentile (VaR) or expected shortfall (ES), while the exact detail (e.g., percentile) of such measure might be unknown. This assumption is sufficient to create a link between realised and forecasted initial margin as both are an exposure tail measures, and all we require is them to be equally 'good' tail measures. It is interesting to note that this makes rather relevant the previous academic and industry discussion of elicitability and backtesting, specifically backtesting of ES, for overview see (Chorniy, TBA) and references therein. In our simple example we build our scoring function using binary indicator, akin to exception counting, i.e. the size of the error is ignored. We assume IM model where scaling is used, so at time zero both forecasts are matched by definition, thus the score of CCP IM at time zero need to be compared with score of forward forecast of internal IM model. Thus we compare observed (or modelled) PnL during the margin period of risk with regards to both realised and forecasted initial margins. This method is similar to comparing two different VaR methodologies and determining their relative conservativeness.

Following notation introduced in Section 7.1 we define for portfolio k forecasted IM at time t and forecast horizon h as $IM^k_{For}(t,h)$ and a realised IM at time $t+h$ as $IM^k_{Re}(t+h)$. We denote p as a margin period of risk (MPoR) and for portfolio k we denote a realised portfolio profit and loss over MPoR as $PnL^k(t+h,p)$ i.e. its MtM move between time points $t+h$ and $t+h+p$.

In order to assess if realised and forecasted IM have the same properties, we count how often they are greater than realised PnL for portfolio k. In case of forecasted IM, $IM^{k,j}_{For}(t,h)$, the count $\varphi^{k,j}_{For}$ is done within each MC scenario j, with realised PnL, $PnL^{k,j}(t+h,p)$, is also calculated on scenario j, i.e. *within* the CCR model, and given by Equation 4. Equation 5 below presents the count, φ^{k}_{Re}, for realised IM of portfolio k which compared with actual realised PnL:

$$\varphi^{k,j}_{For} = \sum_{h\,\epsilon\,H} \mathbb{1}_{\{\,IM^{k,j}_{For}(t,h) > PnL^{k,j}(t+h,p),\}} \qquad \text{Equation 4}$$

$$\varphi^{k}_{Re} = \sum_{h\,\epsilon\,H} \mathbb{1}_{\{IM^{k}_{Re}(t+h)\,) > PnL^{k}(t+h,p)\}} \qquad \text{Equation 5}$$

where H is a set of selected horizons (for example, set of all horizons up to 1 year). Note as the selected horizons are the same for both counts, Equations 4 and 5, it is not significant whether (some) forecasts are overlapping.

This approach corresponds to one of the two main flavours of backtesting, it uses a single point of origin of a forecast with multiple forecast horizons. This flavour is used in CCR model backtesting, see (Chorniy, TBA) for detailed discussion of different backtesting approaches. A more common flavour: single horizon, multiple points of origin, used both in CCR and market risk (e.g., regulatory mandated VaR backtesting), could be utilised as well. In this case forecast at time t for a single horizon h is repeated for the whole period of observation at each forecast point, for example daily (note, overlapping observations are allowed) and the sum in Equation 5 will be not across horizons h, but across all forecast points covering backtested period. Note that each indicator is a single number. Once calculated across all portfolios we obtain a set of distributions of φ_{For}^{j} (one per each MC path) and one distribution for φ_{Re}. For this distribution to be informative the set of portfolios k should be large and well diversified. This requirement could be eased somewhat if observations are sufficiently long, for example three years, as each year can be treated as one of three one-year portfolios.

At this step, once we obtained the distributions of φ_{For}^{j} and φ_{Re}, we require them to be close enough: for our dynamic IM model to be equally 'good' across MC scenarios *and* to have the same properties (to be equally 'good') as the real one. If satisfied, this means that both models are indistinguishable estimators of the PnL distribution tails (neglecting the difference between possible evolutions of markets in real world vs. MC world of CCR engine). Similarity of the distributions between MC scenarios created for the full set of portfolios and also with the real world realised distribution could be assessed by the standard statistical tests like Kolmogorov–Smirnov or other[31]. Interestingly, if forecasted IM is found 'good' only for some MC scenarios j, this may point to the limitations of internal IM model, i.e. under which market scenarios it is likely to fail, something which alpha monitoring alone may take a very long time to reveal.

Let us briefly comment on the point at the end of previous section about the transparency of alternative backtesting approaches for practitioners. Although this comparative approach does not restore the transparency of relative error based monitoring, we suggest that it could be made user friendly if scoring functions are chosen as practical performance measures of IM to be easily understood by practitioners.

Finally we note, that employment of such comparative backtesting procedure and close attention to the variety of dynamic IM models in general, may have additional benefits. It will provide banks and regulators with both toolset and incentives to compare day-to-day performance of VaR/IM models, which will cover not only comparison of VaR methodologies, but production issues and stress period definitions. Perhaps finally the time has come for elicitability-like discussion within the industry.

[31] The choice of a test is also an obvious area of potential future research.

9. Conclusions

In the review part of this paper we have shown that current industry effort in devising models for IM forecasting (except full replication) is based on ability to forecast VaR in the future. As a logical consequence, the proposed model verification (backtesting) of IM forecasting often targets the accuracy of forward VaR forecasts. The industry assumption is that once a good model to forecast forward VaR is built, the job of forecasting of IM is complete. We show that the ability to forecast forward VaR is an important part, but only a part of IM forecasting. Moreover forward VaR backtesting does not verify IM forecasting model, for example, a perfect VaR forecasting model could be (at least in theory) a failure as IM forecasting model.

Initially, we show that IM is not VaR, VaR being only its approximation. Then we show that even with 'IM is VaR' assumption, the forecast of IM is *forecast of a forecast*, which is principally different from 'just' forward VaR forecasting. The difference is both in the methodology of the forecast and in its type. Firstly, the forecast of forward VaR by one model is used to predict another model's VaR forecast, with two models possibly based on different methodologies. Secondly, as IM model produces forecast of a forecast its output is a point forecast (conditional or unconditional). The existing banks' backtesting framework for risk measures such as VaR or PFE relies on their distributional nature, so the existing framework is not directly transferable to point forecasts. Thus we devise and propose a suitable verification framework. We also demonstrate the fundamental limitations of IM forecasting which inform the proposed verification and can recognise the degree of replication of CCP or SIMM IM methodology by the bank's IM model. Finally, as a possible component of the proposed verification framework we introduce elicitability based approach for path dependent IM forecasts.

10. Acknowledgments

We are grateful to our colleagues Ziad Fares and Xavier Lorentz for the discussion of the case when scaling is driven by two correlated stochastic processes and for providing us with the expression for its variability, which we quote.

11. References

Andersen, L. B., Pykhtin, M. & Sokol, A., 2017. *Credit Exposure in the Presence of Initial Margin.* [Online]
Available at: https://ssrn.com/abstract=2806156

Anfuso, F., Aziz, D., Giltinan, P. & Loukopoulos, K., 2017. A sound modelling and backtesting framework for forecasting initial margin requirements. *Risk,* May.pp. 86-91.

Antonov, A., Issakov, S. & McClelland, A., 2018. Efficient SIMM-MVA Calculations for Callable Exotics. *Risk,* September.pp. 102-107.

Antonov, A., Issakov, S. & McClelland, A., 2019. MVA: future IM for client trades and dynamic hedges. *Risk,* August.

Basel Committee on Banking Supervision, 1996. *Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements.* [Online]
Available at: https://www.bis.org/publ/bcbs22.htm

Basel Committee on Banking Supervision, 2010. *Sound practices for backtesting counterparty credit risk models.* [Online]
Available at: https://www.bis.org/publ/bcbs185.pdf

Basel Committee on Banking Supervision, 2015. *Margin Requirements for non-Centrally Cleared Derivatives.* [Online]
Available at: http://www.bis.org/bcbs/publ/d317.pdf

Caspers, P., Giltinan, P., Lichters, R. & Nowaczyk, N., 2017. Forecasting Initial Margin Requirements - A Model Evaluation. *Journal of Risk Management in Financial Institutions,* February, 10(4), p. 365–394.

Chan, J., Zhu, S. & Tourtzevitch, B., 2018. Modelling Forward Initial Margin Requirements for Bilateral Trading. In: L. B. G. Andersen & M. Pykhtin, eds. *Margin in Derivatives Trading.* London: Risk Books.

Chorniy, V. & Arkhypov, S., 2020. *IM forecast quality: certain uncertainties and uncertain certainties.* s.l., Quant Minds.

Chorniy, V. & Greenberg, A., 2015. *Review of Equity-Credit Dependence Studies: Towards Building a Practical Equity-Credit Model for Counterparty Risk.* [Online]
Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2708143

Chorniy, V., TBA. *Backtesting: Literature Review and Discussion of General Methodological Aspects.* s.l.:To be published.

CME Group, 2019. *CME SPAN. Standard Portfolio Analysis of Risk.* [Online].

Eurex Clearing, 2018. *Prisma. Portfolio-based risk management.* [Online].

European Central Bank, 2017. *Guide for the Targeted Review of Internal Models (TRIM).* [Online].

European Central Bank, 2019. *ECB guide to internal models, Release number 2.2,* Frankfurt am Main: Banking Supervision.

Ganesan, N. & Hientzsch, B., 2021. Estimating future value-at-risk from value samples, and applications to future initial margin. *Journal of Risk,* 24(3), pp. 1-42.

Gneiting, T., 2011. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association,* Volume 106, pp. 746-762.

Green, A. & Kenyon, C., 2015. MVA by replication and regression. *Risk,* May.pp. 82-87.

Gurrola-Perez, P. & Murphy, D., 2015. *Filtered historical simulation Value-at-Risk models and their competitors,* Working Paper No. 525: Bank of England.

International Swaps and Derivatives Association, 2018. *ISDA SIMM Methodology.* [Online]
Available at: https://www.isda.org/a/zSpEE/ISDA-SIMM-v2.1-PUBLIC.pdf

Lou, W., 2016. MVA transfer Pricing. *Risk,* July.pp. 72-77.

McWalter, T. A. et al., 2022. Dynamic Initial Margin Estimation Based on Quantiles of Johnson Distributions. *Journal of Credit Risk,* 18(4), pp. 93-116.

Murphy, D., 2023. What can we expect from a good margin model? Observations from whole-distribution tests of risk-based initial margin models. *Journal of Risk Model Validation*, 17(2), p. 1–23.

Simon, M. K., 2002. *Probability Distributions Involving Gaussian Random Variables. A Handbook for Engineers and Scientists.* New York: Springer.

Trillos, C. A. G., Henrard, M. & Marcina, A., 2015. *Estimation of Future Initial Margins in a Multi-Curve Interest Rate Framework.* [Online]
Available at: http://ssrn.com/abstract=2682727

Wilkens, S. & Moran, L., 2017. Capturing Initial Margin in Counterparty Risk. *Journal of Risk Management in Financial Institutions,* Volume 10, pp. 118-129.

12. Appendix

In this appendix we consider the case when bank's IM forecast uses a very close, but not identical replication of actual IM model (SIMM or a particular CCP). In this case as we pointed in Section 6.2, 30% difference between the IM forecast and realisation and correspondently 30% CV of $\alpha(t,0)$ will point to a model or infrastructure failure. In this case the CV bounds should be tighter. It is possible to derive an estimate of these bounds by using a different interpretation of $\alpha(t,0)$. As in this appendix only initial point $(t,0)$ is used, we omit it to simplify notation and only refer to α, *IM* and *VaR*, replacing $\alpha(t,0)$, $IM(t,0)$ and $VaR(t,0)$.

Firstly we note that if both models are closely matched, then one would expect their VaR forecasts to be highly correlated and this correlation could be observed. Next, let us posit there exists in theory a perfect, a 'true' estimate of the VaR, VaR_{prf}, for any particular date. Based on the previous discussion it is reasonable to assume that this perfect model is sufficiently different from any practical model in existence in any bank, CCP or SIMM. In this case for the ratio, α_{CCP}, of CCP's VaR (or SIMM), i.e. the IM, to true VaR_{prf}, IM/VaR_{prf}, has acceptable bound within our framework at 30%. The same applies to ratio, α_{bank}, of the bank's VaR forecast to $\mathrm{VaR}_{\mathrm{prf}}$, VaR/VaR_{prf}. The ratio of $\alpha_{CCP}/\alpha_{bank}$ is then our familiar $\alpha = IM/VaR$. However in this interpretation it is driven by two correlated stochastic processes: α_{CCP} and α_{bank}. Our reader may notice immediately that both random variables α_{CCP} and α_{bank} are not observable, as the 'true' estimate of the VaR, VaR_{prf}, is not observable. On the other hand, as $\alpha = IM/VaR$ it could be viewed as a ratio of IM and VaR, both random variables are observable. At this point we need to discuss how different interpretations relates to our discussion of IM forecast errors in Sections 2 and 3.

We start by noting that our main estimate of acceptable level of the scaling variability comes from Section 2, illustration 3: "models are different, but both fit-for-purpose and industry standards, and use the same data for calibration". This is the observation of relative changes between three different fit-for-purpose VaR models, as shown in Figure 2. It looks at ratio of two VaR directly, in other words α, is *directly observed*, and we therefore do not require any individual information on ratio's components α_{CCP} and α_{bank} or IM and VaR. The only required assumption that model types are *sufficiently different* as per types illustrated. In practice most likely the internal bank's VaR estimate will use Monte Carlo type model and IM will be based on a flavour of either historic (some CCPs), var-covar (SIMM) or fixed scenarios (SPAN) type, so this requirement seems satisfied.

However in the same Section 2 we also considered two other illustrations of potential diversions between fit-for-purpose models. Both consider the error of VaR forecast, which could be interpreted as an error in relation to the non-observable 'true VaR'. Section 3 which considers regulators' view, also estimates the same type of error. For the purpose of our discussion these relate to α_{CCP} and α_{bank} individually, and both Section 2 and 3 suggest that the error below 30% is reasonable. To translate these estimates into the estimate of CV of α we need to treat it as a ratio $\alpha_{CCP}/\alpha_{bank}$ or IM/VaR and thus need to know: the correlation between α_{CCP} and α_{bank} or IM and VaR and the way to calculate the CV of a ratio of two correlated stochastic processes. We start with the former, the correlation, and consider the ratio of α_{CCP} and α_{bank} (our preference for dealing with α_{CCP} and α_{bank} instead of IM and VaR will become clearer as we progress in this section).

At this point we note the exact correlation of α_{CCP} and α_{bank} are not observable. However there are two suitable proxies. For the first proxy, we note that the correlation between IM and VaR is observable: this is correlation between different types of VaR model as shown in Figure 1. Figure 6 shows the correlations (a one-year moving time window) corresponding to 1-day VaRs on Figure 1. The correlations clearly vary with time, but we are interested in the long-term relationship between models; the average (median given in brackets) correlations for this portfolio for FHVaR/HVaR, FHVaR/MCVaR and HVaR/MCVaR are correspondingly 0.82 (0.94), 0.73 (0.83) and 0.7 (0.88) with former two most relevant to practical set-up (bank MC-based forecast vs. CCP). The correlations corresponding to 10-day VaR of the portfolio shown of Figure 1 is given on Figure 7 with average (median) values of 0.9 (0.96), 0.72 (0.87) and 0.7 (0.82) The average correlation varies across portfolios, although less than spot correlation across time. The average (median) correlations[32] across all portfolios as per Section 2 are provided in Table 2 below. The average (median) correlations across all portfolios for 5-day and 10-day are similar to 1-day VaR.

[32] We also calculated correlations for calm period only, as we did for CVs in Section 2, but unlike CVs the correlations for calm period are practically the same as for full period. This statement also applies for 5-day and 10-day VaR.

	Mean Correlation			Median Correlation		
	FHVaR/HVaR	FHVaR/MC VaR	HVaR/MC VaR	FHVaR/HVaR	FHVaR/MC VaR	HVaR/MC VaR
1-day	81%	59%	65%	92%	71%	76%
5-day	82%	57%	63%	92%	69%	75%
10-day	83%	57%	62%	92%	68%	75%

Table 2: Average and median correlations across all portfolios for 1, 5 and 10 day VaR

Figure 1, Figure 6 and Figure 7 show the portfolio with above average correlation and thus provide a convenient example to discuss the meaning of the correlation. We started our discussion in this section with a view that correlation is a rough measure of the level of replication of the actual IM model by the internal bank's VaR model. The assumption is that a higher degree of replication should translate, at least approximately, into higher correlation, especially for a very high, near perfect degree of replication. Thus the first meaning of correlation is a measure of similarity between bank's VaR and (CCP) IM models. The second meaning is purely functional. It is a proxy for another correlation, which is needed for a formalism to convert individual errors of two VaR models into a relative error: an error of one predicting another. Both meanings require correlation to be representative of the methodological relationship between two models and thus should be estimated over long time and across variety of portfolios. In our case we used 12 years of data (July 2007 to July 2019) and 100 portfolios, see Section 2. The meaning of correlation for individual portfolio and over short period of time is different. For example, both Figure 6 and Figure 7 show period of weeks, months, even years when the correlation is almost one (as high as 0.99). This does not mean that over these months, suddenly, MCVaR became perfect replication of FHVaR or HVaR. This means *the difference* between VaR forecasts *is persistent*. We already pointed this out in Section 2: "even under benign market conditions … relative difference between them could vary within the range of 30%, with difference variously realised on time scales from a day-to-day to longer periods, – it can persist days, weeks, months or even years". The period of high correlation is when the difference is locked-in, the period of lower correlation reflects the period when relative error changes, sometimes abruptly. This short term meaning of correlation reiterates that in this section we need to use long term correlation averaged across portfolios.

To conclude the discussion of the correlation between IM and VaR, if used as a proxy of the exact correlation of α_{CCP} and α_{bank} our results suggest that an overall correlation across different fit-for-purpose models with MC VaR as common predictor (expected to be used by bank's internal CCR engine) is about 0.6, but when one historic VaR predict another then higher value of at least 0.8 should be used. We note if former relates to majority on IMM PFE/xVA set ups, the latter could be relevant if IM is forecasted by PFE system built on historic VaR

basis with forward projection. Based on our discussion at the industry forums it appears that for fully collateralised, mostly linear and mostly non-path dependent instruments over relatively short horizons (as, for example, in some prime brokerage activities) a methodology based on market risk VaR, which could be one of historic VaR flavours, is sometimes used by the industry. In the case of bank's IM and CCP IM based on the same flavour of historic VaR the correlation may be much higher than 0.8.

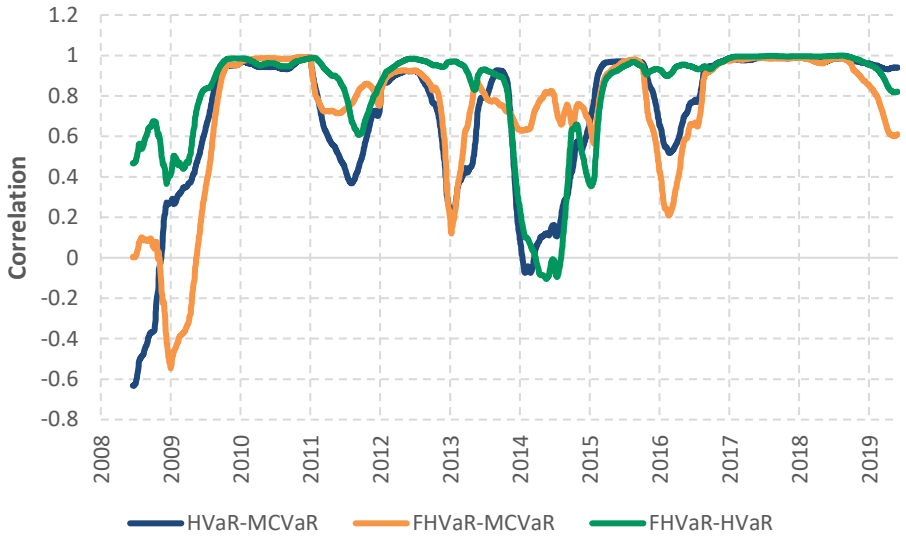

Figure 6: 1-year correlations corresponding to different VaR methodologies for 99th percentile 1-day horizon

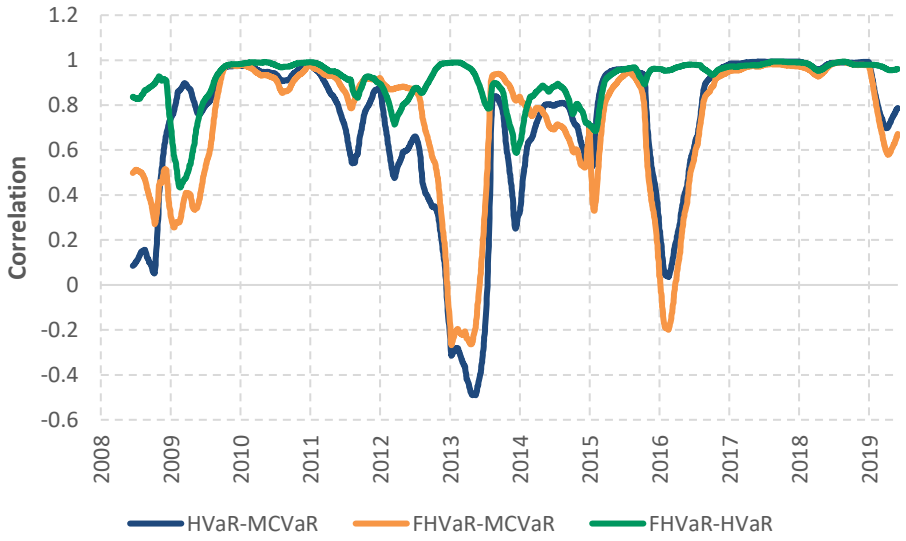

Figure 7: 1-year correlation corresponding to different VaR methodologies for 99th percentile 10-day horizon

The problem with correlation between IM and VaR as a proxy is that it is not clear how well it approximates correlation α_{CCP} and α_{bank}. Thus for the second proxy we approximate 'true'

VaR by averaging the VaR values by the three models on the day and then averaging it over 6 months (thus matching period used in defining spot volatility for normalisation in our filtered historic VaR model, see Sections 2). Then we use this value to obtain α_{CCP} and α_{bank}. For convenience, we abbreviate three alpha values as $\alpha_{VaR1,2,3}$ for three types of VaR. The average (median) correlation calculated across all portfolios across full period of observation are given in Table 3 below.

	Mean Correlation			Median Correlation		
	a_{FHVaR} / a_{HVaR}	a_{FHVaR} / a_{MCVaR}	a_{HVaR} / a_{MCVaR}	a_{FHVaR} / a_{HVaR}	a_{FHVaR} / a_{MCVaR}	a_{HVaR} / a_{MCVaR}
1-day	75%	50%	63%	86%	61%	72%
5-day	76%	49%	61%	86%	59%	69%
10-day	77%	48%	60%	86%	59%	70%

Table 3: Average and median correlations between ratios across all portfolios for 1, 5 and 10 bd VaR

Note that 1-day VaR values are similar for 5-day and 10-day. The values for calm period only are similar to full period and not presented.

To use this proxy we also need to know CV of $\alpha_{VaR1,2,3}$. Also as per our discussion in Section 2 we are specifically interested in CV for a relatively calm of 2013-2019 for which we proposed of 30% as red boundary. Values are provided in Table 4 below. The results for 1-day, 5 and 10-days are all very similar.

		Coefficient of variation		
		a_{FHVaR}	a_{HVaR}	a_{MCVaR}
Calm period	1-day	*12%*	9%	*11%*
	5-day	13%	9%	13%
	10-day	*14%*	10%	*15%*
Full period	1-day	22%	13%	18%
	5-day	21%	13%	18%
	10-day	22%	14%	19%

Table 4: Coefficient of variation across all portfolios for 1, 5 and 10 day VaR

Thus the numbers are, expectedly, similar to CVs of ratio calculated in Section 2. We note that all are well below 30%, which relevant for our next step.

Having discussed the correlation, and CV of each process ($\alpha_{VaR1,2,3}$) now we turn to the calculation of CV of ratio of two correlated stochastic processes. It requires certain approximations. (Simon, 2002) gives exact solution for probability density function for ratio of two correlated Gaussians with non-zero means, from which one can derive CV numerically. Alternative estimate using Taylor approximation for means and variance gives a formula[33] for CV, denoted as ν, directly as function of correlation and CV's of both α_{CCP} and α_{bank}, denoted as ν_{CCP} and ν_{bank}:

$$\nu = \frac{\sqrt{\nu_{CCP}^2+\nu_{bank}^2-2\varrho\nu_{CCP}\nu_{bank}}}{1+\nu_{bank}^2-\varrho\nu_{CCP}\nu_{bank}} \qquad \text{Equation 6}$$

assuming the CV of both α_{CCP} and α_{bank} have equal value of ν_a the equation further simplifies to:

$$\nu = \sqrt{2}\nu_a \frac{\sqrt{1-\varrho}}{1+(1-\varrho)\nu_a^2} \qquad \text{Equation 7}$$

where ϱ is the correlation between a_{CCP}^0 and a_{bank}^0. Both CV from this formula and the one implied by (Simon, 2002) match well for ν_a under 30% (especially for high correlations) and near perfectly at 10%, but we suggest the simple equation is more convenient for quick analysis. Figure 8, Figure 9 and Table 5 are provided below to assist readers. Figure 8 and Figure 9 show ν as function of ϱ at ν_a of 10% and 30% correspondingly. Now, returning to our earlier point, we can state that the convenience of Equations 6 and 7 together with convenience of using CVs which happen to have low values explains our preference for using $\alpha_{VaR1,2,3}$ vs. actual VaR values.

[33] We are grateful to our colleagues Ziad Fares and Xavier Lorentz for the discussion of the case when scaling is driven by two correlated stochastic processes and for providing us with the expression for its variability, which we quote.

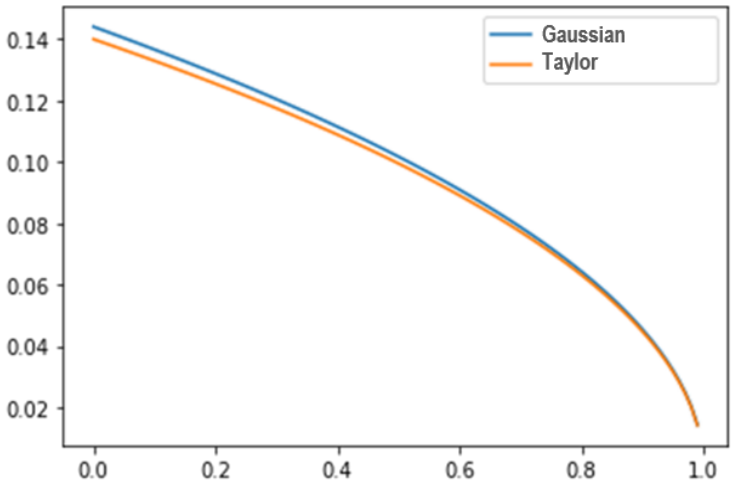

Figure 8: Coefficient of variation of the ratio (Y-axis), depending on correlation (with coefficient of variation of the individual parts at 0.1)

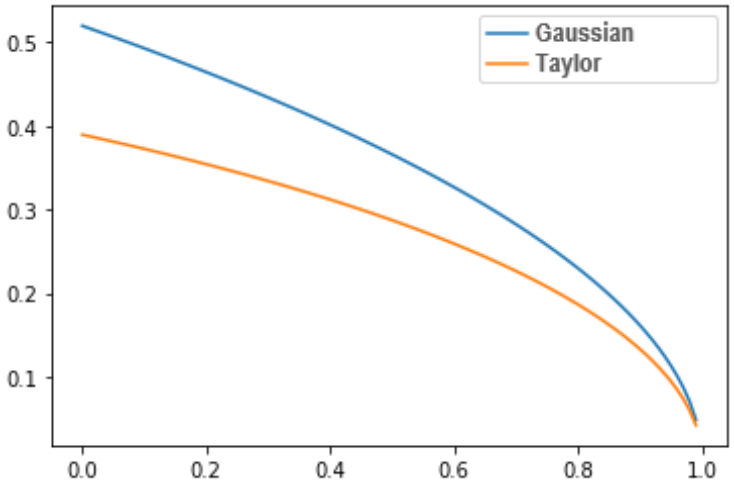

Figure 9: Coefficient of variation of the ratio (Y-axis), depending on correlation (with coefficient of variation of the individual parts at 0.3)

Table 5 shows for the case of high correlation the value of ν derived via (Simon, 2002) as 'Gaussian' and via the Equation 7 as 'Taylor' with ν_a as 'Coefficient of variation ':

	Coefficient of variation: 10%		Coefficient of variation: 30%	
	Correlation: 0.9	Correlation: 0.99	Correlation: 0.9	Correlation: 0.99
Gaussian	4.54%	1.44%	16.13%	4.87%
Taylor	4.47%	1.41%	13.30%	4.24%

Table 5: Coefficient of variation of the ratio at different coefficient of variation and correlation levels

At this point, we observe that ν_a in 10 - 30% range and correlation of about 0.5 to 0.7 result in roughly equal ν and ν_a, and 0.5 to 0.7 is precisely the range of correlation suggested by earlier discussion of fit-for-purpose models with bank using MC approach. Recall that in Section 2 we did not consider scaling as a ratio to define red boundary, which we proposed as 30% for non-stressed markets. However roughly equal ν and ν_a show that if we do treat it as a ratio and then apply other estimates of uncertainty from Section 2 and 3 we still get to similar recommendation of about 30% as a red boundary. All estimates point to same range, thus supporting our broad assessment in Section 6.2, Footnote 19 that the boundary is applicable across asset classes/risk factors, not only equity. In case of two types of historic VaR the 30% value is based on Figure 2 and therefore still stands as long as model types are different, but the ratio arguments may suggest slightly lower number[34]. This brings us to our next point.

We note that the interpretation of a as a ratio of two correlated stochastic processes is most interesting in highly correlated case, as it corresponds to the case of close replication of IM methodology with bank's CCR engine. It also may be relevant in the case when PFE system is built on historic VaR basis with forwards projection as discussed earlier in this section. So let us consider the case of high correlation closer. For ϱ = 0.9 the CV of 10% and 30% translates into much more narrow range of about 4.5% and 15% for the green and red boundaries. Of course, this approach of defining colour boundaries for the case of close internal IM replication leaves open a question how to define or calculate ϱ, so it could act as a key measure of two model similarities or replication's quality. However, the two proxies we discussed earlier suggest that historic correlation between IM and VaR on some test portfolios may be a good basic benchmark.

Alternatively, although it might be possible to use this framework with some sort of a formal derivation of ϱ, a practical rule of thumb could be a good start: classify models into three wide groups: 'generic', 'close replication', and 'near perfect replication', choose red boundary for the 'generic' based on main discussion in this paper (for example, 30%), and then convert it using 0.9 and 0.99 correlations for 'close replication' and 'near perfect replication' models. In our example of 30% 'generic' red boundary translates (rounding the entry of Table 5) to 15% and 4.5% correspondingly, and 10% 'generic' green boundary translates into 5% and 1.5%[35] (the same numbers tabulated for reader's convenience in Section 6.2, Table 1).

[34] The value of 30% is a suggestion and we must repeat that any boundary should be calibrated by each institution to account for inherent specifics embedded into their models and risk management processes.

[35] Assuming equal spacing of red/amber/yellow/green zones, the zone boundaries will be 30/20/10, 15/10/5, 4.5/3/1.5 percent for 'generic', 'close replication', and 'near perfect replication' correspondingly.