Bayesian Backtesting for Counterparty Risk Models

Mante Zelvyte, Matthias Arnsdorf

Quantitative Research, JP Morgan

February 2022

The statements, views and opinions expressed in this article are our own and do not necessarily reflect those of our employer, JPMorgan Chase & Co., its affiliates, other employees or clients.

Abstract

In this article we introduce a new framework for counterparty risk model backtesting based on Bayesian methods. This provides a conceptually sound approach for analyzing model performance which is also straightforward to implement. We show that our methodology provides important advantages over a typical, classical, backtesting set-up. In particular, we find that the Bayesian approach outperforms the classical one in identifying whether a model is correctly specified which is the principal aim of any backtesting framework. The power of the methodology is due to its ability to test individual model parameters and hence identify which aspects of a model are misspecified as well as the degree of misspecification. This greatly facilitates the impact assessment of model issues as well as their remediation.

1 Introduction

Models are imperfect representations of reality and model testing is thus a vital component of model development. Backtesting is a principal method for evaluating the accuracy of model outputs and is used extensively in finance to test predictive models on historical data. In risk management, two prominent areas of focus are value-at-risk (VaR) as well as counterparty credit risk (CCR) models.

A comprehensive backtesting framework is part of the regulatory requirements for risk models. The guidance for market risk VaR models, as prescribed by the Basel Committee (BIS [1996]), is detailed and relies on null hypothesis significance testing using "red-amber-green traffic light" evaluation criteria. For CCR backtesting, regulators provide only high level guidance with the key success criterion defined as the consistency of the observed realizations of risk factors with those forecast by the model (BIS [2010]).

This means that for CCR, firms have considerable freedom in the choice of their backtesting approach and there are a number of discussions in the literature on how this can be designed optimally, see e.g. Ruiz [2014] and Anfuso et al. [2014]. The methods typically lean heavily on the regulatory market risk framework and are based on standard, frequentist, hypothesis testing.

However, it is well known that there are serious shortcomings with the frequentist approach both from a conceptual and practical point of view. A good overview is provided by Wasserstein et al. [2019] and references therein. One alternative is to adopt Bayesian methods. This is widely used in fields as diverse as criminology and medicine but has, to our knowledge, not been applied to counterparty credit model backtesting in finance. This paper aims to fill this gap by presenting a Bayesian framework that provides a more powerful as well as more coherent testing regime while at the same time being straightforward to implement in practice.

1.1 Issues with p-values

Backtesting aims to test how well a model predicted distribution compares to realizations of a particular modeled variable. In the typical frequentist approach this is done by calculating the probability of obtaining results that are at least as extreme as the ones observed assuming that the model being tested is correct. This probability is the so-called *p-value*. We can write this symbolically as $p = p(\text{data}|\text{model})$. The assumption that the model is correct is often referred to as the “null hypothesis”. The p-value is then compared to pre-defined thresholds to determine whether the model should be rejected. The Basel VaR framework recommends that models should be rated “red” for p-values below 0.01%, “amber" for $0.01\% < p < 5\%$ and “green” otherwise. In other words a model should be rejected if the observed outcomes are less than 0.01% likely under the given model.

While this provides a practical and seemingly sensible recipe, there are numerous issues that arise.

- The most immediate question is whether the framework outlined above is able to determine whether a given model is satisfactory. The null hypothesis significance testing is designed to avoid incorrectly rejecting a model if it is correct, i.e. to avoid a type I error or false positive. Therefore, by design, it acts as a safeguard against reporting insignificant findings in scientific applications. While this is certainly a desirable feature, it does not provide confidence that the model is not misspecified. For this we need to ensure that the type II error rate is controlled and that incorrect models are flagged as deficient. The ability of the backtesting to achieve this is referred to as the “power” of the test. Unfortunately, CCR backtesting fundamentally suffers from the limited availability of observations. This means that the power of backtesting is typically low (c.f. e.g. Clayton [2019]).

- One obvious approach to overcome the lack of data in any given observation window is to combine the results of repeated backtesting experiments over time and to group risk factors and portfolios. However, a well-known issue with p-values is that it is very difficult combine and aggregate these in a rigorous way which makes the assessments of true error rates impractical (Lindley [2000]).

- Another common misconception is that the p-value itself is a measure of how well-specified a model is, i.e. that a lower p-value indicates a worse model. In general, however, this is not the case. The p-value only refers to the probability of a particular observation *assuming the model is correct*. As such, it does not tell us how likely it is that the model is correct or to what degree it is misspecified.

This is related to the practical difficulty of translating the p-value into an error metric defined in terms of the ultimate outputs of our model or in terms of actual financial loss.

- Fundamentally, all the issues above stem from the fact that the p-value is the answer to the wrong question. It tells us the probability of the data given the model $P(\text{data}|\text{model})$, but what we need is the probability that the model is correct given the observational data, i.e.: $P(\text{model}|\text{data})$. The p-value can only indicate evidence against the null hypothesis but it cannot be used to demonstrate that our model is valid.

In summary and as expressed cogently in Wasserstein et al. [2019]: “no p-value can reveal the plausibility, presence, truth, or importance of an association or effect. Therefore, a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important”.

1.2 A Bayesian Approach

A solution to the issues above is offered by the Bayesian paradigm which provides an internally consistent approach for model testing. At a basic level the idea is that if we want to study $P(\text{model}|\text{data})$ then we can make use of Bayes rule which tells us that

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})}. \tag{1.1}$$

Here $P(\text{data}|\text{model})$ is referred to as the *likelihood* and $P(\text{model})$ as the *prior*. The result $P(\text{model}|\text{data})$ is the *posterior* distribution which encodes all the information about the model that we can ascertain from the data and prior knowledge.

The symbolic relationship above makes clear that the difference between the frequentist hypothesis testing approach is that in a Bayesian framework we do not assume that our model is correct but instead consider all possible model choices weighted by their prior probabilities. This leads to the following advantages:

- In a Bayesian framework we work directly with $P(\text{model}|\text{data})$ which is the quantity of interest. Because this is a probability distribution it allows us for internally consistent inference and does not suffer from the conceptual issues inherent in hypothesis testing (Lindley [2000]).

- The posterior distribution allows for a very intuitive presentation of results. Success criteria can be imposed directly on the probability of the model being correctly specified given the observed data. In addition, a Bayesian framework provides a direct link between backtesting results and the impact on end-usage metrics such as expected exposure or potential future exposure. This greatly facilitates the decision making process that is the ultimate aim of any backtesting framework.

- Working with probabilities also allows for a coherent way to consistently combine results across time and, in principle, also aggregate results across risk factors or portfolios. This is important given the lack of observed data in any given quarter. A key concept in this regard is that the prior probability distribution can be continuously updated given new information. This is captured by the Bayesian notion that "today's posterior is tomorrow's prior". In the context of a quarterly CCR backtesting program that means that the prior distribution will become more and more informative over time allowing greater ability to test a model's performance.

- Another important feature of the proposed framework is that it allows to isolate and test particular features of a model. For example, if we are interested in the predicted risk factor distribution we may be particularly focused on testing the model volatility of risk factor, the drift or the tail behaviour of the distribution. Using the Bayesian approach we can not only determine whether a model performs adequately overall but also specifically ask which model parameters are less well supported by the data than others.

There are two key issues that are raised in the context of Bayesian statistics. One is that the required modeling can be very complex. The second, more fundamental, criticism is that a prior distribution is by construction subjective and, in principle, unknown. This subjectivity is often viewed as undesirable in a testing framework.

In the context of CCR backtesting the first issue does not arise. As we will see below the modeling involved in the Bayesian backtesting is basic and straightforward. The issue regarding the choice of prior is also well-mitigated for two reasons. Firstly, as we have discussed above (and is also argued by Harvey [2017]), we benefit from the accumulation of market data over time which results in continually increasing understanding of financial markets and, consequently, certainty in our prior beliefs. Secondly, the main parameters of interest in a model will be subject to extensive governance and controls. This means that there will be significant prior information regarding any parameter uncertainty.

In summary, as we will show in this article, Bayesian statistics provide a powerful framework for CCR backtesting that is both intuitive and practical.

2 Classical Backtesting

A typical counterparty credit risk exposure model consists of a system of risk factor diffusions implemented via a high-dimensional Monte Carlo simulation. The risk factor processes are used to calculate counterparty portfolio level quantities such as the future value (also referred to as the mark-to-market or MtM), the expected exposure (EE) or tail metrics such as the potential future exposure[1] (PFE). The aim of CCR backtesting (BT) is to test whether the predicted risk factor distributions are a good approximation of the real-world dynamics. A typical BT framework will consist of tests at the risk factor level as well as tests for the aggregate quantities such as portfolio exposure.

Backtesting will be conducted over a testing window Δ_W, which needs to be long enough to include sufficient data but also short enough to reflect the current model specification and market regime. A typical window period will be between one quarter

[1] The PFE, also referred to as peak exposure, at a future time is typically defined as a quantile or expected shortfall of the simulated MtM distribution.

and one year. Within a chosen window we observe risk factor returns over different forecasting horizons Δ_H. This means that the observation window is sub-divided into a set of $n+1$ observation dates $t_0, ..., t_n$ such that $t_n - t_0 = \Delta_W$, $t_i - t_{i-1} = \Delta_H$ and $n = \Delta_W / \Delta_H$.

The risk factor return over the future period $[t_{i-1}, t_i]$ is a random variable denoted by R_i. The actual, real-world, distribution of R_i is not known. We denote the real-world cumulative distribution function (CDF) by $F_i^R(x) \equiv P^R(R_i \leqslant x)$ and the corresponding density by $f^R(x)$. The risk factor model will assume a distribution for R_i which may not be the same as f_i^R. We denote the model CDF by $F^M(x) \equiv P^M(R_i \leqslant x)$ and the density by $f^M(x)$. The aim of the backtesting framework is to assess how close f_i^M is to f_i^R.

One of the complications of testing CCR models is that models and markets are not static and observations depend on the state of the market at the start of the observation window. This means that the model distributions $f_i^M(x)$ can be different at each observation date. In the classical framework this issue is addressed by making use of the "probability integral transform" (PIT) of the observed data. This acts as a standardization of observations by removing the dependence on market state and model changes.

Given a set of observations of realized risk-factor returns r_i over the periods $[t_{i-1}, t_i]$ we define the probability integral transform as the set of transformed variables $\{F_i^M(r_i)\}$ obtained by applying the model CDF F^M to all observations r_i:

$$F_i^M(r_i) = \int_{-\infty}^{r_i} f_i^M(x) \mathrm{d}x. \tag{2.1}$$

A fundamental result is that the set $\{F_i^M(r_i)\}$ will be uniformly distributed if and only if the model distributions f_i^M correspond to the actual distributions of r_i over each period, i.e. $f_i^M(x) = f_i^R(x)$ (the idea dating back to Rosenblatt [1952], see also Diebold et al. [1998]). Therefore the discrepancy of the distribution of $f_i^M(r_i)$ compared to a uniform distribution can be used as a measure of the difference between the probability densities underlying the model and real-world dynamics.

In practice the PIT is calculated using the Monte-Carlo risk factor simulation. At each observation date t_i we simulate N estimates of the risk factor return under the distribution f_i^M. This results in a set of modeled values $r_{i,j}^M$ with $0 < j \leqslant N$. The empirical cumulative distribution $\hat{F}_i^M$ is then constructed from the realized values as follows:

$$y_i \equiv \hat{F}_i^M(r_i) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{\{r_{i,j}^M \leqslant r_i\}}. \tag{2.2}$$

Standard distributional tests of uniformity can then be performed on the transformed data. Some of the most commonly used tests are the binomial, multinomial or chi-squared (χ^2) goodness-of-fit tests. The discretized versions of the Anderson-Darling, Cramer-Von-Mises and Kolmogorov-Smirnov tests are also cited in the counterparty credit risk backtesting literature (Anfuso et al. [2014], Clayton [2019]). The choice of the test statistic depends on the specific part of the distribution we wish to analyse.

Given a test statistic for a set of observed data, one can define the p-value as the probability of the statistic being as or more extreme than the one observed. The p-value is then compared to pre-defined thresholds to determine any remediation action. More details will be provided in later sections.

3 Bayesian Backtesting Framework

The Bayesian backtesting framework we introduce here assumes the same test set-up as in the classical case. As before, we are interested in the performance of a simulated risk factor for which we have a set of n return observations r_i. Using the PIT these are transformed into the data set $\boldsymbol{y} \equiv \{y_i\}_{i=1}^n$. Our aim is to calculate the probability that our model specification is correct given the data $\boldsymbol{y}$.

To do this we first assume that the model implied risk factor distributions are fully parameterized at each time point t_i by a set of m model parameters $\boldsymbol{\beta}_i \equiv \{\beta_i^j\}_{j=1}^m$. This means we can write

$$f_i^M(x) = f_i(x; \boldsymbol{\beta}_i) \tag{3.1}$$

for some parametric distribution f_i. Similarly, we assume that the actual, real-world, distribution f_i^R is of the same form[2] but with potentially different and unknown parameters $\boldsymbol{\gamma}_i = \{\gamma_i^j\}_{j=1}^m$, i.e.:

$$f_i^R(x) = f_i(x; \boldsymbol{\gamma}_i). \tag{3.2}$$

The choice and complexity of the parameterization used in practice will depend on which aspects of the risk factor distribution we wish to analyze.

We would like to calculate the posterior distributions $P(\boldsymbol{\gamma}_i|y_i)$. Bayes theorem tells us that these can be calculated as

$$P(\boldsymbol{\gamma}_i|y_i) = \frac{P(y_i|\boldsymbol{\gamma}_i)\pi(\boldsymbol{\gamma}_i)}{P(y_i)}, \tag{3.3}$$

where π is a prior distribution on the parameters $\boldsymbol{\gamma}_i$ which we assume does not depend on the period i for simplicity. The likelihood function $P(y_i|\boldsymbol{\gamma}_i)$ can, in principle, be calculated by noting that

$$\begin{aligned}
P(Y \leqslant y_i|\boldsymbol{\gamma}_i) &= P(F_i^M(R_i) \leqslant y_i; \boldsymbol{\gamma}_i) && (3.4)\\
&= P(F_i(R_i; \boldsymbol{\beta}_i) \leqslant y_i; \boldsymbol{\gamma}_i) && (3.5)\\
&= P(R_i \leqslant F_i^{-1}(y_i; \boldsymbol{\beta}_i); \boldsymbol{\gamma}_i) && (3.6)\\
&= F_i(F_i^{-1}(y_i; \boldsymbol{\beta}_i); \boldsymbol{\gamma}_i) && (3.7)
\end{aligned}$$

and hence:

$$P(y_i|\boldsymbol{\gamma}_i) = F_i'(F_i^{-1}(y_i; \boldsymbol{\beta}_i); \boldsymbol{\gamma}_i), \tag{3.8}$$

where $F_i(\cdot)$ denotes the cumulative distribution function of $f_i(\cdot)$ and F_i' is the derivative of F_i with respect to y_i.

[2] Note that for any two distributions it is always possible to define a more general distribution that contains both distributions as special cases.

It is convenient to remove the dependency on the model parameters $\boldsymbol{\beta}_i$ and work instead in terms of a relative misspecification θ_i^j of β_i^j vs. γ_i^j. This is important because estimating the model parameters is expensive if they are not known explicitly and also because we will need variables that are time-invariant. In general, the misspecifications $\boldsymbol{\theta}_i = \{\theta_i^j\}_{j=1}^m$ can be defined implicitly with respect to a fixed and time-independent set of base parameters $\boldsymbol{\alpha} = \{\alpha^j\}_{j=1}^m$ by demanding

$$F_i(F_i^{-1}(y_i;\boldsymbol{\alpha});\boldsymbol{\theta}_i) = F_i(F_i^{-1}(y_i;\boldsymbol{\beta}_i);\boldsymbol{\gamma}_i) \tag{3.9}$$

for all y_i. In practice, the above equation does not always lead to an explicit relationship and the parametric distribution needs to be chosen carefully. Two key examples are provided below where we will see that the base parameter is 1 or 0 and the resulting misspecification is the ratio or the difference of the real-world parameter to the model parameter.

We can express the posterior distribution at time t_i in terms of the misspecification as

$$P(\boldsymbol{\theta}_i|y_i) = \frac{P(y_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)}{P(y_i)} = \frac{P(y_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)}{\int_{\boldsymbol{\theta}_i} P(y_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i} \tag{3.10}$$

with

$$P(y_i|\boldsymbol{\theta}_i) = F_i'(z_i;\boldsymbol{\theta}_i) = \frac{f_i(z_i;\boldsymbol{\theta}_i)}{f_i(z_i;\boldsymbol{\alpha})} \tag{3.11}$$

and

$$z_i = F_i^{-1}(y_i;\boldsymbol{\alpha}). \tag{3.12}$$

To specify the prior distribution we will make the simplifying assumption that the individual parameters are independent and hence

$$\pi(\boldsymbol{\theta}_i) = \prod_{j=1}^m \pi^j(\theta_i^j), \tag{3.13}$$

where the individual priors π^j are an input to the backtesting as will be discussed in the next section.

In order to extend this to multiple periods we assume that the misspecifications are the same across our observation window and hence $\boldsymbol{\theta}_i = \boldsymbol{\theta}$ for all i. We should think of the misspecifications as the long run average relationship between the model and real-world parameters. Assuming in addition that returns are independent over each horizon allows us to define the total likelihood for the observation window as

$$P(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^n P(y_i|\boldsymbol{\theta}). \tag{3.14}$$

The posterior distribution for the entire observation window is then given by

$$P(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{P(\boldsymbol{y})} = \frac{P(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} P(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{3.15}$$

We note that this can also be interpreted as an experiment where we perform repeated observations and for each subsequent observation take the new prior to be the posterior estimated previously.

We now discuss the key terms in the above expression in more detail.

3.1 The Likelihood

The main computational effort in the Bayesian backtesting approach is the calculation of the likelihood function. Here we provide two basic examples which will be used in our empirical analysis later on.

3.1.1 Gaussian Example

The simplest example is when we assume that the model and real-world distributions can be parameterized by Gaussian distributions. We also assume for simplicity that the distributions are the same across time windows. In this case we have

$$f^M(x) = \frac{1}{\sigma^M}\phi\left(\frac{x-\mu^M}{\sigma^M}\right) \tag{3.16}$$

and

$$f^R(x) = \frac{1}{\sigma^R}\phi\left(\frac{x-\mu^R}{\sigma^R}\right), \tag{3.17}$$

where $\phi(\cdot)$ is the standard normal density and σ^M, σ^R, μ^M, μ^R are the model and real-world volatilities and mean levels. According to equation (3.8) we can then express the likelihood as:

$$P(y_i|\mu^R, \sigma^R, \mu^M, \sigma^M) = \frac{1}{\sigma^R}\frac{\sigma^M}{\phi(\Phi^{-1}(y_i))}\phi\left(\frac{\sigma^M\Phi^{-1}(y_i) - (\mu^R - \mu^M)}{\sigma^R}\right) \tag{3.18}$$

with the cumulative normal distribution denoted by $\Phi(\cdot)$. It is hence natural to define the misspecification parameters as

$$\theta_\sigma = \frac{\sigma^R}{\sigma^M} \quad \text{and} \quad \theta_\mu = \frac{\mu^R - \mu^M}{\sigma^M} \tag{3.19}$$

which represent the relative volatilities and mean levels and corresponds to choosing $\alpha_\sigma = 1$ and $\alpha_\mu = 0$. With this we have:

$$P(y_i|\theta_\mu, \theta_\sigma) = \frac{1}{\theta_\sigma\phi(\Phi^{-1}(y_i))}\phi\left(\frac{\Phi^{-1}(y_i) - \theta_\mu}{\theta_\sigma}\right). \tag{3.20}$$

3.1.2 Student-t Example

In order to study the tails of the risk factor distributions we need to resort to more complex distribution parameterizations. Here we provide the likelihood in the case where the real-world realizations follow a Student-t distribution while the model assumes a Gaussian distribution. Hence we have

$$f^M(x) = \frac{1}{\sigma^M}\phi\left(\frac{x-\mu^M}{\sigma^M}\right) \tag{3.21}$$

and

$$f^R(x) = \frac{1}{\hat{\sigma}^R}t_{\nu^R}\left(\frac{x-\mu^R}{\hat{\sigma}^R}\right), \tag{3.22}$$

where $t_{\nu^R}(\cdot)$ is the standard Student-t distribution with ν^R degrees of freedom. The standard deviation σ^R of f^R is related to $\hat{\sigma}^R$ by

$$\sigma^R = \hat{\sigma}^R \sqrt{\frac{\nu^R}{\nu^R - 2}} \tag{3.23}$$

which is defined for $\nu^R > 2$.

As in our previous example, the likelihood is then derived as

$$P(y_i|\mu^R, \hat{\sigma}^R, \nu^R, \mu^M, \sigma^M) = \frac{1}{\hat{\sigma}^R} \frac{\sigma^M}{\phi(\Phi^{-1}(y_i))} t_{\nu^R} \left(\frac{\sigma^M \Phi^{-1}(y_i) - (\mu^R - \mu^M)}{\hat{\sigma}^R} \right). \tag{3.24}$$

We can define the misspecifications for the mean and volatility as before:

$$\theta_\sigma = \frac{\hat{\sigma}^R}{\sigma^M} \quad \text{and} \quad \theta_\mu = \frac{\mu^R - \mu^M}{\sigma^M}, \tag{3.25}$$

but we note that there is no equivalent expression for the degrees of freedom parameter since we have assumed that the model uses a Gaussian distribution which can be thought of as a Student-t distribution with infinite degrees of freedom. Hence we set $\theta_\nu = \nu^R$ which leads to our final expression for the likelihood:

$$P(y_i|\theta_\mu, \theta_\sigma, \theta_\nu) = \frac{1}{\theta_\sigma \phi(\Phi^{-1}(y_i))} t_{\theta_\nu} \left(\frac{\Phi^{-1}(y_i) - \theta_\mu}{\theta_\sigma} \right). \tag{3.26}$$

This example can be generalized to more complex distributions but that is out of the scope of this article.

3.2 The Prior

The prior acts as a “weighting” of the likelihood function for different values of the model parameters. It represents our prior knowledge (or lack of it) and is a key distinguishing feature of Bayesian analysis that facilitates the estimation of the densities of parameter misspecifications. The choice of prior distribution depends on the parameter type and our confidence in the model specification. For parameters like the mean level, which can take on any negative or positive value, we can use a Gaussian distribution as a prior. For parameters like volatility, which are positive, a typical choice of prior is the Gamma distribution.

The initial mean of the prior is set to our best guess of the true parameter value and the width of the distribution reflects our uncertainty. In the backtesting context, for example, we should have high confidence in our model parameters when we are valuing standard instruments on highly liquid assets since market observability will be good. Conversely, when dealing with highly exotic products in illiquid markets confidence will be lower. In absence of any prior knowledge we can choose a distribution that gives roughly equal weight to all parameter values. The parameters that govern the prior distribution shape are referred to as *hyperparameters*.

A powerful feature of our backtesting approach is that the prior distribution can be naturally updated over time. In particular, we can use the posterior of one observation window as the prior for the next window. This provides a consistent way of incorporating all available historical information in our analysis and partially resolves the problem of lack of observations. Given that backtesting is an ongoing requirement this means also that any choice of initial prior will become irrelevant over time. This will be examined in detail in later sections.

We can also accommodate the fact that historical results should have less weight than the current ones by weighting the likelihood of the historical data to arrive at a modified prior, e.g. by defining the prior in the n^{th} observation window as

$$\pi_n(\boldsymbol{\theta}) \propto \prod_{k=1}^{n-1} P(\boldsymbol{y}_{n-k}|\boldsymbol{\theta})^{\omega^k} \pi_0(\boldsymbol{\theta}),$$

where $P(\boldsymbol{y}_{n-k}|\boldsymbol{\theta})$ is the likelihood from observation window $n-k$ and $\pi_0(\cdot)$ is the initial prior. The weights ω^k with $0 \leqslant \omega \leqslant 1$ determine how much importance we place on the historical data.

Finally, we note that unless prior and posterior distributions are conjugate, the functional form of the prior and posterior distributions for each observation window will be different. To retain the same functional form of the priors, one can fit the prior to the posterior via moment matching.

3.3 The Posterior

The posterior distribution incorporates all the information that can be used to evaluate the performance of our model given the available data and our prior knowledge. In particular, the posterior allows us to infer the best estimates for our model parameters. This is a key difference to the classical approach which only assesses the overall validity of the model without providing any information about which aspects of the model may be misspecified.

To analyze individual parameters we need to estimate the marginal posterior distributions for each parameter of interest which can be done by marginalizing over all other parameters. For example, in the Gaussian case the posterior distribution for the volatility misspecification θ_σ is obtained by integrating over the mean misspecification θ_μ:

$$P(\theta_\sigma|\boldsymbol{y}) \equiv \int_{-\infty}^{\infty} P(\theta_\sigma, \theta_\mu|\boldsymbol{y}) \mathrm{d}\theta_\mu. \tag{3.27}$$

Typical measures that are used in analyzing the posterior distribution are:

- The mean or median of the distribution. This provides the best estimate of each model parameter.
- The standard deviation or highest posterior density interval (HPD) of the distribution. These are measures of the uncertainty in our parameter estimates. For a confidence level p and a parameter θ, the HPD is defined as the smallest interval $[a, b]$ such that $P(a < \theta < b|\boldsymbol{y}) = p$.

A main goal of any backtesting framework is to determine whether the model under consideration performs satisfactorily. This requires a criteria to accept or reject the model, similarly to what we have in the classical framework. The advantage of the

Bayesian framework is that we can base decisions directly on the probability of the model being correct. In addition, we can determine the size of any misspecification.

The posterior distribution allows us to define many possible criteria depending on the desired use case. In this article we consider the probability of model parameters being correct up to a certain tolerance. More specifically, for each parameter θ and chosen tolerance ϵ we can calculate the probability of the parameter being correct as

$$P(\bar{\theta} - \epsilon < \theta < \bar{\theta} + \epsilon | \boldsymbol{y}),$$

where $\bar{\theta}$ corresponds the non-misspecified value of the parameter. To accept our model we can then demand that this probability exceeds a chosen threshold, e.g. 95%. The error tolerance ϵ and the probability threshold need to be chosen based on the accuracy requirements for our model. It is worth re-emphasizing the following points:

- The probability threshold is on the probability of the model being correct. Accepting the model means that there is a high probability that the model is not misspecified. This is in contrast to the classical approach where the threshold is on the probability of observing the data assuming the model is correct. Accepting the model in this case does not mean that the model is correct.

- The tolerance range for the misspecification can be directly related to the amount of loss incurred as a function of the model misspecification. Typically, sensitivities of end-usage metrics to the main model parameters will be available to do this. More generally, we can asses any quantity that is a function of the model parameters in terms of our posterior distribution. This could be, for example, the counterparty level expected exposure or potential future exposure.

We will compare the performance of the Bayesian and classical approaches in the next section.

4 Testing Framework

We now look at the practical implementation of the framework laid out in the previous section and compare this to a traditional backtesting approach. This will demonstrate that, in a backtesting context, Bayesian techniques can be applied naturally and can provide important benefits as compared to the classical set-up.

The model we wish to backtest is a risk factor simulation model. We assume that the model distribution over each observation window is a standard normal distribution $N(0, 1)$. Our aim is then to identify misspecifications in this model with respect to a simulated data set of risk factor realizations. We will explore examples of increasing complexity to show how different aspects of the model distribution can be tested.

In our test set-up, we will consider two consecutive observation windows consisting of 50 independent observations corresponding roughly to a year of weekly returns. While 50 observations is fairly high and not always available in practical settings for long forecasting horizons, e.g. when backtesting is performed quarterly, the power of classical tests with less than 50 observations is very low which would make any comparison to the Bayesian approach less informative. The consideration of two observation windows will allow us to examine the effect of the sequential updating of priors in the Bayesian framework.

At each observation date we simulate a realization of the risk factor. We will consider different distributions for the realizations in order to test how well the backtesting can identify different types of model misspecification. The simulated risk factor returns are PIT transformed to arrive at the simulated data set $\boldsymbol{y}$. The data set is then analyzed using our Bayesian framework as well as the classical approach, which will form our benchmark comparison.

Our testing consists of two main parts. First we analyze how well the Bayesian framework performs in inferring the model parameters. We do this by assessing the means and standard deviations of the posterior distributions and comparing those to the real-world parameters in the given scenario. Parameter inference is not part of the standard classical backtesting approach and hence a direct comparison is not provided here.

In the second part we analyze how well the the Bayesian and classical approaches perform in identifying model misspecifications, i.e. in model acceptance and rejection. To do this we define Bayesian success criteria and then compare type I and type II errors when evaluating our model in the various scenarios.

4.1 Misspecification Scenarios

Our testing will be performed across the following three risk factor scenarios:

Scenario 1: Vol misspecification The real-world risk factor distribution is Gaussian with mean 0 and volatility 1.2. This is the simplest example and aims to test a typical case where model volatilities are misspecified. The volatilities are usually the main drivers of the risk factor dynamics in exposure models.

Scenario 2: Mean misspecification The real-world risk factor distribution is Gaussian with mean 0.4 and volatility 1. Here we test whether we can detect a mean level misspecification which is typically more difficult than identifying issues with the volatility. This is particularly relevant in the context of CCR backtesting as the exposure models will often be defined in a risk-neutral measure and hence the model drift will differ from the real-world one.

Scenario 3: Tail and vol misspecification The real-world risk factor distribution is generalized Student-t with mean 0, volatility 0.6 and 6 degrees of freedom. This is a more complex scenario where we explore whether we can identify the misspecification of more subtle aspects of the distribution shape. Here we consider offsetting effects of a low volatility combined with a heavy tail. This is relevant as model volatilities may be set conservatively which will mask issues with tail modeling.

4.2 Bayesian Approach

In the Bayesian approach we aim to calculate a marginal posterior distribution for all our parameters given the (simulated) real-world data. This will be analyzed in order to infer the best estimate for the model parameters and also to assess whether the model passes our chosen acceptance criteria.

To calculate the posterior we need to specify parameterized forms for the real-world distributions of our risk factor as well as priors for all parameters. Given the likelihood and prior, the posterior can then be calculated following equation (3.15). Depending on the scenario under investigation we will either choose a normal or Student-t distribution as the basis for our likelihood calculation. Initial priors for the first observation window will be uninformative or weakly-informative and centered around the model parameter values. In the subsequent observation windows we will use the previous posterior distribution to inform our new prior.

More details on the risk factor distribution parameterizations are provided below.

4.2.1 Normal Distribution

If we are mainly interested in testing the mean and volatility specification of our model then it is simplest to assume that the true generating process is normally distributed with mean μ^R and volatility σ^R.

Since the model is given by a standard normal distribution, it follows that the misspecification parameters are equal to the real-world ones, i.e. $\theta_\mu = \mu^R$ and $\theta_\sigma = \sigma^R$. The likelihood function $P(\boldsymbol{y}|\theta_\mu, \theta_\sigma)$ is hence given by equation (3.18) which can be computed in closed form.

The initial prior distributions are chosen to be weakly informative and are defined by

$$\pi(\theta_\mu) = N(m, s^2), \tag{4.1}$$

$$\pi(\theta_\sigma) = \Gamma(a_\sigma, b_\sigma), \tag{4.2}$$

where $\Gamma(a, b)$ denotes the Gamma distribution with mean a/b and variance a/b^2.

In the first observation window, the hyperparameters m, s, a and b are set as follows:

- Mean θ_μ: $m = 0$, $s = 0.2$ which implies that the mean misspecification is centered around 0 with standard deviation of 0.2.
- Volatility θ_σ: $a_\sigma = b_\sigma = 10$, which implies that the volatility misspecification is centered around 1 (i.e. no misspecification) and standard deviation is ≈ 0.32.

For subsequent observation windows we update the prior by moment matching the posterior distribution of the previous window. In figures 1a and 1b we show how the volatility and mean priors are updated over four consecutive observation windows in the case that the real-world risk factor distribution is Gaussian with mean 0 and volatility 1.2. We can see clearly how the prior becomes more informative and centered around the correct parameter values with each update.

In order to analyze the sensitivity to our prior choice we also specify the following four test priors for the volatility:

Prior A: Uninformative prior with mean 1 and standard deviation of 32.

Prior B: Weakly-informative prior with mean 1 and standard deviation of 0.32. This corresponds to our default prior.

Prior C: Stronger prior with mean 1 and standard deviation of 0.16.

Prior D: Stronger prior with mean 1.2 and standard deviation of 0.17.

4.2.2 Student-t Distribution

To test the tail behaviour of our risk factor dynamics we can use a generalized Student-t distribution with mean μ^R, volatility σ^R and degrees of freedom ν^R. Again, we know that the misspecifications are equal to the real world parameters, i.e. $\theta_\mu = \mu^R$, $\theta_\sigma = \sigma^R$ and $\theta_\nu = \nu^R$. The likelihood $P(\boldsymbol{y}|\theta_\mu, \theta_\sigma, \theta_\nu)$ is given by equation (3.26).

The prior for the mean level and volatility are as described in the previous section. Gamma prior is chosen for the degrees of freedom parameter:

$$\pi(\theta_\nu) = \Gamma(a_\nu, b_\nu) \tag{4.3}$$

with hyperparameters $a_\nu = 2$ and $b_\nu = 0.1$, which implies a mean for θ_ν of 20 and standard deviation of 14 (see Juárez and Steel [2010] for a discussion of why this is a suitable uninformative prior). For $\theta_\nu \gg 20$ the Student-t distribution is approximately Gaussian and hence close to the model specification.

As before, the posterior distribution is derived according to equation (3.15).

4.3 Benchmark Approach

In the classical benchmark approach we define a model rejection criterion based on the probability of the observed data given the model specification as described in section 2. In our testing we will use a three-bin chi-square (χ^2) test in order to analyze the data.

Given N observations of our (PIT transformed) data $\boldsymbol{y}$ the χ^2 test is based on a partition of the unit interval into a number of bins $b = \{(0, k_1], (k_1, k_2], \ldots, (k_{n-1}, 1]\}$. The T statistic is then defined as

$$\hat{T}_{\chi^2} \equiv \sum_{i=1}^{k} \frac{(O_{b_i} - E_{b_i})^2}{E_{b_i}}, \tag{4.4}$$

where O_{b_i} is the realized number of observations in the bin $b_i = [k_{i-1}, k_i]$ and $E_{b_i} = N(k_i - k_{i-1})$ is the expected number of observations given the model. It is well-known that the statistic tends to a χ^2 distribution with $k-1$ degrees of freedom, i.e. $T_{\chi^2} \sim \chi^2_{k-1}$. This allows us to calculate the p-value p which is the probability of observing T_{χ^2} values as or more extreme than those predicted under the null hypothesis $\mathbb{H}_0$ that the model is correct, i.e.:

$$p \equiv P(T_{\chi^2} \geqslant \hat{T}_{\chi^2}|\mathbb{H}_0).$$

Under the classical framework, the model is flagged if the p-value falls below the specified type I error level (a typical choice being 5%). The power of the test (i.e. the probability of flagging the model when it is in fact misspecified) is sensitive to to the size of model misspecification as well as the number of observations as can be seen in figure 2a for the case of a volatility misspecification.

The chi-square test is also sensitive to the choice of bins and there is no optimal choice for the bin width since it depends on the type of misspecification (e.g. mean versus volatility). Most reasonable choices should produce similar, but not identical, results. In figure 2b we show how the power of the benchmark test depends on the bin choice. We can see that a choice of three bins is optimal in the volatility misspecification case. Figure 2c shows how the power varies with the bin sizes in the three-bin case. We see that the choice of $[0.05, 0.95, 0.05]$ is close to optimal. This is hence the choice we use when considering scenarios one and three. When considering the mean misspecification, the optimal bin configuration has narrower bin sizes around the center of the distribution. Hence in the following we use the bin specification $[0.1, 0.8, 0.1]$ in scenario two.

5 Test Results

5.1 Inference Tests

In this set of tests we investigate to what extent our Bayesian approach is able to estimate the size of individual parameter misspecifications in our three test scenarios. In each scenario we simulate the risk factors for two subsequent observation windows with 50 independent observations each. Given the simulated data and a parameterized form for the risk factor dynamics, we calculate the marginal posterior distributions for each relevant parameter. In scenarios one and two we assume Gaussian dynamics and in scenario three we use the student-t distribution. The prior choice is as described in the previous section.

We analyze the posterior distributions by calculating the mean as well the 68% and 95% HPDs of the posterior. The further the bulk of the posterior distribution is away from the model parameter value, the more likely it is that the parameter is misspecified.

The simulated examples represent a typical output from a Bayesian backtest. To demonstrate how representative these single examples are, we also examine the distribution of the posterior mean using ten thousand simulations. We show how the sample standard mean and sample standard deviation relate to the single example posterior distribution.

5.1.1 Scenario 1. Volatility Misspecification: $\theta_\mu = 0$, $\theta_\sigma = 1.2$

The prior and posterior distributions for the volatility in the first window are shown in figure 3a. We can see that the posterior distribution is clearly shifted towards the correct level of volatility and is also more informative than the initial prior. The mean of the posterior is around 1.3, the 68% HPD interval is $[1.15, 1.4]$ and the 95% HPD interval is $[1.1, 1.55]$. We hence see in this example the actual volatility falls within the 68% HPD and the model volatility outside the 95% HPD which provides strong evidence that the model volatility is misspecified.

In figure 5a we show the result of repeating the analysis over 10k simulations. We see that the sample mean (of posterior mean estimates) is close to the sampling distribution volatility of 1.2. We also see that the standard deviation of the sample distribution is close to the average of the standard deviations of the posterior distributions (with the proximity increasing in the second window). This suggests that the posterior distribution provides a reliable inference of the volatility as well as the uncertainty around the estimation.

In figures 3b and 5b we see how the posterior evolves in the second window where the prior is taken to be the (moment-matched) posterior of the first window. We can see clearly that the uncertainty in the volatility estimate has reduced with the 68% HPD interval now being $[1.25, 1.4]$.

We repeat this analysis for the estimation of the mean which is not misspecified in this scenario. Figures 3c and 3d show the prior and posterior distributions for the first and second window examples, and the histograms of the sample distribution of the posterior means over 10k simulations are shown in figures 5c and 5d. We see that the mean is correctly estimated at 0 with a sample standard deviation of 0.07 in the second window.

In figure 4 we examine the sensitivity of the posterior distribution to the prior choice for the volatility parameter misspecification. We consider three sets of 50 risk factor simulations drawn from the normal distribution with volatility misspecification size 1.2. In particular, we choose a sample (a) with high standard deviation (1.35); (b) with standard deviation in line with the risk factor volatility (1.2) and (c) with low standard deviation (1.15). Unsurprisingly, the uninformative prior (A) tends to adapt quickly to the observed evidence and results in estimates very close to the standard deviations of the samples. The strong priors produce narrow posteriors with mean levels which remain closer to the initial prior means. The weekly informative default prior provides a balance between these extremes in terms of weighting the data and prior knowledge.

We conclude that our Bayesian approach is able to estimate the misspecification of both the mean and volatility together with the associated uncertainty in this scenario.

5.1.2 Scenario 2. Mean misspecification: $\theta_\mu = 0.4$, $\theta_\sigma = 1.0$

In this scenario the risk factors still follow a Gaussian distribution. However it is now the mean that is misspecified and not the volatility.

The prior and posterior distributions for the volatility in the representative sample are shown 6a and 6b. We note that, in contrast to scenario one, the posterior mean is now closer to the actual level $\sigma^R = 1$, particularly in the second window. This is seen more clearly when considering the distribution of the posterior mean over 10k simulations as shown in figures 7a and 7b. The sample mean in the second window is given by 1.03. We also note that the sample standard deviation is close to the average standard deviation of the posterior across the 10k simulations, which demonstrates that the posterior distribution provides a good estimate of the parameter level as well as the estimation uncertainty.

The prior and posterior distributions for the mean are given in figures 6c and 6d. We see that the posterior shows a clear shift towards the mean level of the sampling distribution $\mu^R = 0.4$.

The histograms of the posterior means are shown in figures 7c and 7d. We observe that the sample mean moves closer to the true value over consecutive observation windows. In the second window the sample mean is 0.31+/- 0.08 and hence the true value is at just around one standard deviation from the sample mean. We expect this to improve in subsequent observation windows. More data is needed for an accurate parameter estimation since the initial prior is centered at $\mu = 0$ with a standard deviation of 0.2. This means that the model misspecification is large with respect to the prior confidence and hence more information will be needed to shift the posterior. Finally, we note that, as before, the sample standard deviation and the average of the posterior standard deviations are close.

5.1.3 Scenario 3. Tail and volatility misspecification: $\theta_\mu = 0$, $\theta_\sigma = 0.6$, $\theta_\nu = 6$

We now analyze the more complex scenario where the risk factor is distributed according to a Student-t distribution with mean 0, volatility 0.6 and 6 degrees of freedom. The actual volatility is thus low compared to the model specification, however, the tails of the distribution are heavier. These are two compensating effects which are typically hard to disentangle.

In figures 8a and 8b we see the posterior distributions for the volatility misspecification. The mean of the posterior is close the the correct level of 0.6 even in the first observation window.

Figures 8c and 8d show the priors and posteriors for the degrees of freedom. This is more difficult to estimate (partially due to the choice of uninformative prior) but we see that in the second window the posterior mean has shifted significantly and is close to the correct level of 6.

This result demonstrates that the Bayesian approach is able to identify the source of a model misspecification even in the presence of multiple effects. Another way to see this is by analyzing the posterior distributions for the expected exposure and potential future exposure. Here we assume that the risk factor distributions represent the value (MtM) of a portfolio. A priori it is not clear whether the model misspecification will result in a lower or higher EE or PFE than what the real-world distribution would predict.

In figure 9 we see that the Bayesian approach is able to clearly identify that that the actual EE and PFE should be lower than what is implied by the model. The inference becomes stronger in the second window, as expected, and the model EE and PFE lie outside the 95% HPD. We also note that the misspecification has a stronger impact on the PFE which is to be expected given that the tail of the distribution is directly impacted via both volatility and degrees of freedom parameters in the scenario misspecification.

5.2 Power Tests

To benchmark the performance of Bayesian framework against classical null hypothesis significance testing, we consider a simulation study across two observation windows with the same set-up as before: 50 observations per window, $10k$ simulations in scenarios one and two and $2K$ simulations in scenario three. In each of our three misspecification scenarios we will then compare the rates of model rejection and acceptance given suitable performance criteria.

Flagging model misspecifications requires an assessment of the accuracy and precision of our testing procedure, which commonly relies on the concepts of the type I and type II errors as well as the test power. The type I error corresponds to flagging a misspecification when the model is not misspecified whereas the type II error corresponds to not flagging a misspecification when the model is misspecified. The power is defined as the complement of the type II error, i.e. the probability of correctly flagging a misspecification when the model is misspecified. Stricter flagging criteria will naturally lead to a higher power at the expense of a higher type I error. The challenge for any testing framework is hence to increase power while keeping the type I error fixed.

5.2.1 Classical Power

In the classical approach a model is flagged if the chi-square p-value falls below a chosen threshold. In the limit of infinite repeated experiments, the type I error is equal to this flagging threshold. However, in our analysis we have a finite set of simulations in each scenario and the equivalence does not hold exactly. Hence, for a set of simulated risk factors, we calibrate the p-value threshold so that the flagging rate for a correctly specified model is equal to the target type I error rate.

Given a p-value threshold and a set of risk factor simulations, the test power is determined by the flagging rate of the misspecified model. The type II error is one minus the power. We note that in the classical approach we can only flag the entire model in aggregate. There is no ability to identify the misspecification of individual parameters.

When analyzing the test power in the second observation window we need to choose how data from the first window is used. In the classical case we consider two approaches. In the first approach we estimate the p-value in the second window on a stand-alone basis without taking any previous data into account. This corresponds most closely to what would happen in a typical backtesting set-up. The model is then flagged overall if it is flagged in either the first or the second window. In the second approach we calculate a single p-value for the entire data across both windows consisting of a total of 100 observations. This makes use of the maximum information available in the simulated data and is hence expected to lead to a higher power.

5.2.2 Bayesian Power

While under the classical framework the concept of power is well-established, the same does not exist in Bayesian analysis. However, in order to compare approaches, we can define equivalent performance criteria as described below.

We first look at how we can flag the misspecification of individual parameters which is possible in the Bayesian approach. For each parameter we set an acceptable error tolerance level ϵ for the misspecification. Given a posterior distribution for the parameter, we can calculate the probability p_ϵ that the misspecification lies in the the range $[\bar{\theta} - \epsilon, \bar{\theta} + \epsilon]$ where $\bar{\theta}$ corresponds to no misspecification. We can now flag a parameter misspecification if p_ϵ falls below a desired confidence threshold which in the following we fix at 95%.

With this criterion we can define type I and type II error rates as usual. For a given set of simulated risk factors we will calibrate to a chosen type I error rate target by adjusting the error range ϵ.

While we do consider the power of flagging each parameter separately, for a fair comparison with the power of the classical hypothesis test we also define an aggregate power measure. To do this we first note that a model is misspecified in aggregate if at least one parameter is misspecified. We deem a test to have correctly flagged an overall misspecification if at least one misspecified parameter is flagged and no correct parameter is flagged.

Different combinations of individual parameter tolerance levels can result in the the same type I error and there is hence no unique calibration. For a given type I error we will thus choose the tolerance level combination that leads to the maximum power[3].

When considering the second observation window we use the prior updating described in previous sections, i.e. the prior of the second window is the (moment-matched) posterior of the first one for each model parameter. Hence we are making use of the entire set of data across both windows.

In the following we present a comparison of the Bayesian and classical approaches in our three misspecification scenarios.

5.2.3 Scenario 1. Volatility Misspecification: $\theta_\mu = 0$, $\theta_\sigma = 1.2$

In figures 10a and 10b we show how the type I error and power depend on the tolerance interval for both the risk factor volatility and mean. We note that the type I error for the volatility and the mean decreases as the interval increases, as expected. A 5% type I error corresponds roughly to a tolerance level of 0.34 for the volatility and 0.39 for the mean in the first window. In the second window this reduces to 0.25 for the volatility and 0.31 for the mean.

[3] Given that the number of tolerance intervals is finite, this result in non-continuous power function. To overcome this we a) calculate a rolling average of the power across a small interval of type I error $(+/-0.01)$ and b) use a cubic spline interpolation with knots every 0.02 steps (with not-a-knot boundary condition for the first and last two polynomials).

The power for the identification of the volatility misspecification also decreases as the tolerance increases. Again, this reflects the fact that a larger interval implies that we flag an incorrect model less often and hence increase the type II error.

Given that the mean is correctly specified in this scenario, the power of testing the mean misspecification is defined as the probability of not flagging the mean. This increases with increasing interval size.

In order to compare the Bayesian and classical approaches we plot the power of the chi-square test as well as the Bayesian backtest as a function of the type I error in figures 10c and 10d.

We see clearly that the Bayesian estimation outperforms the classical one. In the first window the power for a 5% type I error rate is around 37% in the classical case but 58% and 97% for the Bayesian volatility and mean estimation, respectively, and 60% for both parameters combined. In the second window the Bayesian power is increased further to 80% (volatility), 98% (mean) and 80% (both parameters combined). Using the combined data for classical testing (100 observations) leads to an increase in power to 55% which is expected given that the power of hypothesis testing should increase with the amount of data available. We note, however, that it would not be feasible in practice to use ever increasing sets of data whereas prior updating is always possible.

Finally, we note that the power of the combined flagging in the Bayesian case is close to the power of the flagging of the volatility on its own. This is because the power is defined as flagging the volatility and not flagging the mean. The power of the mean test however is very high, i.e. the probability of not flagging the correctly specified mean is close to 1.

We also consider the impact of prior choice on the power. We see in figure 11 that the power as a function of type I error is fairly close across all priors and hence sensitivity to the prior choice is low overall. As expected, the informative prior (D), which is centered on the misspecified value, results in a slightly higher power.

In summary, we observe that the Bayesian approach not only has greater power but is also able to evaluate the accuracy of individual model parameters which is not possible in the classical approach.

5.2.4 Scenario 2. Mean misspecification: $\theta_\mu = 0.4$, $\theta_\sigma = 1.0$

Here we compare the power in flagging a misspecification of the mean level. In this scenario the volatility is correctly specified. The power as a function of the tolerance level is shown in figures 12a and 12b for windows one and two respectively. As before, we see that the type I error and power decreases with increasing interval size. To achieve a 5% type I error rate we need an error tolerance for the mean of about 0.39 in the first window and 0.31 in the second. Given that volatility is specified correctly, we we find that, as expected, the power of the volatility test increases with increasing tolerance level.

The power of our test as a function of the type I error for both the Bayesian and classical approach is given in figures 12c and 12d. We note that overall the power for mean misspecification flagging is higher than in the previous case where the volatility was misspecified. This reflects the fact that the misspecification of the mean in scenario two is larger than the misspecification of the volatility in scenario one, as reflected in a higher Kullback-Leibler divergence for the former (80 vs 30)[4].

The power of the Bayesian approach is notably higher than in the classical case in both observation windows. At 5% type I error the Bayesian power for mean estimation is at 77% in the first window and over 95% in the second. The power for the volatility test (which is correctly specified) is also high at around 92% in the second window. As observed in scenario one, the aggregate power of model flagging in the Bayesian case is close to the power of the mean flagging on its own and hence over 95% in the second window. This should be compared to the classical power of 83% for the combined data case and 51% for the split case which are both significantly lower.

Hence we conclude that the Bayesian approach outperforms the classical one in the scenario two also.

5.2.5 Scenario 3. Tail and volatility misspecification: $\theta_\mu = 0$, $\theta_\sigma = 0.6$, $\theta_\nu = 6$

In our final scenario we investigate the ability of backtesting to identify misspecifications of the tails of our risk factor distribution. The analysis is based on 2000 simulations from a Student-t distribution. The model distribution, which is standard Gaussian, can be thought of a Student-t distribution with $\nu^M >> 20$ degrees of freedom. The tolerance interval we use for the degrees of freedom is directional, i.e. the interval for model parameter ν^M with tolerance ϵ is $[20+\epsilon, \infty)$.

We first examine how the type I error varies with the tolerance for the degrees of freedom. This is shown in figures 13a and 13b for windows one and two. We see that for a type I error of 5% we need a tolerance of $\epsilon = -15$ in the first window and $\epsilon = -14$ in the second. These ϵ levels results in rather wide tolerance intervals which can be attributed to the use of uninformative prior for degrees of freedom parameter, thus making posterior convergence to the misspecification value slow - see figure 1c for the convergence towards 20 degrees of freedom in four windows that indicates little impact on the width of the prior/posterior distribution in the second window.

The power in the Bayesian and classical case is shown in figures 13c and 13d. The plots are less smooth than in the previous cases due to the fact that we use a reduced set of samples in this scenario leading to increased Monte Carlo noise. Nevertheless, we can draw several conclusions. Firstly, we note that the power of the classical test is lower in this scenario than in previous ones. At 5% type I error it is close to 0 in the first window and only increases by using combined data in the second window to around 40% which indicates that the test is not satisfactory. This is to be expected given the fact that we have two competing effects: a thinner tail but higher volatility in the model vs. the real-world distribution.

In the Bayesian case we see that the power of identifying the misspecification in the degrees of freedom on its own is also low at 30% in the first window and just

[4] The Kullback-Leibler divergence, also known as the relative entropy, is a common measure of the distance between two distributions. It is a generalization of the squared distance and aims to quantify the difference in information between two distributions.

above 40% in the second for a type I error of 5%. However the power of identifying the volatility misspecification and the mean non-misspecification remain very high. Given that in this scenario we accept the model as being correctly flagged if either volatility or degrees of freedom misspecification is flagged and mean misspecification is not flagged, this results in a high power in flagging the model overall which is at 95% in the first window and close to 100% in the second.

This is a remarkable result and shows that the Bayesian approach is able to distinguish the effect of volatility and the degrees-of-freedom. Hence in this case the Bayesian approach is superior both qualitatively and quantitatively to the classical one.

6 Conclusion

In this article we have introduced a general Bayesian framework for backtesting counterparty risk models. This has clear benefits in terms of performance in comparison to the standard classical approach while also having a firm conceptual basis and being straightforward to implement.

The main outputs of our framework are the posterior distributions of the parameter misspecification in our model conditional on the observed data. These directly encode the probability of the model being correctly specified which should be the focus of backtesting. By contrast, in classical hypothesis testing we can only determine the probability of the observed data conditional on a chosen model specification being correct and do not have any evidence for model correctness itself.

A key distinguishing feature of Bayesian statistics is the need to specify prior distributions of the model parameters. These encode any prior knowledge that we may have regarding the level and uncertainty for a given parameter and can be chosen naturally in a backtesting context. Priors can be based on expert judgment (e.g. regarding parameter liquidity) but, more importantly, they will reflect the information gathered in previous testing cycles. Thus the Bayesian framework incorporates all the information gathered over the course of the ongoing backtesting process.

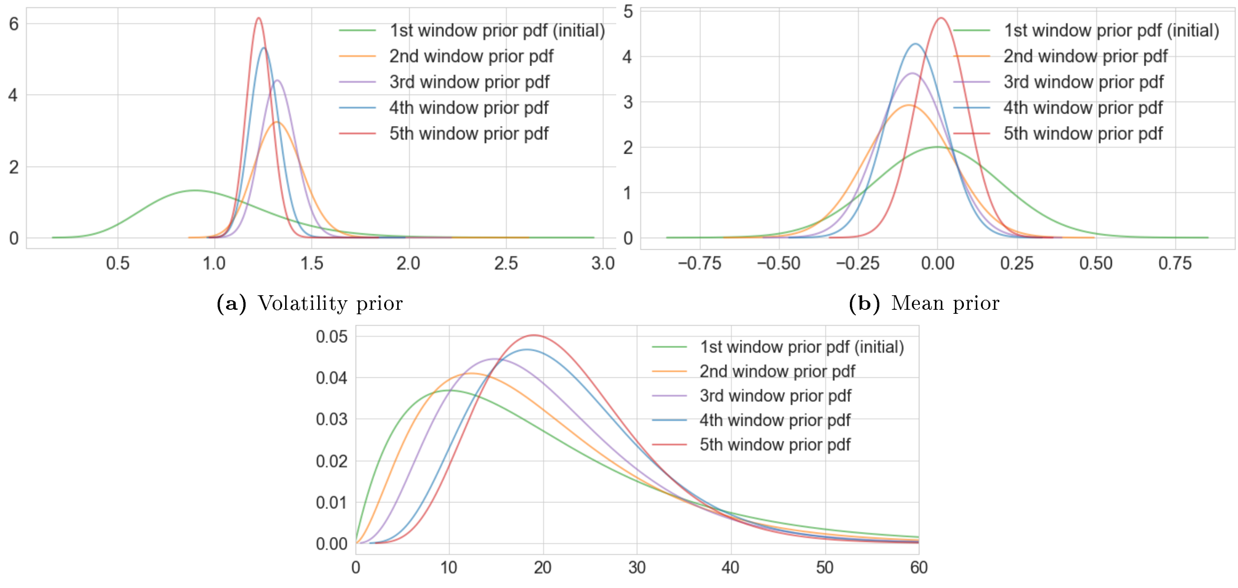
An important feature of our Bayesian approach is that we are able to analyze the source of model issues by estimating all relevant model parameter misspecifications given the data and our prior knowledge. Thus we not only can determine whether a model is correctly specified in aggregate, as is done in the classical case, but also which aspects of the model may be misspecified. This allows for rapid diagnosis and remediation of problems flagged in backtesting.

The fact that we can directly estimate the level of an individual parameter misspecification also allows us to express the results of the backtesting in terms of the end-usage metrics of the model, e.g. expected exposure or potential future exposure. This means that we can assign loss estimates to any identified issues which allows for an economic assessment of their impact and an optimal prioritization of any follow-up action.

In a direct comparison of our framework with a classical backtesting set-up, we have shown that the Bayesian approach has greater power in rejecting or accepting a given model. This is a key success criterion of any backtesting framework and is particularly important in the counterparty risk context as the lack of observational data means that the power of testing is typically low in the classical case.

The outperformance of the Bayesian framework is due to several factors. Firstly, we fully take any prior knowledge into consideration via the prior distributions of the parameters. This will augment the information gathered through multiple observation windows. Secondly, the Bayesian approach is able to disentangle the affect of different parameters on the model output which may be offsetting. This allows us to more accurately determine overall model performance.

Finally, we note that whilst we designed the Bayesian approach with counterparty risk backtesting in mind, the framework is far more general. In particular, the framework can be adapted to the backtesting of any predictive time-series or forecasting model that is sensitive to distributional assumptions. It is thus well suited to be a key tool at the heart of risk modeling.

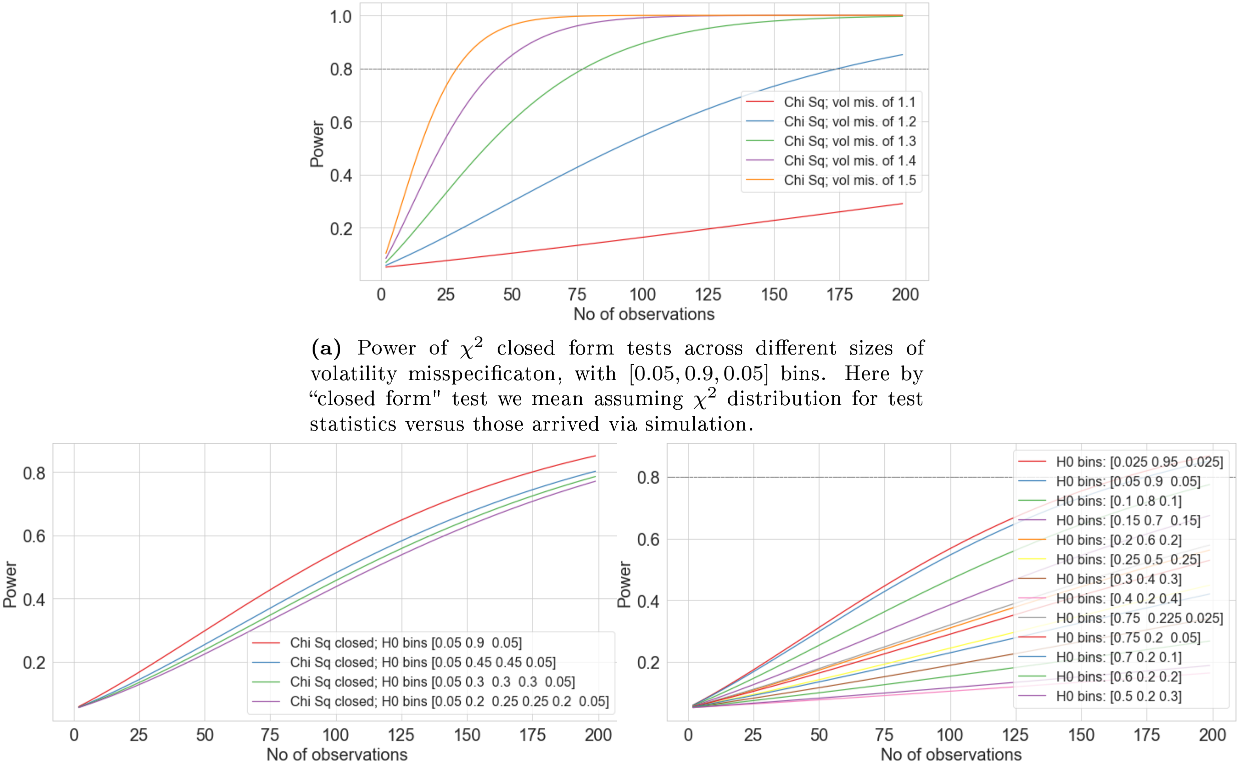


(a) Volatility prior

(b) Mean prior

(c) Degrees of freedom prior

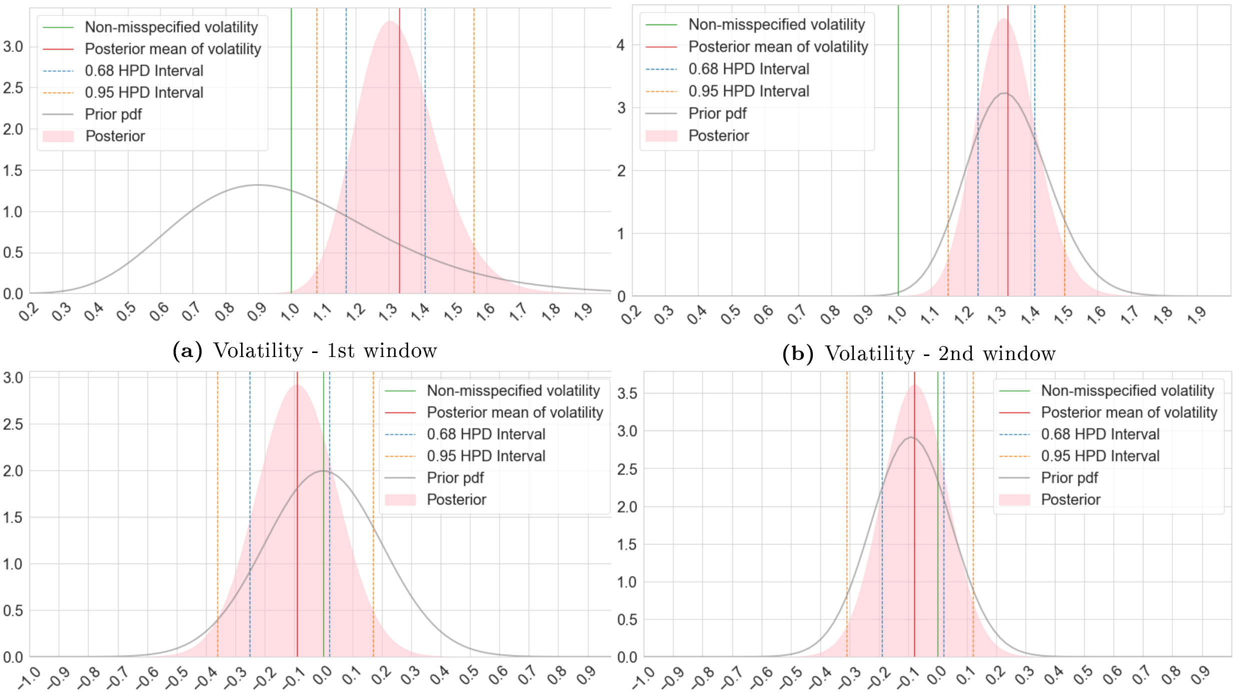
Figure 1: The impact of prior updating given four quarters with 50 observations (annual frequency) each where volatility is misspecified by 1.2. Note the shrinkage of the priors as well as convergence towards the true misspecification level.



(a) Power of χ^2 closed form tests across different sizes of volatility misspecificaton, with $[0.05, 0.9, 0.05]$ bins. Here by “closed form" test we mean assuming χ^2 distribution for test statistics versus those arrived via simulation.

(b) Power of χ^2 closed form tests across different number of bins

(c) Power of χ^2 closed form tests for different bin size specifications for 3 bins



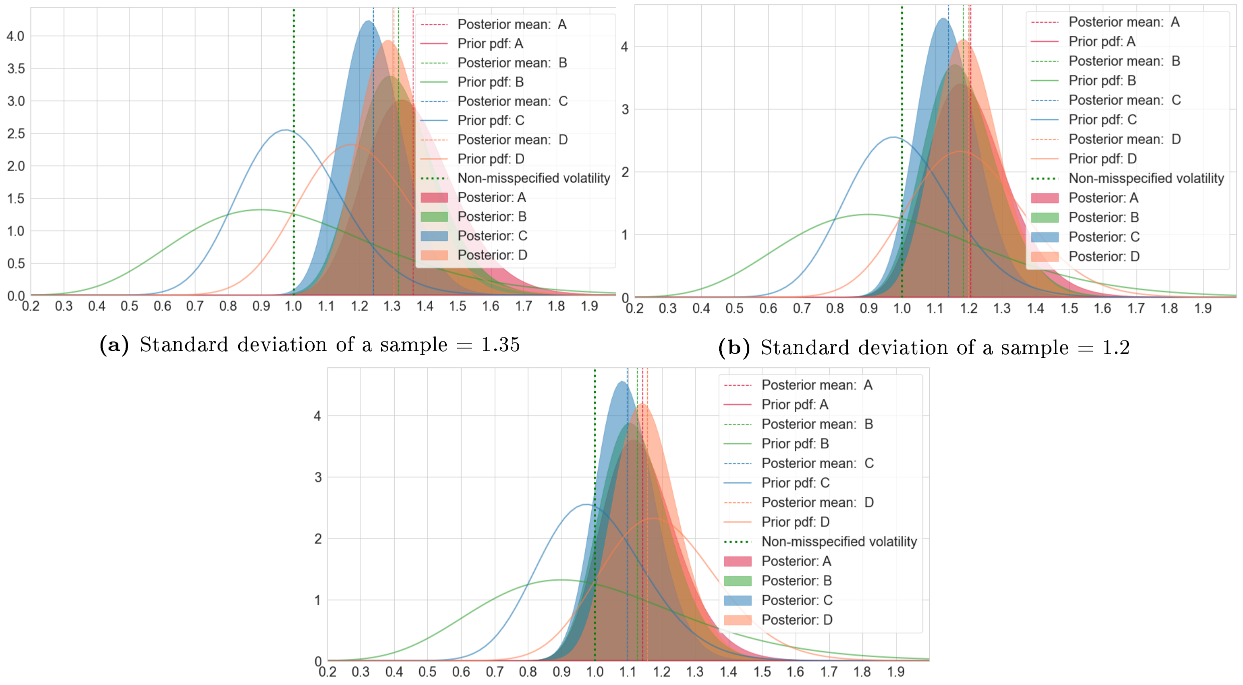
(a) Volatility - 1st window

(b) Volatility - 2nd window

(c) Mean - 1st window

(d) Mean - 2nd window

Figure 3: Inference for parameters - scenario 1: volatility misspecification

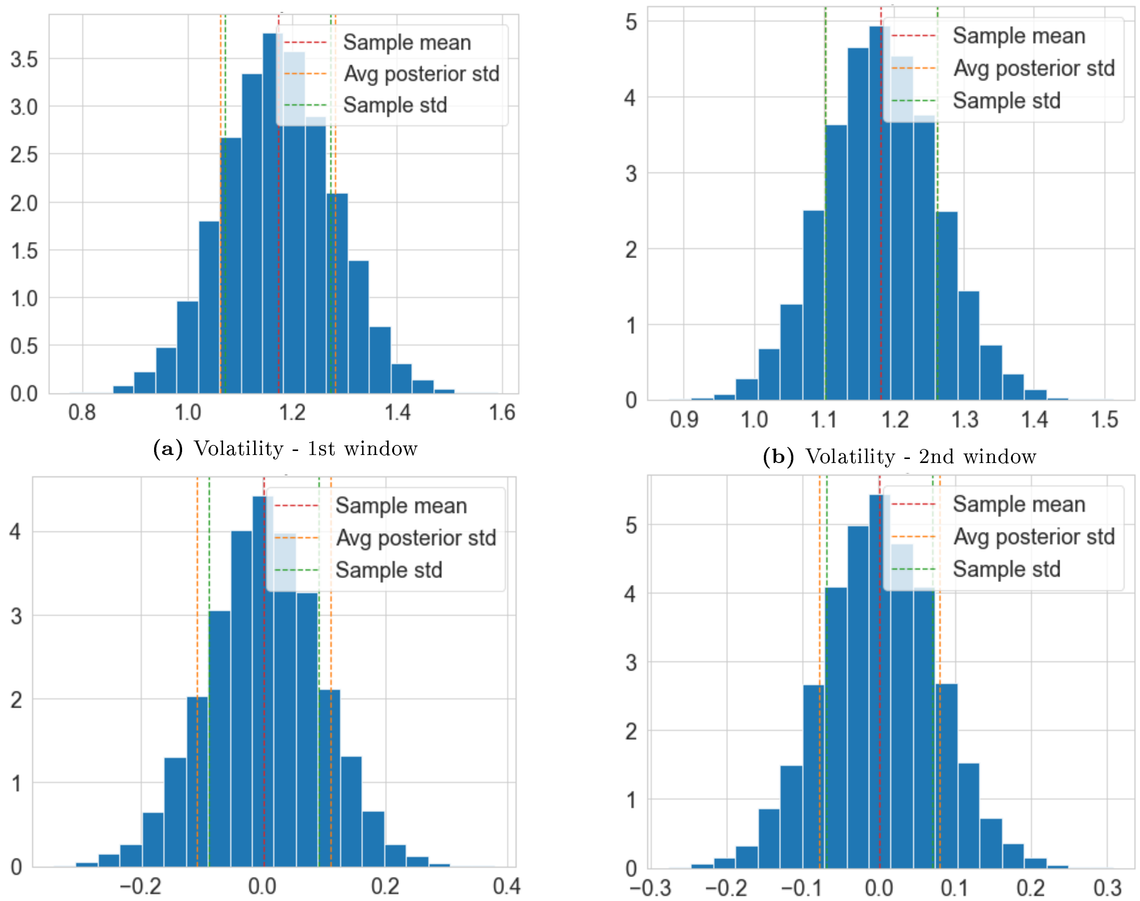


(a) Standard deviation of a sample = 1.35

(b) Standard deviation of a sample = 1.2

(c) Standard deviation of a sample = 1.15

Figure 4: The impact of prior choices for inference where volatility is misspecified by 1.2 for varying sample standard deviation levels (50 observations). Priors: A - mean=1, std=32, B - mean=1, std=0.32, C - mean=1, std=0.16, D - mean=1.2, std=0.17.



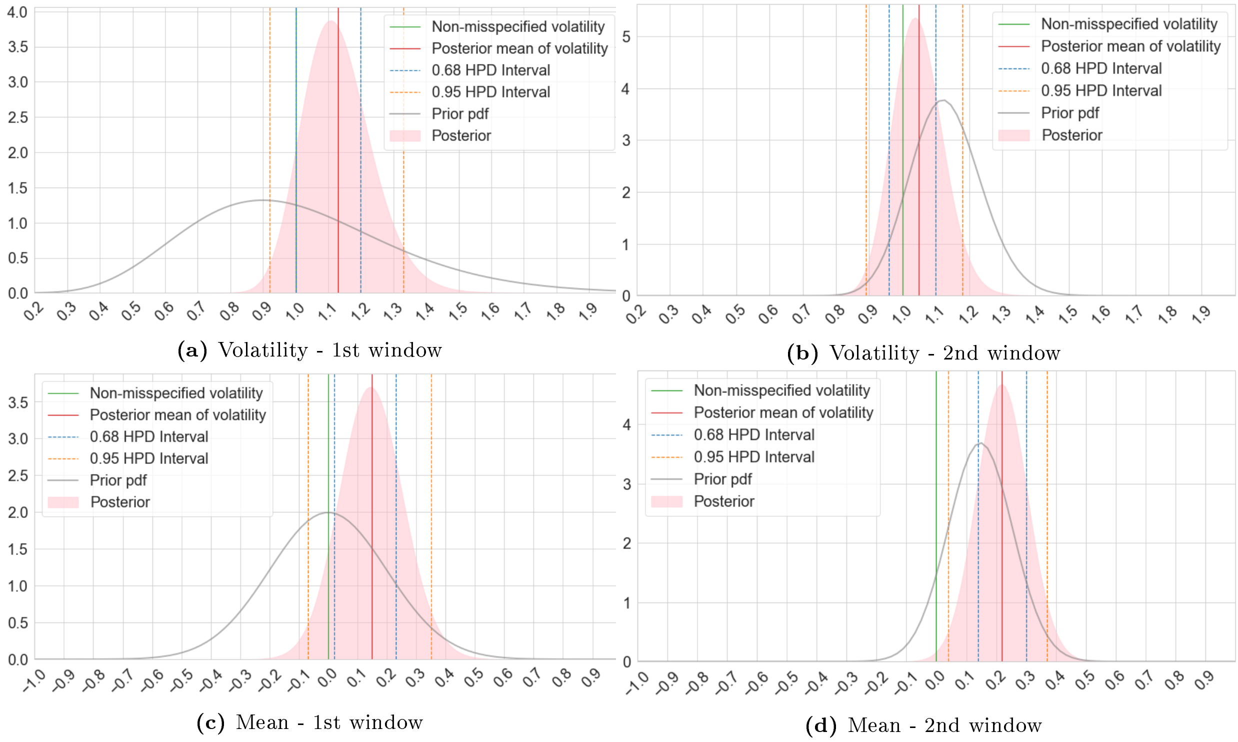
(a) Volatility - 1st window

(b) Volatility - 2nd window

(c) Mean - 1st window

(d) Mean - 2nd window

Figure 5: Posterior std - scenario 1: volatility misspecification



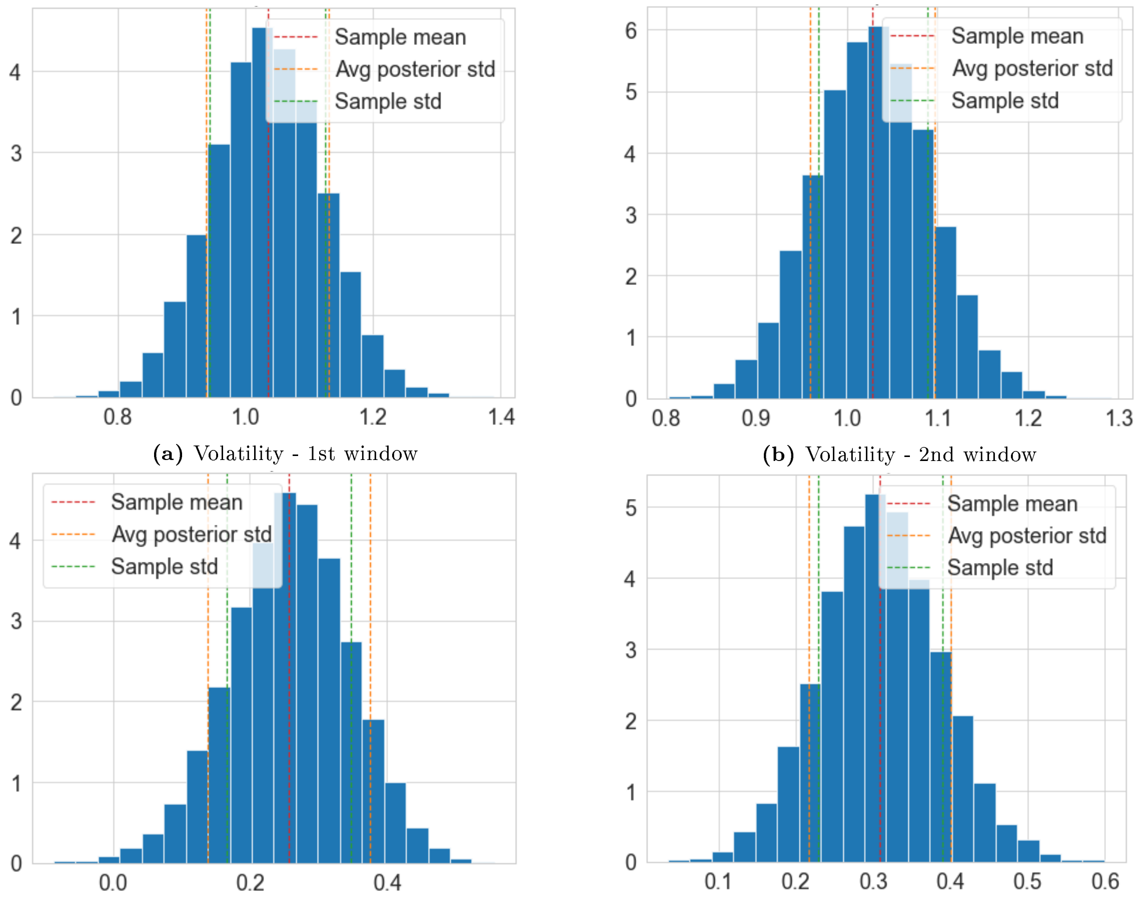
(a) Volatility - 1st window

(b) Volatility - 2nd window

(c) Mean - 1st window

(d) Mean - 2nd window

Figure 6: Inference for parameters - scenario 2: mean misspecification



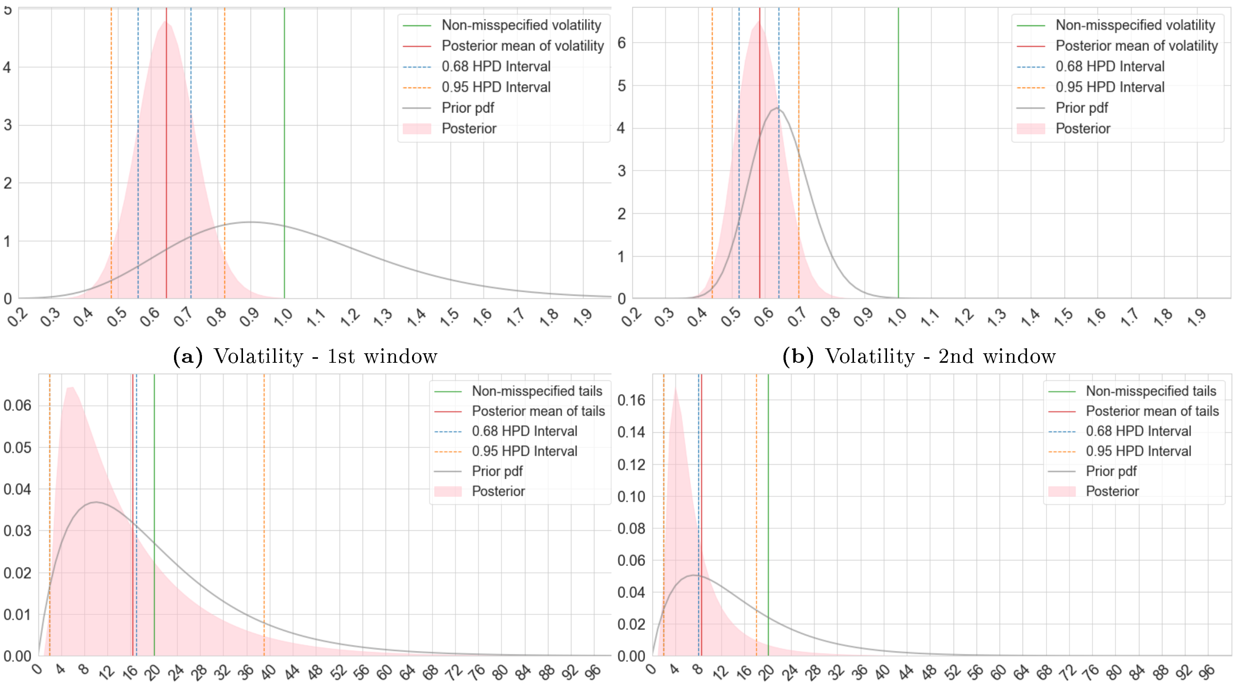
(a) Volatility - 1st window

(b) Volatility - 2nd window

(c) Mean - 1st window

(d) Mean - 2nd window

Figure 7: Posterior std - scenario 2: mean misspecification



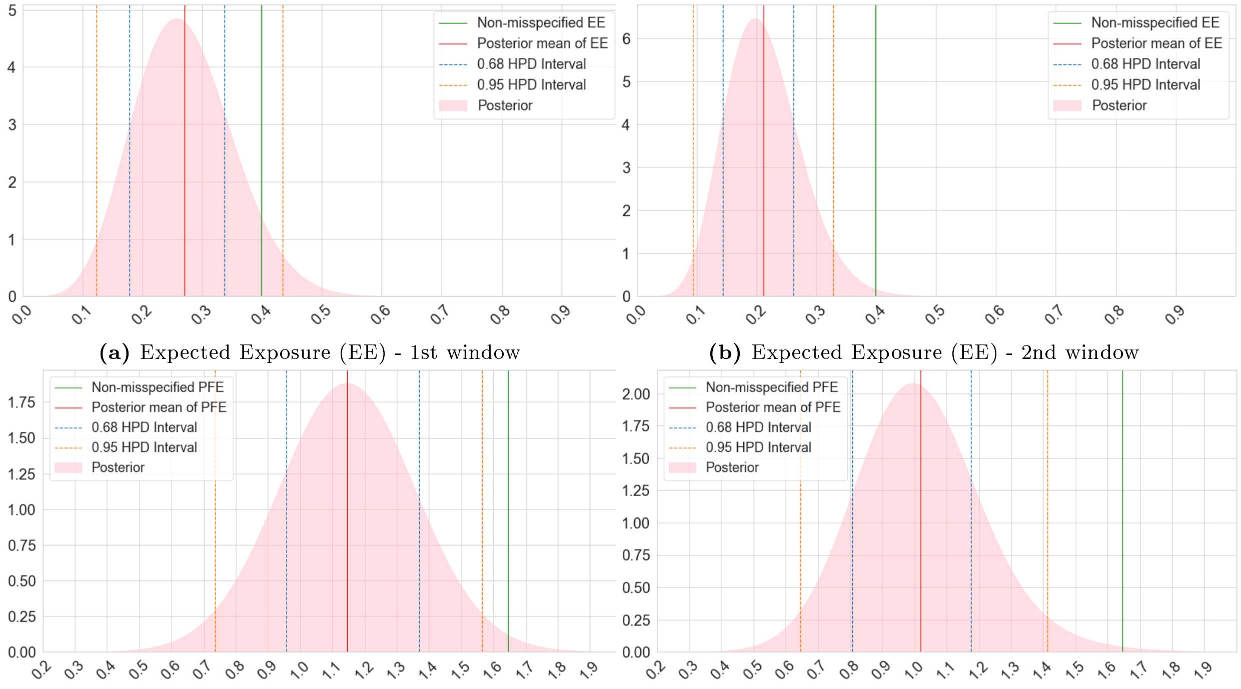
(a) Volatility - 1st window

(b) Volatility - 2nd window

(c) Tail (degrees of freedom) - 1st window

(d) Tail (degrees of freedom) - 2nd window

Figure 8: Inference for parameters - scenario 3: volatility and tail misspecification



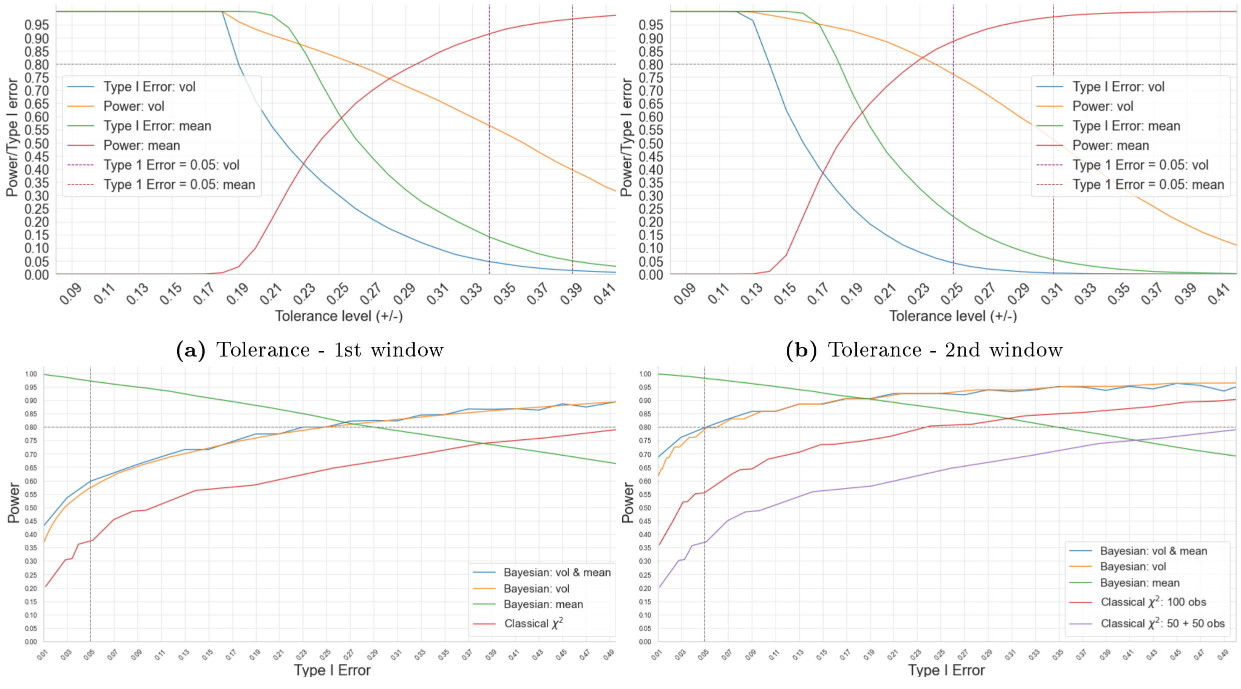
(a) Expected Exposure (EE) - 1st window

(b) Expected Exposure (EE) - 2nd window

(c) Potential Future Exposure (PFE) - 1st window

(d) Potential Future Exposure (PFE) - 2nd window

Figure 9: Inference for exposure metrics - scenario 3: volatility and tail misspecification



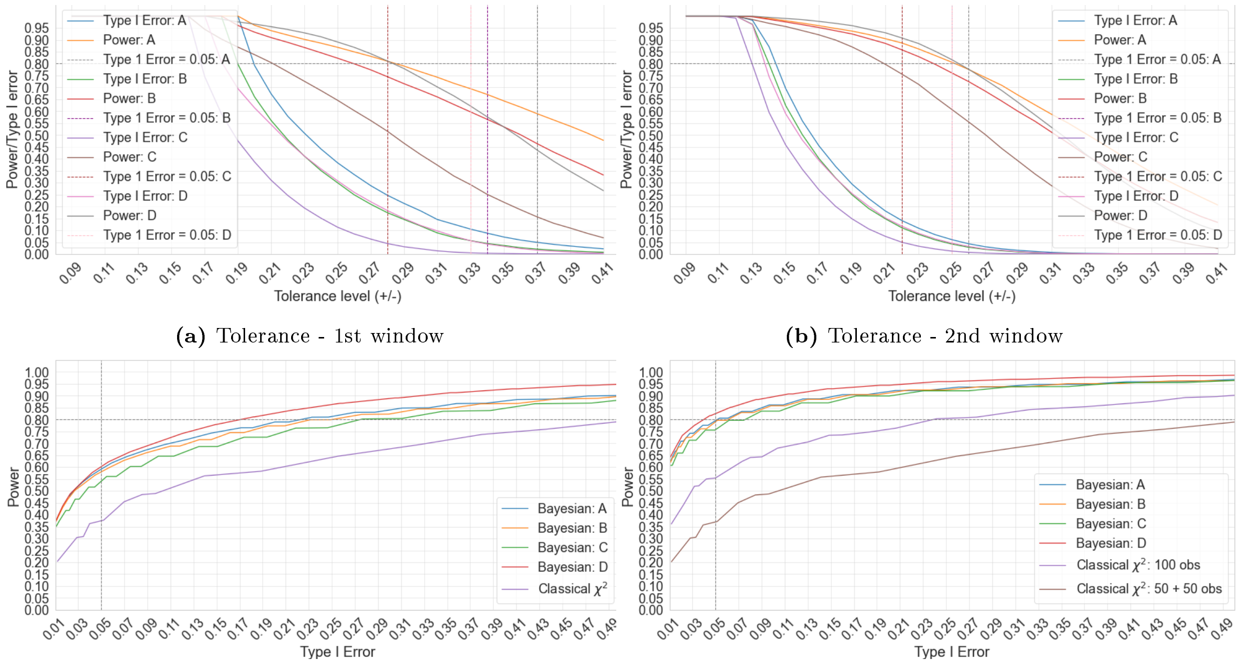
(a) Tolerance - 1st window

(b) Tolerance - 2nd window

(c) Power - 1st window

(d) Power - 2nd window

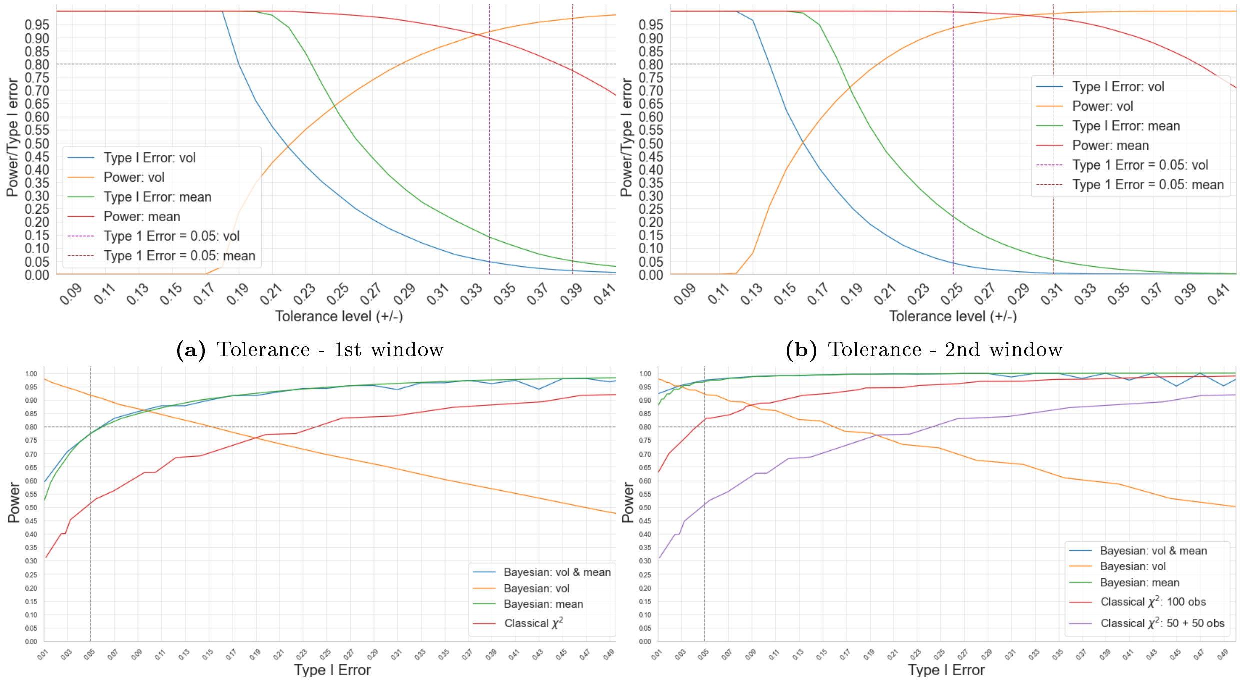
Figure 10: Power - scenario 1: volatility misspecification



(a) Tolerance - 1st window **(b)** Tolerance - 2nd window

(c) Power - 1st window **(d)** Power - 2nd window

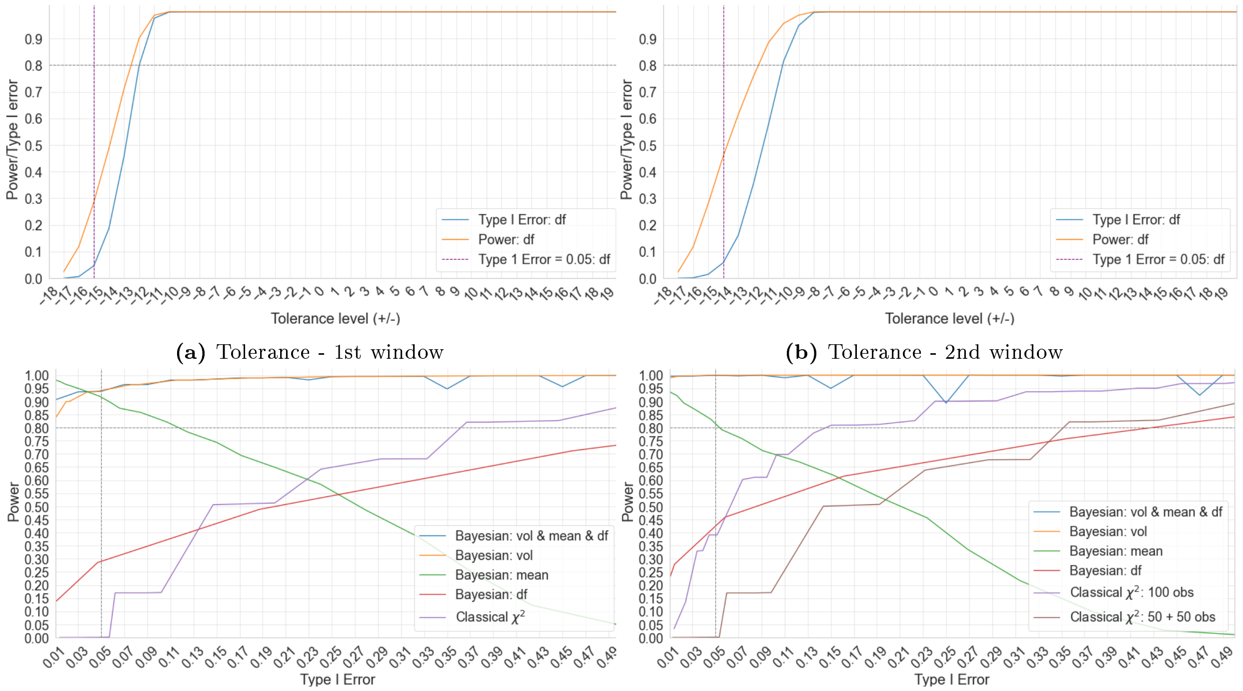
Figure 11: The impact of prior choices for power where volatility is misspecified by 1.2 (50 observations). Priors: A - mean=1, std=32, B - mean=1, std=0.32, C - mean=1, std=0.16, D - mean=1.2, std=0.17.



(a) Tolerance - 1st window **(b)** Tolerance - 2nd window

(c) Power - 1st window **(d)** Power - 2nd window

Figure 12: Power - scenario 2: mean misspecification



(a) Tolerance - 1st window

(b) Tolerance - 2nd window

(c) Power - 1st window

(d) Power - 2nd window

Figure 13: Power - scenario 3: tail and volatility misspecification

References

Anfuso, F., D. Karyampas, and A. Nawroth (2014). A Sound Basel III compliant framework for backtesting Credit Exposure Models. *Risk, August*.

BIS (1996). Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements. Technical report, Basel Committee on Banking Supervision.

BIS (2010). Sound practices for backtesting counterparty credit risk model. Technical report, Basel Committee on Banking Supervision.

Clayton, M. A. (2019). Backtesting Volatility Assumptions using Overlapping Observations. *Available at SSRN 3342541*.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, 863–883.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance 72*(4), 1399–1440.

Juárez, M. A. and M. F. Steel (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics 28*(1), 52–66.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*(3), 293–337.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics 23*(3), 470–472.

Ruiz, I. (2014). Backtesting counterparty risk: how good is your model? *Journal of Credit Risk 10*(1).

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a World Beyond "p $<$ 0.05". *The American Statistician 73*(sup1), 1–19.