EUROPEAN CENTRAL BANK

EUROSYSTEM

Working Paper Series

F. Coppens, M. Mayer, L. Millischer, F. Resch, S. Sauer and K. Schulze

Advances in multivariate back-testing for credit risk underestimation

No 1885 / February 2016

Abstract

When back-testing the calibration quality of rating systems two-sided statistical tests can detect over- and underestimation of credit risk. Some users though, such as risk-averse investors and regulators, are primarily interested in the underestimation of risk only, and thus require one-sided tests. The established one-sided tests are multiple tests, which assess each rating class of the rating system separately and then combine the results to an overall assessment. However, these multiple tests may fail to detect underperformance of the whole rating system. Aiming to improve the overall assessment of rating systems, this paper presents a set of one-sided tests, which assess the performance of all rating classes jointly. These joint tests build on the method of Sterne [1954] for ranking possible outcomes by probability, which allows to extend back-testing to a setting of multiple rating classes. The new joint tests are compared to the most established one-sided multiple test and are further shown to outperform this benchmark in terms of power and size of the acceptance region.

Keywords: credit ratings; probability of default; back-testing; one-sided tests; minP approach; Sterne test;

JEL classes: C12, C52, G21, G24

Non-technical summary

Well-performing credit assessment systems play an important role in contributing to an efficient and stable financial system. They produce adequate credit ratings of, e.g., sovereigns, companies, or specific financial instruments.

The importance of ratings necessitates a regular evaluation of their quality, for which several approaches exist in the literature. The comparison of estimated with actually observed numbers of defaults within a credit assessment system, called *back-testing*, is the most wide-spread method. Most of the statistical tests used for back-testing are 'two-sided' because they consider over- and underestimation of credit risk. Such tests are relevant for example for banks, because both sides imply financial losses for banks, either from greater than expected losses on granted loans or from missed business opportunities and higher capital charges. In contrast, risk-averse investors or regulators tend to embrace a 'one-sided' perspective in that they focus on detecting underestimation of credit risk, but are more or less indifferent with respect to credit risk overestimation. Hence, they consider ratings as appropriate only if a rated entity's estimated probability of default does not indicate a better credit quality of the entity than its actual payment behaviour. These users of ratings require 'one-sided' tests that are sensitive only to credit risk underestimation and have a greater power than two-sided tests to identify well-performing systems.

The key contribution of this paper is a set of novel one-sided statistical tests that allow the assessment of credit assessment systems from a holistic perspective. In particular, the proposed tests assess the quality of all rating grades jointly, instead of the existing straightforward, but often less powerful, approach to assess each rating grade independently and then to combine the results.

We show that our novel joint tests have greater probability to identify miscalibrated credit assessment systems than the most established one-sided test, i.e. they have a greater statistical power. Our tests outperform the existing test also in terms of the number of possible observations of defaults that lead to the conclusion that a credit assessment system is not well-performing. This is an innovative performance criterion which is intuitively beneficial when little is known about the true probabilities of debtors' defaults produced by a miscalibrated credit assessment system, which is usually a realistic situation. However, the increased performance of our novel tests comes at the expense of varying degrees of increased implementation complexity and computation time, so that the user can choose her optimal combination of statistical performance and ease of implementation from our set of novel tests.

Our novel tests may also be useful in other areas of applied statistical analysis, such as medical science, as the usually limited sample sizes in these areas are more in line with our statistical assumptions than with the assumptions in the existing literature.

1 Introduction

Assessing the credit quality of debtors is a key task of the financial sector in order to enable an efficient allocation of credit and to ensure the stability of any individual financial institution as well as that of the whole financial system. To this end credit assessment systems use quantitative and qualitative information to produce estimates of debtors' creditworthiness, also called *ratings*. These ratings are typically associated with probabilities of default which are used not least for pricing, risk management and regulatory purposes.

Being in the core of financial intermediation, ratings necessitate a regular evaluation of their quality. Several approaches exist for validating the quality of rating systems (for an overview see Basel Committee on Banking Supervision [2005]): the most widespread method is back-testing the calibration of a rating system by comparing ex post-realised default rates with ex ante-estimates of probabilities. Other approaches include tests of the discriminatory power (see Lingo and Winkler [2008]) or the comparison of ratings from different sources, called benchmarking (see Hornik et al. [2007]). The focus of this paper is on back-testing the calibration quality.

Poor calibration of credit assessments may result in either an overestimation or an underestimation of credit risk. Both situations can be associated with financial risks for the user: the underestimation of credit risks can lead to explicit financial losses because more debtors than expected will default on average. The overestimation of credit risks can lead to missed business opportunities, e.g. because competitors with better credit assessment systems will be able to provide more attractive offers to potential creditworthy borrowers. Jankowitsch et al. [2007] and Blöchlinger and Leippold [2006] demonstrate the impact of miscalibrated ratings systems on the profitability of banks. Hence, the existing back-testing literature, as summarised e.g. in Basel Committee on Banking Supervision [2005], focuses on statistical tests that are 'two-sided', i.e. tests that detect both over- and underestimation of credit risk.

However, many users of credit assessment systems are negatively affected only by an underestimation of credit risk, whereas they do not suffer from an overestimation of credit risk. To the contrary, these users may even appreciate fewer than predicted defaults, as they typically imply lower credit losses on a portfolio. This applies in particular to third-party users of banks' internal ratings-based systems (IRBs): for example, institutions taking assets as collateral that were assessed by IRBs, such as some central banks including the ECB, do not suffer themselves from lost business opportunities following an underestimation of risk by the IRB.[1] Also banking supervisors may have a tendency to be more interested in the underestimation of credit risk than its overestimation for financial stability reasons. Credit rating agencies use information from IRBs for the assessment of asset-backed securities and may be particularly concerned about lower than expected quality of the underlying assets.[2] Furthermore, any bond investor who relies on external ratings would suffer from an underestimation of risk but might even profit from an overestimation as this might result in a lower bond price but would on average not lead to the predicted losses.

The key contribution of this paper is a set of novel one-sided statistical

[1]The underestimation may lead to a better than expected financial risk protection for the collateral takers.

[2]See, e.g. DBRS [2013], Fitch [2014], Moody's [2015], and Standard & Poor's [2013].

tests for users of credit assessment systems that are primarily interested in a potential underestimation of risks. Such one-sided tests are common for single-dimensional problems such as the classic 'Lady tasting tea' experiment by Fisher [1935] or trials testing for positive effects of a drug against placebos (e.g. Fisher [1991]). However, credit assessment systems usually allocate debtors to several rating classes, so that multi-dimensional tests are needed to test the whole system. Similar situations may arise in clinical trials where different doses or multiple points in time of the treatment are tested upon (see e.g. Dmitrienko and Hsu [2004]). So far, the literature (e.g., Döhler [2010]) has focused on tests that identify an underestimation of the whole system by assessing each rating class independently and then combining the results. These so-called 'multiple tests' are discussed in Section 3.2 below.

While multiple tests have their merits if the user is interested in the performance of the individual rating classes, these tests may fail to identify miscalibrations which are not significant for any rating class but lead to an underestimation of the PD when considering all classes jointly. The existing literature on this question includes only two-sided joint tests, such as the approaches by Hosmer and Lemeshow [1980] or Aussenegg et al. [2011]. In the latter paper the authors suggest a multivariate version of the Sterne [1954] test. This Sterne test is an exact test and can replicate a number of established approximate tests including the test of Hosmer and Lemeshow [1980]. The basic idea behind the Sterne test is to rank all possible outcomes, i.e. debtors' defaults in the case of credit ratings, by their probability of occurrence in increasing order. Starting with the lowest probability, outcomes are assigned to the rejection region of the test until the pre-defined significance level of the test is exploited.

In order to extend this idea also to one-sided tests, this paper introduces three novel one-sided versions of the joint Sterne test in Section 4: it begins with the theoretically optimal version that maximizes the number of rejected outcomes among one-sided tests. However, this test is computationally very demanding, so that two computationally more feasible alternatives are proposed: first a one-sided iterative Sterne test that assigns the outcomes with the lowest probability to the rejection region step by step. Second, a one-sided test that maximizes the size of the rejection region among one-sided tests containing a two-sided Sterne test. In addition, Section 4.3 discusses an enhanced version of the multiple test that shares important features of joint tests and is computationally more efficient than the Sterne-based tests.

In order to perform the comparison of the different tests, we measure performance in the traditional way by the probabilities to identify a poorly calibrated rating system, the so-called *power* of the tests, and in a more innovative way by the relative size of the acceptance region of the tests. Furthermore, the computation time required for the different tests is reported in order to assess how easily they can be implemented. The comparison of the tests shows that the one-sided joint Sterne tests perform best in terms of power for most of the studied scenarios and for all studied scenarios in terms of acceptance region size. However, this additional performance comes at the expense of a significantly increased computational complexity.

The following Section 2 sets out the probabilistic framework and the statistical hypotheses studied in this paper. Section 3 discusses two-sided joint and one-sided multiple tests which have been established in the literature. The most established one-sided test serves as a benchmark for the novel one-sided tests that jointly assess the performance quality of several rating classes and are introduced in Section 4. Section 5 compares the costs and benefits of these novel joint tests against the benchmark. Section 6 concludes.

2 Theoretical Framework

This section introduces the notation and probabilistic framework of the paper. We apply standard notation building on Aussenegg et al. [2011]. Our statistical model is given by:

$$(\Omega, \mathcal{F}, \mathbf{P}_p : p \in \mathcal{P}) = \left(\bigtimes_{c=1}^{C} \Omega_c, \bigtimes_{c=1}^{C} \mathcal{F}_c, \mathcal{B}_C(\mathbf{p}) : \mathbf{p} \in [0,1]^C \right), \tag{1}$$

where $\mathcal{B}_C(\mathbf{p})$ denotes the C-variate binomial-distribution and $\mathbf{p} = (p_1, ..., p_C)$ is its vector of Bernoulli-probabilities. Furthermore $\bigtimes_{c=1}^{C} \Omega_c$ denotes the product sample space of the individual sample spaces $\Omega_c := \{0, ..., n_c\}$ containing observable defaults in rating class c and the σ-algebra $\mathcal{F}_c$ is the power set of the sample space, i.e. $\mathcal{F}_c := 2^{\Omega_c}$.

Our model captures rating systems with a finite number of C rating classes and a finite number of n_c obligors in each rating class c. Assuming independence between default events within and across rating classes $c = 1, ..., C$ is in line with most of the existing literature, see e.g. Frey and McNeil [2003]. Under this independence assumption the probability mass function of a default pattern $\mathbf{D} = (D_1, ..., D_C)$ with realized value $\mathbf{d} = (d_1, ..., d_C)$ under the measure $\mathbf{P}_\mathbf{p}$ is the product of the binomial marginal distributions:

$$\mathbf{P}_\mathbf{p}(\mathbf{D} = \mathbf{d}) = \mathcal{B}_C(\mathbf{d}; \mathbf{n}, \mathbf{p}) = \mathbf{P}_\mathbf{p}(D_1 = d_1, ..., D_C = d_C) =$$
$$= \prod_{c=1}^{C} \mathbf{P}_{p_c}(D_c = d_c) = \prod_{c=1}^{C} \mathcal{B}_1(d_c; n_c, p_c) = \prod_{c=1}^{C} \binom{n_c}{d_c} p_c^{d_c} (1 - p_c)^{n_c - d_c},$$

where $\mathbf{p} = (p_1, ..., p_C)$ denotes the vector of true and latent default probabilities and $\mathbf{n} = (n_1, ..., n_C)$ denotes the vector of sizes of sample spaces.

Consistent with Krahnen and Weber [2001] we define a rating system as a function: $R : \{companies\} \to \{ratingclasses\}$. This means that a rating system R assigns each element of a set of companies to a rating class, denoted for example by $\{A, B+, B, B-, ...\}$. The assignment of companies to rating classes is based on an ex-ante estimated default probability denoted by $\hat{\mathbf{p}} = (\hat{p}_1, ..., \hat{p}_C)$ and ensures that all obligors within a rating class are reasonably homogeneous with respect to their estimated probability of default (PD).[3]

As we aim to find statistical evidence in favor or against the calibration quality of a rating system we next turn to hypothesis testing. In our statistical model a hypothesis can be formulated as the subset of the parameter space $\mathcal{P}$ on which the hypothesis holds true. Thus, we separate the parameter space $\mathcal{P}$

[3]Some rating systems produce a continuum of PDs and assign an interval of PDs to a rating class. This situation may warrant adjustments of the existing and new tests discussed in this paper; these adjustments are not the focus of the present paper. Furthermore, $\hat{\mathbf{p}}$ refers to the predicted probability of default of a rating system, which does not need to be the same as the average realised default rate over a certain sample in the model development or testing.

into the null hypothesis H_0 and the alternative H_1, where for the latter the hypothesis does not hold true. Thus, we have $\mathcal{P} = H_0 \cup H_1$ and $H_0 \cap H_1 = \varnothing$. We will formulate two hypotheses regarding the rating system's performance. From the perspective of loss avoidance, we call a rating system as *well-performing*, if there is no under-estimation of credit risk, i.e. if the true probabilities of default are equal to or less than the predicted default probabilities. In contrast, a rating system is *under-performing* if the true probability of default exceeds the predicted one in at least one rating class. This is formulated by the following *one-sided composite* (null and alternative) *hypotheses*:

$$H_0: \quad \forall c \in C: \quad p_c \leqslant \hat{p_c}. \tag{2}$$

$$H_1: \quad \exists c \in C: \quad p_c > \hat{p_c}.$$

From the perspective of calibration quality, we call a rating system *well-calibrated*, if there is no under- or over-estimation of credit risk, i.e. if the true default probabilities are equal to the predicted default probabilities. This is formulated by following *two-sided composite* (null and alternative) *hypotheses*:

$$H_0: \quad \forall c \in C: \quad p_c = \hat{p_c}, \tag{3}$$

$$H_1: \quad \exists c \in C: \quad p_c \neq \hat{p_c}.$$

For both, the one-sided and the two-sided hypothesis, we can draw two conclusions: either to reject the null hypothesis or not to reject it. In either case the conclusion is based on the observed default pattern $\mathbf{d} \in \Omega$ and derived by a statistical test. Formally, the *test* is a random variable

$$\phi: \quad (\Omega, \mathcal{F}) \quad \longrightarrow \quad (\{0,1\}, 2^{\{0,1\}}),$$

which maps the observation $\mathbf{d} \in \Omega$ to the probability that the test concludes to reject the null hypothesis under this observation. Here we consider only non-randomized tests, as it simplifies notation and randomized tests are typically not applied in testing credit rating systems and do not seem to add significant value. The observation space Ω can be separated into observations yielding a rejection of the null hypothesis and observations not doing so. We denote the *rejection region* of a test ϕ by $\mathcal{R}_\phi$ and it is given by $\mathcal{R}_\phi := \{\mathbf{d} \in \Omega \mid \phi(\mathbf{d}) = 1\}$. Analogously the acceptance region of a test ϕ is given by $\mathcal{A}_\phi := \{\mathbf{d} \in \Omega \mid \phi(\mathbf{d}) = 0\}$. As a consequence, it holds that $\Omega = \mathcal{R} \cup \mathcal{A}$. Hence, a non-randomized test can equivalently be defined in terms of its rejection or acceptance region:

$$\phi(\mathbf{d}) = \mathbf{1}_{\mathcal{R}_\phi}(\mathbf{d}) = 1 - \mathbf{1}_{\mathcal{A}_\phi}(\mathbf{d}),$$

where $\mathbf{1}$ denotes the indicator function.

The two-sided hypothesis of well-calibration is usually, but not necessarily, tested by a two-sided test. The one-sided hypothesis of well-performance is typically tested by a one-sided test. A test is called *one-sided*, if it holds for its acceptance region $\mathcal{A}$ for all $c = 1, ..., C$

$$\mathbf{d} \in \mathcal{A}, \ \mathbf{d} - e_c \in \Omega \implies \mathbf{d} - e_c \in \mathcal{A},$$

where $e_c := (0, .., 1, ..., 0)^\top$ denotes the c-th unit-vector. In words, if a one-sided test does not reject a default pattern, then it does not reject a default pattern with fewer or equal defaults in each rating class either.

We next turn to type I and type II errors of a test. For $\mathbf{p} \in H_0$ the probability of a type I error of a test ϕ, i.e. the probability to reject the null hypothesis H_0 even though it is true, is given by:

$$\mathrm{E}_{\mathbf{p}}\phi = \sum_{\mathbf{d} \in \mathcal{R}_\phi} \mathbf{P}_{\mathbf{p}}(\mathbf{d}), \quad \mathbf{p} \in H_0.$$

A test is of significance *level* α, if it holds

$$\mathrm{E}_{\mathbf{p}}\phi \leqslant \alpha \quad \forall\, \mathbf{p} \in H_0.$$

Analogously, for $\mathbf{p} \in H_1$ the probability of a type II error of a test ϕ is given by:

$$\mathrm{E}_{\mathbf{p}}(1-\phi) = 1 - \sum_{\mathbf{d} \in \mathcal{R}_\phi} \mathbf{P}_{\mathbf{p}}(\mathbf{d}) = \sum_{\mathbf{d} \in \mathcal{A}_\phi} \mathbf{P}_{\mathbf{p}}(\mathbf{d}), \quad \mathbf{p} \in H_1.$$

The probability that the test ϕ rejects the null hypothesis H_0 for $\mathbf{p} \in H_1$, i.e. in case the alternative is true, is called the *power* of the test and it is given by:

$$\mathrm{E}_{\mathbf{p}}\phi = \sum_{\mathbf{d} \in \mathcal{R}_\phi} \mathbf{P}_{\mathbf{p}}(\mathbf{d}), \quad \mathbf{p} \in H_1.$$

The ideal test would have a power of 1 for all $\mathbf{p} \in H_1$ and a power of 0 for all $\mathbf{p} \in H_0$ and hence it would have a zero-probability for type I and II errors. However, usually such a test does not exist as $\mathbf{p}$ cannot be observed, and hence the test must infer information about the true $\mathbf{p}$ from the observation $\mathbf{d}$ which was generated by some $\mathbf{P}_{\mathbf{p}}$. Consequently there is a probability of incorrect decisions made by the test and there is a trade-off between minimizing the probability of type I and type II errors. In this respect it is standard practice to bound the type I error and then minimize the type II error, i.e. to restrict to tests of a certain level and then to choose the test with the highest power on the alternative hypothesis.

3 A Review of Established Tests

3.1 Existing Tests for Single Rating Classes

Many of the procedures for testing the calibration quality of rating systems that are proposed by the academic literature or applied in practice are designed for a single rating class, see for example Coppens et al. [2007] and Basel Committee on Banking Supervision [2005]. When testing the one-sided hypothesis of equation (2) for the performance of a single rating class it follows from the Neyman-Pearson lemma[4] that the *one-sided binomial test* is the uniformly most powerful test, i.e. it has the highest power among all non-randomized tests of the same level α for any given alternative H_1. When testing a single rating class c the acceptance region of the one-sided binomial test for level α is given by all observations not contained in the upper α-quantile of the distribution:

$$\mathcal{A}^c_{\mathcal{B}} := \left\{ d_c \in \Omega_c \,\middle|\, \sum_{i=0}^{d_c-1} \mathcal{B}_1(i; n_c, \widehat{p}_c) < 1 - \alpha \right\}.$$

[4] See for example DeGroot and Schervish [2002].

As regards tests of the two-sided hypothesis of equation (3) for single rating classes, there is a number of approximate tests relying on normal approximations as well as a few exact tests that have been proposed for calibration quality testing. In the following we will concentrate only on those which lay the foundation for section 4.

The "gold standard" for confidence intervals for binomial distributions is defined by Clopper and Pearson [1934]. By inverting this procedure a calibration quality test for the two-sided hypothesis in equation (3) is obtained where the rejection region is "symmetric" around the median in the sense that the acceptance region lies between the lower and upper $\alpha/2$ quantiles of the distribution:

$$\mathcal{A}^c_{CP} := \left\{ d_c \in \Omega_c \middle| \frac{\alpha}{2} < \sum_{i=0}^{d_c} \mathcal{B}_1(i; n_c, \widehat{p}_c), \sum_{i=0}^{d_c-1} \mathcal{B}_1(i; n_c, \widehat{p}_c) < 1 - \frac{\alpha}{2} \right\}.$$

However, Reiczigel [2003] shows that the definition of confidence intervals by Sterne [1954] is preferable to that of Clopper and Pearson [1934]. The associated Sterne test aims at finding a minimal acceptance region, i.e. an acceptance region containing the lowest number of default patterns possible under level α. The acceptance region of the Sterne test can be constructed by starting with the outcome with the highest probability of occurrence and then adding the outcome with the next highest probability of occurrence until the sum of all probabilities outside the acceptance region is as close as possible and just above $1 - \alpha$. The acceptance region of the *Sterne* test is defined as:

$$\mathcal{A}^c_{Sterne} := \left\{ d_c \in \Omega_c \middle| \sum_{i \in \Omega_c:\ \mathcal{B}_1(i; n_c, \widehat{p}_c) \leqslant \mathcal{B}_1(d_c; n_c, \widehat{p}_c)} \mathcal{B}_1(i; n_c, \widehat{p}_c) > \alpha \right\}.$$

In general, this acceptance region need not be a connected set. For uni-modal distributions such as the binomial distribution, however, the acceptance region is two-sided and it becomes one-sided for highly skewed distributions.

Whereas the Sterne test is an exact test, it can also be linked to numerous approximate tests. In the approximate test the binomial distribution is approximated, typically by variants of the normal distribution.

When approximating the binomial distribution by a normal distribution, applying the Sterne method gives the *Score test*. Also, a variant of the Score test which addresses issues stemming from the discreteness of the binomial can easily be obtained by applying a continuity correction. Finally the *Wald test* and its modification by Agresti and Coull [1998] can also be replicated. The details and precise connections of these approximate tests to the exact Sterne test are presented in Vollset [1993].

3.2 Existing Multiple Tests for Rating Systems

Tests for single risk classes can be used directly to test the performance quality of a whole rating system. This involves the combination of the individual test results, typically derived by the one-sided binomial test, using an appropriate multiple-testing procedure in order to test the one-sided composite hypothesis in equation (2). An equivalent approach can also be followed for the two-sided composite hypothesis in equation (3).

This section gives an overview of multiple-testing procedures that are applicable in the context of rating system validation.[5]

Defining the individual null-hypotheses that rating class c is *well-performing* by:

$$H_0^c: \quad p_c \leqslant \hat{p_c},$$

the hypothesis in equation (2) for testing the performance of a whole rating system can also be written in terms of the individual null-hypotheses:

$$H_0 = H_0^1 \cap ... \cap H_0^C.$$

Hence, we will reject the global null hypothesis H_0 and conclude that a rating system is underperforming if we can reject the individual null hypothesis H_0^c for at least one rating class c. As a consequence, making a type-I error for the global null hypothesis is equivalent to falsely rejecting at least one of the individual null hypotheses. The probability to reject the global null hypothesis even though it is true, is also referred to as the family-wise error rate (FWER).

3.2.1 The multiple-testing problem

A statistical phenomenon that occurs when conducting multiple tests is the so called *multiple-testing problem* or *alpha inflation*.[6] In order to sketch this phenomenon consider the following example. Suppose that a rating system consists of seven rating classes and that for each rating class $c = 1, ..., 7$ the individual null hypothesis H_0^c is tested at a significance level of 5%. If all individual null hypotheses are true then the probability to falsely reject at least one of them is given by $1-(1-0.05)^7 = 0.30$. Hence, even though each individual hypothesis is tested at a significance level of 5% the type-I error probability for the global hypothesis amounts to 30%. In the context of the validation of credit rating systems it means that even if all rating classes are perfectly calibrated the probability to falsely conclude that the rating system is underperforming can be substantially higher than the significance level chosen for the tests of individual rating classes. In order to control the FWER when conducting multiple tests one can either decrease the significance levels for the individual tests (in the example discussed above this would imply that each individual null hypothesis is tested at a significance level of $1-(1-0.05)^{1/7} = 0.007$) or, equivalently, adjust the p-values of the individual tests upwards. The literature mainly follows the latter approach.

3.2.2 Multiple-testing procedures

In order to address the multiple-testing problem as outlined in the previous section the literature has developed a number of procedures aimed at controlling the FWER. In this section we will shortly review two of them: the classic Bonferroni adjustment and the min-P approach by Westfall and Wolfinger [1997]. For a more comprehensive discussion of multiple-testing procedures in the context of the validation of credit rating systems we also refer to Döhler [2010].

[5] A more detailed description of multiple-testing procedures can be found for example in Döhler [2010].

[6] See for example Lehmann and Romano [2006].

In general, multiple-testing procedures aim at controlling the probability of one or more false rejections, i.e. the FWER, at a multiple significance level α. The methods discussed below guarantee *strong control of the FWER* meaning that it holds that $FWER \leqslant \alpha$ for all possible constellations of true and false hypotheses. In the following let pv_c, $c = 1, ..., C$, denote the observed p-value corresponding to the individual null hypotheses H_0^c.

The classic Bonferroni adjustment maintains strong control of the FWER by adjusting p-values according to: $pv_c^{'} = \min(C \cdot pv_c, 1)$, where $pv_c^{'}$ denotes the adjusted p-value for rating class c. This is equivalent to increasing the significance level α as described at the end of section 3.2.1. Hence, all individual hypotheses with $pv_c^{'} \leqslant \alpha$ are rejected and the global null hypothesis is rejected if $\min(pv_1^{'}, ..., pv_C^{'}) \leqslant \alpha$. The Bonferroni method is derived from Boole's inequality, it does not require independence between default events, and it is the most conservative method among the multiple-testing procedures discussed in the literature in the sense that it makes the strongest p-value adjustments.

Westfall and Wolfinger [1997] point out that the classic Bonferroni method as well as more recent multiple-testing procedures that build on this approach can be especially conservative[7] when the p-values follow a discrete distribution. The latter is the case in our problem setting where, as discussed above, the number of defaults in a rating class follows a binomial distribution. Westfall and Wolfinger [1997] argue that power improvements can be gained by taking into account the discreteness of the distribution of test statistics and they suggest the min-P approach where adjusted p-values are computed as:

$$pv_c^{'} := P(\min(PV_1, ..., PV_C) \leqslant pv_c).$$

Here PV_c denotes the p-value of rating class c considered as a random variable. Hence, the adjusted p-value for rating class c is the probability that the minimum p-value is smaller than the observed p-value for rating class c. If default events are assumed to be independent, as we did in section 2, the adjusted p-values can be calculated as:

$$pv_c^{'} = 1 - [1 - P(PV_1 \leqslant pv_c)] \cdot ... \cdot [1 - P(PV_C \leqslant pv_c)] \, .$$

As for the Bonferroni method, all individual hypotheses with $pv_c^{'} \leqslant \alpha$ are rejected and the global null hypothesis is rejected if $\min(pv_1^{'}, ..., pv_C^{'}) \leqslant \alpha$.

3.2.3 The Benchmark: Multiple Test

One example for the application of the min-P approach in the context of the validation of credit rating systems is the multiple test which is used by central banks in the euro area to validate their in-house credit assessment systems. This section outlines the multiple test which will be used as a benchmark for the novel one-sided joint back-testing procedures that we introduce below.

In the following we explicitly highlight the dependence of p-values on the number of defaults: $PV_c = PV_c(d_c)$. The p-value for testing rating class c is given by:

$$PV_c(d_c) = 1 - F^{(0)}(d_c - 1),$$

where $F^{(0)}(x)$ denotes the CDF of a binomial distribution with parameters $\mathcal{B}(x; n_c, \widehat{p}_c)$ under the assumption that H_0^c is true. Under our assumption of

[7] Conservative in the statistical sense, i.e. leading to a small type-I error.

inter-class independence of default events (see Section 2), the min-P adjusted p-values are given by:

$$PV_c^{'}(d_c) = F^{minP}(PV_c(d_c)),$$

where $F^{minP}(x) = 1 - [1 - P(PV_1 \leqslant x)] \cdot ... \cdot [1 - P(PV_C \leqslant x)]$. Hence, the min-P adjusted p-values can also be written as:

$$PV_c^{'}(d_c) = F^{minP}(1 - F^{(0)}(d_c - 1)).$$

Note that the adjusted p-values $pv_c^{'}$ depend negatively on the number of defaults in rating class c as well as positively on the number of rating classes C and the PDs under H_0^c, i.e. $\widehat{p}_c$, of all rating classes $c = 1, ..., C$ included in the multiple test. Finally, we define the acceptance region $\mathcal{A}_{mult}$ for testing the global null hypothesis H_0:

$$\mathcal{A}_{mult} = \Big\{ \boldsymbol{d} = (d_1, \ldots, d_C) \in \Omega \mid \forall\, d_c : \;\; PV_c^{'}(d_c) > \alpha \Big\}. \tag{4}$$

The figures below illustrate the case of two rating classes, taking two rating classes with 90 issuers each as an example;the PDs under H_0 are set to 32% and 35% in the first and second class, respectively. Figure 1 depicts the bivariate binomial probability distribution for observed defaults under H_0. Figure 2 represents the acceptance and rejection regions of the multiple test.

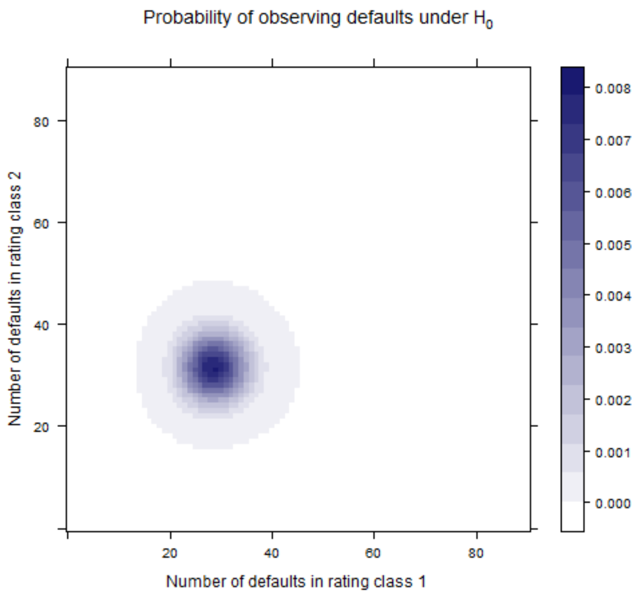

Figure 1: Probability distribution of observed defaults under H_0 in a scenario with 90 issuers in each of two rating classes and PDs under H_0 set to 32% and 35% respectively.

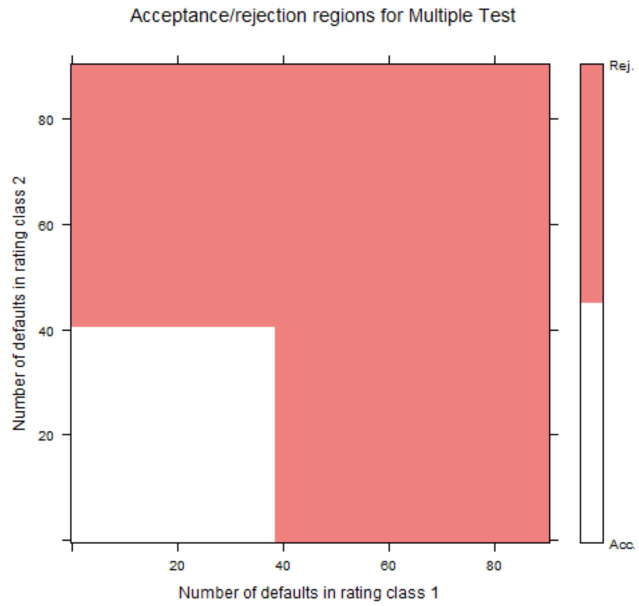

Figure 2: Observation space split in acceptance (white) and rejection (red) regions in a hypothetical two-dimensional scenario with 90 issuers in each rating class and PDs under H_0 set to 32% and 35% respectively.

3.3 Existing Two-Sided Joint Tests for Rating Systems

When testing the calibration quality of a rating system given by the two-sided composite hypothesis of equation (3) the literature established a number of typically two-sided tests, see e.g. Aussenegg et al. [2011] for an overview. In contrast to the multiple testing procedures discussed in the previous Section 3.2, these *joint tests* are not based on a separate assessment of all rating classes which are then aggregated, but assess the calibration of all rating classes of the system jointly.

Unfortunately, the "gold standard" of univariate two-sided tests, the Clopper-Pearson test described in subsection 3.1 cannot be extended to a multivariate setting as shown e.g. in Aussenegg et al. [2011].

The concept of the univariate Sterne test, however, i.e. defining the acceptance region based on a ranking of outcomes by probability, can be applied to multiple dimensions. The resulting multivariate *two-sided Sterne test* is defined by its acceptance region

$$\mathcal{A}_{Sterne} := \left\{ \mathbf{d} \in \Omega \;\middle|\; \sum_{\mathbf{i} \in \Omega:\ \mathcal{B}_C(\mathbf{i};\mathbf{n},\hat{\mathbf{p}}) \leqslant \mathcal{B}_C(\mathbf{d};\mathbf{n},\hat{\mathbf{p}})} \mathcal{B}_C(\mathbf{i};\mathbf{n},\hat{\mathbf{p}})) > \alpha \right\}.$$

Aussenegg et al. [2011] derive the multivariate Sterne test and show that it converges to the test by Hosmer and Lemeshow [1980] which is widely used for backtesting the calibration quality of credit assessment systems. Furthermore, they show how the multivariate versions of the Wald and Score tests and the

Score test with continuity correction are related to the multivariate Sterne test. We base our novel joint tests on the Sterne test, as it is the exact version of various approximate tests known in the literature. While the Sterne test is precise for all sample sizes, the approximate tests are biased for small sample sizes for which the normal distribution is not an appropriate approximation of the binomial distribution.

Besides its close links to numerous two-sided tests, the Sterne test has another appealing property: by construction the Sterne test has the smallest acceptance region among all tests of a given level. Note that this does not imply that the Sterne test is the most powerful test for our problem setting. In fact, for composite alternatives such as the hypotheses of equations (2) and (3) no uniformly most powerful test exists. Hence, identifying an optimal test in terms of power in our problem setting would require additional assumptions on the composite alternative, such as (i) selecting a specific parameter vector for the alternative: $H_1 = \mathbf{p}$, which leads to the likelihood-ratio test,[8] but neglects power under all other parameters, or (ii) the assumption of a distribution of the parameters under H_1, implying different weights to the parameters under H_1. However, assumptions on the composite alternative are hard to justify, if there is no prior belief about the alternative.

In light of the non-existence of a uniformly most powerful test we consider another intuitive criterion in order to compare different tests, namely the size of the acceptance region. A smaller acceptance region is intuitively appealing for several reasons: first, for any given H_1 (simple or composite), decreasing the size of the acceptance region by removing observations can obviously never decrease the power of a test and will in most cases increase the power. Second, minimizing the size of the acceptance region does not require any assumption about H_1. Consequently, for these two reasons, maximally minimizing the acceptance region of a test is beneficial in the absence of any prior knowledge about the probability distribution of H_1. To our knowledge, this criterion to compare tests by the size of the acceptance region is new to the literature except for Reiczigel et al. [2008] who show that this criterion implies confidence sets with good coverage properties outperforming conventional confidence sets. We will further analyse this concept in section 5 when we benchmark our new one-sided joint tests against the multiple test.

4 New One-Sided Joint Tests

Section 3.3 summarizes standard joint tests which are able to assess the calibration quality of all rating classes jointly by testing the composite two-sided hypothesis of equation (3). In the medical statistics literature, several one-sided tests comparing multivariate hypotheses similar to that of equation (2) were developed by Bartholomew [1959], Perlman [1969], and O'Brien [1984], in particular for the use in clinical trials where treatment success is measured by several indicators simultaneously. In particular, the likelihood ratio test by

[8] Note that according to the Neyman-Pearson lemma, the likelihood-ratio test is the most powerful test for a simple alternative of the form: $H_1 = \mathbf{p}$, i.e. it outperforms all tests of a given level in terms of power under this simple specification of the alternative. Note further, that the acceptance region of this test is bounded by a hyperplane, which depends on the parameter $\mathbf{p}$. This implies that there does not exist an uniformly most powerful test for composite hypotheses of the form of equations (2) and (3).

Perlman [1969] has evolved to be one of the standard procedures for such problems. Making the assumption of underlying multivariate normal distributions, the test allows for a closed form solution for the p-value of the likelihood ratio test. However, the assumption of a multivariate normal distribution is often not applicable to questions involving discrete variables, in particular the credit risk estimation problem studied in this paper where the probabilities and the sample sizes are low. An extension of the likelihood ratio test to the multivariate binomial distribution would be straight forward from an analytical view point. The practical implementation of this test, however, would still need to be defined and would require computationally intensive procedures. To the best of our knowledge, the literature does not include multivariate joint tests to assess the performance of rating systems based on multivariate binomial distributions formulated by the composite one-sided hypothesis of equation (2). This section presents four novel multivariate one-sided tests for the joint performance of all classes of a rating system, with the aim to close this gap. Of the four presented tests the first three are one-sided variants of the Sterne test and the last an enhanced joint version of the multiple test presented in section 3.2.3. These one-sided tests are particularly relevant when the focus lies on loss prevention.

4.1 One-sided Optimal and Iterative Sterne Tests

As described in Section 3.3 it is not possible to find a test which outperforms all other tests in terms of power if there is no prior knowledge about the composite alternative. We found that the Sterne test outperforms all other tests in terms of minimizing the size of the acceptance region. When addressing one-sided tests, employing the same optimality criterion allows us to find an optimal one-sided test. Hence this subsection presents the *one-sided optimal Sterne test* $\phi_{1OptSterne}$. This test is defined by the one-sided test of level α which has the lowest number of observations in its acceptance region, i.e.

$$\phi_{1OptSterne} := \arg\min_{\phi} \{ \ \#(\mathcal{A}_\phi) \mid \phi \text{ is one-sided}, \ \phi \in \Phi_\alpha \ \},$$

where $\#(\mathcal{A}_\phi)$ denotes the number of observations in the acceptance region $\mathcal{A}_\phi$.[9] This test $\phi_{1OptSterne}$ is optimal in minimizing the acceptance region among all one-sided tests of the same level α, noted Φ_α. Thus it tends to achieve a high power under all parameters of the composite alternative, as explained in Section 3.3. It further serves as the conceptual basis for the one-sided variant of the Sterne test in Section 4.2, which approximates this optimal test.

However, this test is very hard to implement, since the rejection areas of all tests for a given level must be compared. In particular, deriving all one-sided rejections regions for higher dimensions C is computationally very intensive.

An alternative which is easier to compute is an iterative procedure to derive an acceptance region that yields a one-sided test while applying the Sterne method. It can be summarised as follows: at each iteration step, all candidate observations which, when excluded from the acceptance region, ensure that it is still one-sided, are considered and the observation with lowest probability is ex-

[9] It is possible, though unlikely, that this test is not uniquely defined. In this case, one chooses the test with the highest power at the alternative hypothesis $H_1 = 2\hat{\mathbf{p}}$; this arbitrarily chosen H_1 does not affect any of our results.

cluded. The steps used to derive the *one-sided iterative Sterne test* $\phi_{1IterSterne}$ are detailed in Annex A.

We do not present the results of this iterative test as it is still computationally rather intensive and might not be optimal in terms of having the smallest acceptance region; it might thus not coincide with the one-sided optimal Sterne test $\phi_{1OptSterne}$.

4.2 One-sided Sterne Envelope Test

Given the complexity of the computational implementation of the *one-sided optimal Sterne test* described in section 4.1, we finally consider the one-sided variant of the two-sided Sterne test: its one-sided 'envelope'. We denote this *one-sided Sterne envelope test* by $\phi_{1envSterne}$. The intuition of this test in two dimensions is the following: starting from a two-sided Sterne test (Figure 3) which is optimal in terms of size of the acceptance region, we make it one-sided as shown on Figure 4. The significance level of the two-sided input test then is increased to ensure the one-sided outcome test reaches the desired significance level. Using a rounded top right corner potentially improves the power of the test for a wide range of alternatives compared to the multiple test.

In order to define it formally, we first consider all one-sided tests, the acceptance region of which contain the acceptance region of the two-sided Sterne test of level α', i.e.

$$\Phi_{\alpha'}^{env} := \{ \phi \mid \phi \text{ is one-sided and } \mathcal{A}_{\phi_{2Sterne}} \subseteq \mathcal{A}_\phi \text{ for } \phi_{2Sterne} \in \Phi_{\alpha'} \}$$

The acceptance region of the one-sided Sterne envelope test $\phi_{1envSterne}$ is defined by the intersection of all one-sided acceptance regions containing the two-sided acceptance region, i.e.

$$\mathcal{A}_{1envSterne} := \bigcap_{\phi \in \Phi_{\alpha'}^{env}} \mathcal{A}_\phi.$$

Further we set the rejection region $\mathcal{R}_{1envSterne} := \Omega - \mathcal{A}_{1envSterne}$.

As the rejection region of the one-sided Sterne envelope $\mathcal{R}_{1envSterne}$ is typically smaller than the rejection region of the original two-sided test $\mathcal{R}_{2Sterne}$, the one-sided Sterne envelope test has a lower level α than the two-sided test it is based on, i.e.

$$\alpha = \sup_{p \in H_0} \mathrm{E}_{\mathbf{p}} \phi_{1envSterne} < \alpha'.$$

The difference between α and α' depends on the distributions under H_0. It is higher, if the distance of the modes of these distributions from the origin is larger. For a given level α, we can find the level α' which implies a level for one-sided Sterne envelope test close to α, i.e.

$$\sup_{p \in H_0} \mathrm{E}_{\mathbf{p}} \phi_{1envSterne} \approx \alpha.$$

As the following proposition shows this test has appealing conceptual properties: It is one-sided and it only rejects rating-systems which are rejected by the two-sided Sterne test at level α'.

Proposition 1. *Let* $\phi_{2Sterne}$ *denote the two-sided Sterne test for a given level* α'. *For the one-sided Sterne envelope test* $\phi_{1envSterne}$ *it holds* *(i)* $\phi_{1envSterne}$ *is one-sided and it holds* $\mathcal{A}_{\phi_{2Sterne}} \subseteq \mathcal{A}_{\phi_{1envSterne}}$.

(ii) $\phi_{1envSterne}$ *is optimal in terms of having the largest rejection region among all one-sided tests whose acceptance region contain the two-sided acceptance region, i.e.*

$$\phi_{1envSterne} := \arg\max_{\phi}\{ \#(\mathcal{R}_\phi) \mid \phi \text{ is one-sided with } \mathcal{A}_{\phi_{2Sterne}} \subset \mathcal{A}_\phi\}.$$

Proof. (i) As it holds $\mathcal{A}_{\phi_{2Sterne}} \subset \mathcal{A}_\phi$ for all $\phi \in \Phi^{env}$, it also holds for the intersection and thus for $\phi_{1envSterne}$.

To see that $\phi_{1envSterne}$ is one-sided, consider some $\mathbf{d} \in \mathcal{R}_{1envSterne}$. There exists some test $\phi^r \in \Phi^{env}$ with $\mathbf{d} \in \mathcal{R}_{\phi^r}$, else it cannot hold $\mathbf{d} \in \mathcal{R}_{1envSterne}$. Since ϕ^r is one-sided, it holds $\mathbf{d} + e_c \in \mathcal{R}_{\phi^r}$, thus $\mathbf{d} + e_c \notin \mathcal{A}_{\phi^r}$, so $\mathbf{d} + e_c \notin \mathcal{A}_{1envSterne}$, implying $\mathbf{d} + e_c \in \mathcal{R}_{1envSterne}$ for $c = 1, .., C$. The implication for some $\mathbf{d} \in \mathcal{A}_{1envSterne}$ follows by the one-sidedness of all $\phi \in \Phi^{env}$.

(ii) For a proof by contraposition we assume the existence of some $\phi^c \in \Phi^{env}$ with

$$\#(\mathcal{R}_{\phi^c}) > \#(\mathcal{R}_{\phi_{1envSterne}}).$$

By $\#(\mathcal{A}_{\phi^c}) = \#(\Omega) - \#(\mathcal{M}_{\phi^c}) - \#(\mathcal{R}_{\phi^c})$ and $\#(\mathcal{A}_{1envSterne}) = \#(\Omega) - \#(\mathcal{R}_{1envSterne})$, it follows $\#(\mathcal{A}_{\phi^c}) < \#(\mathcal{A}_{1envSterne})$. However, by $\phi^c \in \Phi^{env}$, it follows $\mathcal{A}_{1envSterne} \subset \mathcal{A}_{\phi^c}$, and thus $\#(\mathcal{A}_{\phi^c}) \geqslant \#(\mathcal{A}_{1envSterne})$, which is a contradiction.

□

After these conceptual considerations, we turn to the numerical implementation of this test. The main difficulty here is deriving the level α' of the two-sided Sterne test, which ensures the given level α for the one-sided Sterne envelope test. Since the observed target α for the one-sided Sterne envelope test is essentially monotonically increasing in the level α', this can be found by numerical iteration techniques.

Figures 3 and 4 illustrate the construction of the one-sided Sterne envelope test in a hypothetical two-dimensional scenario with 90 issuers in each rating class and PDs under H_0 set to 32% and 35% respectively. Figure 3 depicts the acceptance and rejection regions of the two-sided Sterne test $\phi_{2Sterne}$ and figure 4 the acceptance and rejection regions of the one-sided Sterne envelope test, the test with the smallest acceptance region containing that of $\phi_{2Sterne}$.

4.3 Enhanced multiple test

The multiple test of section 3.2.3 controls the $FWER$ and corrects for alpha-inflation. However, due to the fact that each rating class is tested separately, its acceptance region is represented by a box in the observation space (i.e. a rectangle in a setting where two rating classes are tested, a cuboid with three classes as shown on figure 6). Because of this rigid specification and the discreteness of the binomial distribution, in general it holds that $FWER < \alpha$. This allows to remove observations from the acceptance region $\mathcal{A}_{mult}$ of the multiple test defined in equation (4) without exceeding the level α. We use this fact to define a simple test, which shares features of joint tests. In order to define this *enhanced*

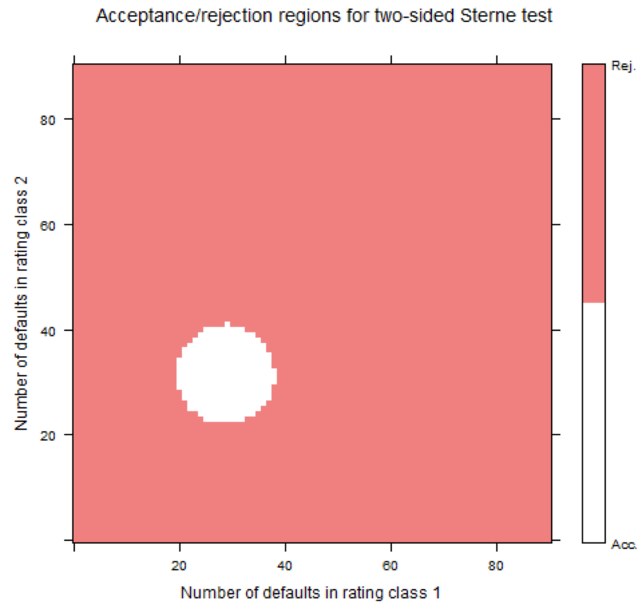

Figure 3: Two-sided Sterne test acceptance (white) and rejection (red) regions in a hypothetical two-dimensional scenario with 90 issuers in each rating class and PDs under H_0 set to 32% and 35% respectively. A significance level of $\alpha' = 11\%$ was chosen for the two-sided Sterne, leading to a significance level of $\alpha = 5\%$ for the one-sided envelope test.

multiple test, we consider the hyper-pyramid containing default patterns which are accepted by the multiple test and whose total number of defaults over all classes exceeds m:[10]

$$H(m) = \left\{ \mathbf{d} \in \mathcal{A}_{mult} \middle| \sum_{i=1}^{C} d_i \geqslant m \right\}$$

Define m_0 as the minimum m for which $\mathbf{P}_{\hat{\mathbf{p}}}(H(m)) \leqslant \alpha - FWER$, the acceptance region of the enhance multiple test then is

$$\mathcal{A}_{mult+} = \mathcal{A}_{mult} - H(m_0)$$

Note that by construction the enhanced multiple test rejects all default patterns that are rejected by the multiple test. It further rejects default patterns with low performance in *all* rating classes. The enhanced multiple test is therefore uniformly more powerful than the multiple test, i.e. it is more powerful for any specification of an alternative hypothesis. This is achieved while still controlling

[10]It should be noted that the region $H(m_0)$ which is removed from the multiple test's acceptance region could also be chosen to be of a different shape than a hyper-pyramid. Thus the enhanced multiple test can still be optimised in terms of power or size of the acceptance region by considering another shape for the region $H(m_0)$. However, in a discrete context such as here, the hyper-pyramid allows easy understanding and implementation.

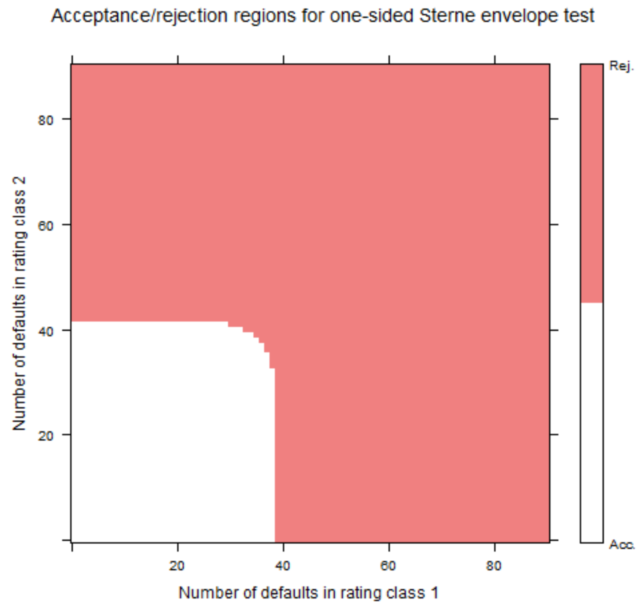

Figure 4: One-sided Sterne envelope test acceptance (white) and rejection (red) regions in a hypothetical two-dimensional scenario with 90 issuers in each rating class and PDs under H_0 set to 32% and 35% respectively. The significance level is $\alpha = 5\%$.

the $FWER$ at the level α. A comparison of power of the two tests is presented in Section 5.1.

Figures 5 depicts the acceptance and rejection regions of the enhanced multiple test in a hypothetical two-dimensional scenario with 90 issuers in each rating class and PDs under H_0 set to 32% and 35% respectively. In this scenario, the multiple test rejects the rating system if 40 or more defaults are observed in the first and if 42 or more defaults are observed in the second rating class. The enhanced multiple test further rejects those observations where the sum of defaults exceeds 72.

4.4 Stylised Comparison between Multiple and Joint Tests

This subsection presents a stylized comparison between the multiple and a typical joint test in order to illustrate cases where the two tests reach different decisions.

It is the defining element of the multiple test that it assesses each class of a rating system separately. The global null hypothesis of a well-performing rating system is then rejected if at least one class is rejected, while the exact number of rejected rating classes is irrelevant.

In contrast, a joint test assesses the joint performance of the whole rating system simultaneously. In particular, a rather poor performance in one rating class can be balanced by a good performance in other classes. On the other hand, if only medium performance is found in the majority of classes the system can

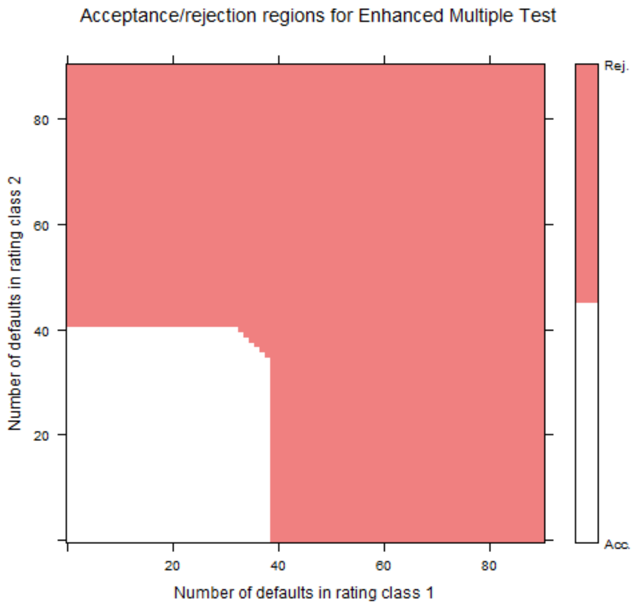

Figure 5: Enhanced multiple test acceptance (white) and rejection (red) regions in a hypothetical two-dimensional scenario with 90 issuers in each rating class and PDs under H_0 set to 32% and 35% respectively. The significance level is 5%.

be rejected, although no class performs very poorly when assessed individually. More generally, a system can be rejected by *any* combination of classes with poor performance.

This is illustrated in Figure 6 which restricts to the case of three rating classes for the ease of presentation. The figure shows the acceptance region of the multiple (orange) and a joint test (yellow). Each axis represents the number of observed defaults in each class: the higher these numbers, the more evidence there is for underperformance.[11] As the multiple test assesses each class separately, its acceptance region is necessarily a cuboid, whereas the acceptance region of the joint test is more flexible. We consider the following default patterns to illustrate when the decision of the multiple deviates from a joint test:

- Point A (no defaults in classes 1 and 2) represents a case where the rating system is rejected based solely on the poor performance in class 3. Point A is rejected by the multiple and the joint test.
- For any point lying between the segment [BC] and the rounded region of the yellow surface, the system is rejected based on a combined poor performance of classes 2 and 3 by a joint test. However, these points are not rejected by the multiple test.
- For points lying between point D and the rounded yellow surface, the

[11] Note that it is possible that the acceptance region of the joint test (yellow) fits perfectly in that of the multiple test (orange) but that in the most general case (represented here) it partially protrudes.

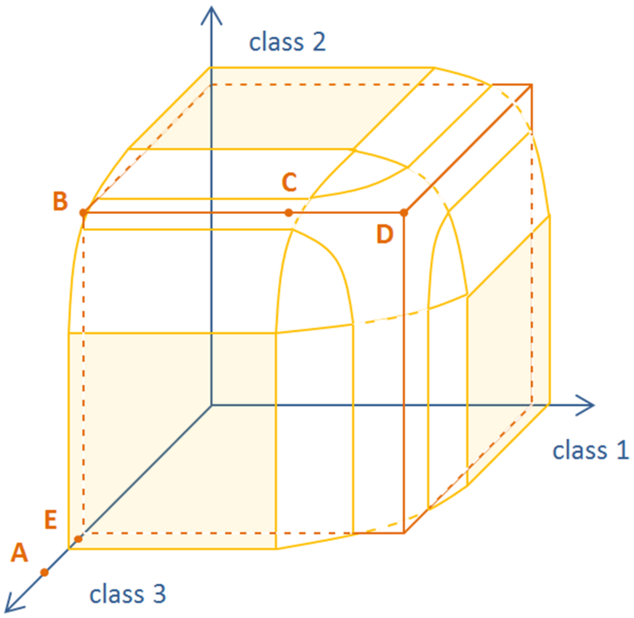

Figure 6: Acceptance region of the multiple (orange) and a joint test (yellow) for a rating system with three classes. Each axis represents the number of observed defaults in each class: the higher these numbers, the more likely it is to fall outside the acceptance region, i.e. is to see the rating system rejected.

combined defaults observed in all three classes would lead to a rejection of the rating system by the joint test. Again, these points are not rejected by the multiple test.

- Point E is a case with rather poor performance only in class 3. This point is rejected by the multiple test, but not by the joint test, as the performance in the classes 1 and 2 is very good.

5 Cost-Benefit Analysis of Different Tests

This Section compares the new set of joint tests proposed in Section 4 with the multiple test of subsection 3.2 as a benchmark. It analyses if the new tests can outperform the benchmark test with respect to the benefits that characterise a good test in general, which are a high power of identifying underperformance and a small acceptance region. The analysis is performed in a baseline scenario of a standard rating system in Subsections 5.1 and 5.2, as well as for further rating systems with differently-sized rating classes in Subsection 5.3. Finally, potential costs of the new joint tests in terms of a more complex implementation and communication are compared in Subsection 5.4.

5.1 Comparison of Benefits in Terms of Power in the Baseline Scenario

Following the standard approach in the literature, we first control the probability to wrongly reject our one-sided hypothesis of well-performance given in equation (2) by bounding it by the significance level of the test. In our case, the significance level α is set to 5%. The standard second step is then to look for the test that maximises the probability to correctly reject H_1, i.e. the power of the test. It is important to note that the calculation of the power of a test requires an assumption about the parameter vector for H_1.

The choice of the parameter vector for H_1 is a specific challenge in our testing problem. In some problems such as van Dyk [2014] the statistical test helps discriminating between two well-specified hypotheses and the H_1 parameter vector is uniquely defined. However, in many other testing problems, including our tests of the performance quality of rating systems, the conductor of the test does not know the parameter vector of H_1. In order to ensure the robustness of our results, we compare the different tests under a variety of assumptions about H_1 in this section.

5.1.1 Data Description

The baseline scenario for the comparison of the tests uses five rating classes. Each rating class reflects one credit quality step following the Basel framework for banks' capital requirements purposes. The European Banking Authority (EBA) has published draft mapping reports European Banking Authority (EBA) [2015] for all credit rating agencies that are recognised by the European Securities and Markets Authority. We use the EBA mapping report for Standard & Poor's as the basis to assign Standard & Poor's rating grades to the five rating classes.[12]

We then use the statistics for global corporate ratings in Standard & Poor's [2012] to determine the number of rated entities ('size') and the PD over a one-year horizon for each rating class. Standard & Poor's number of global corporate ratings in 2012 and the average one-year realised default rate between 1981 and 2012 serve as proxies for the size and the PD, respectively. Table 1 shows the realised default rates used as PDs for the null hypothesis of equation (2) in this section.

5.1.2 Specifications for Alternative Hypotheses

We consider three alternative specifications of the parameter vector under H_1 to compare the power of the different tests with this baseline scenario. All three specifications ensure that the power of the tests is in a range that allows a meaningful comparison of the tests.[13]. Table 2 summarises these specifications and its technical details are presented in detail in Appendix B.

[12] Credit quality step 6, equivalent to a rating by Standard & Poor's between CCC and C, is neglected in the back-testing analysis because it is the last rating class prior to default with only 154 rated corporates and an average default rate of 26.85%.

[13] If the parameter vector under H_1 is chosen too close to H_0 or too far from H_0, all tests will yield a power close to the significance level α or (close to) 1, respectively. It is thus not possible to show in this case that one test outperforms another test for a large region of H_1.

Table 1: Null hypothesis PD ($\mathbf{p}$) of equation (2) used for the study, based on Standard & Poor's (S&P) global corporate ratings and other portfolios. The data are taken from Tables 14, 33 and 51 of Standard & Poor's [2012]. The realised default rate for rating class 1 is the rounded weighted average of the realised default rate for Standard & Poor's rating grades 'AAA' and 'AA'.

Rating class (c)	S&P ratings	Realised default rate (%)
1	AAA/AA	0.02
2	A	0.07
3	BBB	0.22
4	BB	0.86
5	B	4.28

Table 2: Three alternative specifications for the alternative hypothesis are studied in order to compare the power of the different tests.

	Description
H_1A	Alternative PDs are obtained by increasing the null hypothesis PDs of Table 1 towards $\mathbf{p} = 1$ in all rating classes as explained in Appendix B.
H_1B	Alternative PDs are obtained by increasing the null hypothesis PD of only *one* class at a time.
H_1C	Alternative PDs are drawn from a multivariate normal distribution, centered at the null hypothesis PDs.

Abstracting from technical details, the main difference between the alternative specifications is that the first specification H_1A reflects an underestimation of default risk in *all* rating classes simultaneously and the second specification H_1B assumes underestimation of default risk in exactly *one* rating class. The third specification H_1C allows for any combination of rating classes featuring underestimation while giving most weight to two and three rating classes featuring underestimation. Whether the underestimation of risk is more likely in a broad range or in a small number of rating classes depends on the source of the miscalibration. For example, if an unobservable common factor that is not included in the rating model but correlated with the explanatory variables turns negative, it will lead to more defaults than anticipated in all rating classes. In contrast, the use of expert judgment for rating adjustments after their calibration to PDs may lead to systematic underestimation of risk in some (low quality) rating classes.

In the first specification H_1A, we calibrate the parameter vector to ensure a power of 50% for the multiple test, which serves as our benchmark. To that end Appendix B describes formally how the parameter vector is adjusted in *all* rating classes to reach the power of 50% for the multiple test. The first specification H_1A thus reflects an underestimation of default risk in all rating classes simultaneously. In Figure 6, specification H_1A gives most weight to default patterns located close to point D. Table 4 shows the results for this baseline scenario. The enhanced multiple test slightly increases power from 50% to 53.4%. The Sterne envelope increase power significantly to 69.7%. The greater power of the two joint tests confirms the expected result that the joint tests can better identify underperformance occurring in all rating classes compared with the multiple test.

The second specification H_1B is similar to H_1A, except that the parameter vector is adjusted only in *one* of the five rating classes. For each of the five rating classes the alternative is derived, which gives the power of 30% for the benchmark test.[14] For the two joint tests power is then computed as the average of power over these five alternatives. This gives a representative estimate of the power of the different tests if credit risk is underestimated in exactly one rating class. The second specification H_1B results in an average power of 31.4% for the enhanced multiple and 29.6% for the Sterne envelope test (see Table 4). Thus power differences between the three tests are very small and might be negligible. This is a promising result for the two joint tests, as intuition suggests that they perform worse than the benchmark test because of their construction if underperformance occurs in only one rating class (see also Section 4.4). In the context of Figure 6 this implies that default patterns similar to point E rarely occur.

For the third specification of the alternative hypothesis, H_1C, we define a multivariate normal distribution on the complete set of H_1 on the parameter space $\mathcal{P}$. The distribution, which is described in detail in Appendix B states the hypothetical probability for each potential parameter of H_1 to be the true parameter. Since all parameters have a positive probability, all possible parameters are (marginally) accounted for. For all tests power is then computed by integrating the power of each alternative over the parameter space with this distribution. The third specification H_1C thus reflects an underestimation of

[14]See Appendix B for a more formal description.

default risk distributed over any combination of rating classes, with most probability mass reflecting a underperformance in two and three rating classes. Under specification H_1C the enhanced multiple test increases power compared with the multiple test from 7.7% to 8.3%. The Sterne envelope increase power significantly to 11.3%. Expressing these power increases in relative terms instead of absolute terms shows that they are significant, as the Sterne envelope increases the power of the benchmark test by 46.8%. The greater power of the joint tests compared with the multiple test provides a strong robustness check to the same result under the first specification H_1A. It suggests that the one-sided joint tests can better identify a broad range of miscalibrations resulting from any number of rating classes in underperformance.

5.2 Comparison of Benefits in Terms of Size of Acceptance Region in the Baseline Scenario

Section 3.3 highlights that the size of the acceptance region $\mathcal{A}^{Size}$ can serve as an additional criterion on the basis of which different tests can be compared, in particular in the case of a lacking knowledge on the alternative hypothesis. Indeed, this criterion does not require an explicit specification of the parameter vector of H_1. The reduction of the size of the different acceptance regions can be expressed in terms of the reduction of the number of default patterns included in the acceptance region compared to the benchmark test relative to the size of the acceptance region of the benchmark test, i.e. for test j as $\left(\mathcal{A}_j^{Size} - \mathcal{A}_{multiple}^{Size}\right)/\mathcal{A}_{multiple}^{Size}$.

Table 4 shows that the enhanced multiple test reduces the size of the acceptance region of the benchmark test slightly by 3% and the Sterne envelope test reduces the size significantly by 72%.

In line with our findings for the power comparison, the size criterion confirms our result that the enhanced multiple test slightly outperforms the multiple test and that the Sterne envelope test significantly outperforms the multiple test in the studied baseline scenario.

The analysis in the following Section 5.3 shows the robustness of these results to different size scenarios. However, the degree of this gain has to be assessed against the additional costs in terms of computational efficiency and simplicity of implementation (see Section 5.4).

5.3 Comparison of Benefits in Terms of Power and Size of Acceptance Region for Other Rating Portfolios

This section considers whether the results for the baseline scenario are robust for rating systems whose classes have different sizes. For this purpose, we repeat the evaluation of the baseline scenario in Section 5.1 with the same H_0 given by equation (2) and Table 1 for a

1. portfolio of rated entities that is biased towards greater PDs, represented by Standard & Poor's ratings for non-financial corporations only (Scenario 'Non-financials'),

2. portfolio of rated entities that is biased towards lower PDs, represented

by Standard & Poor's ratings for insurance companies only (Scenario 'insurance'),

3. hypothetical small rating system with only 100 rated entities in each rating class (Scenario 'small'),
4. hypothetical large rating system with 5,000 rated entities in each rating class (Scenario 'large').

Table 3 summarises the studied size scenarios, i.e. the distribution of obligors per rating class.

Table 3: Studied size scenarios: number of obligors per rating class based on Standard & Poor's (S&P) global corporate ratings and other portfolios. The data are taken from Tables 14, 33 and 51 of Standard & Poor's [2012]. The 'small' and 'large' size scenarios are hypothetical.

Size scenarios	Size ($\mathbf{n}$) of rating class: 1	2	3	4	5
Baseline	374	1,330	1,637	1,047	1,471
Non-financials	100	563	1,084	836	1,277
Insurance	148	387	188	48	27
Small	100	100	100	100	100
Large	5,000	5,000	5,000	5,000	5,000

Table 4 presents the power and sizes of acceptance regions of the three tests under the four alternative size scenarios. It confirms the results of the baseline scenario. The portfolio bias towards higher quality (Scenario 'Insurance') or lower quality (Scenario 'Non-financials') rated entities has very limited influence on the outcome of the comparison of the different tests. Abstracting from the second specification H_1B which results in very limited power differences, in all combinations of power specification and size scenario,[15], the enhanced multiple test slightly outperforms the benchmark test while the Sterne envelope test significantly outperforms the benchmark test. The gains in power and size reduction are less pronounced for the two hypothetical scenarios 'Small' and 'Large' compared to the three real-world size scenarios.

To conclude, the comparison of the different tests on a purely statistical basis suggest that one-sided joint tests can outperform the multiple test for most combinations of sizes and alternative hypotheses H_1. The statistical gains from joint tests appear particularly pronounced for intermediate size scenarios and specifications of H_1 that are tilted towards underperformance in many or all classes, as can be expected from the different designs of the multiple and joint tests (see in particular Section 4.4). Furthermore, among the two studied joint tests, the Sterne envelope test seems to clearly outperform the enhanced multiple test. These results are qualitatively not affected by using alternative parameter vectors $\mathbf{p}$ for H_0 (not reported, but available upon request).

[15]Specification H_1B combined with the hypothetical scenario 'Small' is the only exception.

Table 4: Comparison of power and size of acceptance region for all size scenarios (Table 3) and all specifications of the alternative hypothesis. For each of the three tests (m) multiple, (em) enhanced multiple and (ev) envelope Sterne test the power and size of the acceptance region are given and the best performing test is highlighted in blue.

Size scenarios	H_1A	(power) H_1B	H_1C	(size) Acc. region	Type of test
Baseline	50.0%	30.0%	7.7%	123,930	(m)
	53.4%	31.4%	8.3%	-3%	(em)
	69.7%	29.6%	11.3%	-72%	(ev)
Non-financials	50.0%	30.0%	6.2%	42,336	(m)
	52.0%	30.7%	6.5%	-2%	(em)
	64.4%	31.1%	8.4%	-67%	(ev)
Insurance	50.0%	30.0%	5.4%	216	(m)
	54.2%	30.6%	5.6%	-22%	(em)
	74.3%	31.9%	7.6%	-61%	(ev)
Small	50.0%	30.0%	7.3%	240	(m)
	51.1%	30.7%	7.5%	-12%	(em)
	57.5%	29.9%	6.9%	-47%	(ev)
Large	50.0%	30.0%	28.4%	2,279,088	(m)
	50.7%	30.4%	28.5%	± 0%	(em)
	62.2%	26.5%	35.2%	-24%	(ev)

5.4 Comparison of Costs in Terms of Implementation and Communication

The comprehensive assessment of statistical tests from a practitioner's perspective goes beyond purely statistical properties such as power or the size of the acceptance region. The multiple test can be seen as a good test for practitioners: The multiple test is easily implementable in the available IT infrastructure and has a short computation time; it can even be implemented with a standard spreadsheet software. The multiple test is also intuitive enough to allow the communication of the results in a simple and transparent way on the basis of a well-founded and widely accepted methodology.

At the same time, the statistical analysis in Section 5.1 has shown that the joint tests can improve upon the multiple test for a wide range of alternative hypotheses H_1. The R-package "validateRS" available with this paper[16] makes the tests easily implementable. This paper has developed the formal foundation of these tests and shown that the results of the test can be easily communicated. The joint tests, in particular the iterative Sterne test, have only some limitations in terms of computational efficiency if the number of dimensions C, the size of individual classes N_c or some elements of the probability $\mathbf{p}$ under H_0 become very large. Table 5 illustrates the computation time for the different tests and scenarios.

[16] The R-package can be downloaded from the Comprehensive R Archive Network `https://cran.r-project.org/`.

Table 5: Comparison of computation time for all scenarios. Both (i) the time required to define the test, i.e. to determine the acceptance region and (ii) the time required to compute the power for a single given alternative hypothesis are reported for each of the size scenarios defined in Table 3. For each of the three tests (m) multiple, (em) enhanced multiple and (ev) envelope Sterne test the times are compared. The shortest computations times (market in blue) are always observed for the multiple test.

Size scenarios	(milliseconds) Determining accept. region	(milliseconds) Power computation	Type of test
Baseline	240	0.3	(m)
	360	5.1	(em)
	68,970	9.9	(ev)
Non-financials	70	0.3	(m)
	170	2.0	(em)
	36,380	5.0	(ev)
Insurance	50	0.2	(m)
	140	1.2	(em)
	1,430	0.4	(ev)
Small	40	0.2	(m)
	160	1.2	(em)
	1,260	0.4	(ev)
Large	60	0.3	(m)
	13,970	99	(em)
	1,358,500	1,334	(ev)

6 Conclusion

This paper presents a new set of one-sided multivariate tests for the ex-post detection of credit risk underestimation of rating systems. Existing one-sided multivariate tests are based on an assessment of the rating performance in each of the system's rating classes separately. The novelty of the presented tests consists in the joint assessment of the performance in all rating classes. The rejection of a rating system can therefore not only be triggered by a higher-than-expected default rate in a single class but by a poor performance in any combination of rating classes.

The new tests are shown to outperform the established one-sided multivariate test by Westfall and Wolfinger [1997] in terms of power for a variety of Standard & Poor's rated portfolios. The concrete gain in power depends on the specification of the alternative hypothesis. When compared in terms of the size of the acceptance region, which is a novel measure that is beneficial when little is known about the alternative hypothesis, the new tests significantly outperform the benchmark test. However this increased performance comes at the expense of increased implementation complexity and computation time.

Appendix A Steps of the One-sided Iterative Sterne Test

This Annex explains the iterative procedure mentioned in Section 4.1 to derive an acceptance region that yields a one-sided test while applying the Sterne method. It can be summarised as follows: at each iteration step, all candidate observations which, when excluded from the acceptance region, ensure that it is still one-sided, are considered and the observation with lowest probability is excluded. The steps used to derive the *one-sided iterative Sterne test* $\phi_{1IterSterne}$ are detailed below.

- Initial step. The methods starts by setting the acceptance region of the test to the complete observation space i.e. $\mathcal{A}_0 = \Omega$. Then the the most extreme observation $\bar{\mathbf{d}} = (n_1, ..., n_C)$, which forms a corner point the observation space Ω, is considered. If it holds that $\mathbf{P}_{\hat{\mathbf{p}}}(\bar{\mathbf{d}}) > \alpha$, the final step is performed and the acceptance region is set to Ω. If on the other hand it holds that
$$\mathbf{P}_{\hat{\mathbf{p}}}(\bar{\mathbf{d}}) \leqslant \alpha,$$
then $\bar{\mathbf{d}}$ is excluded from $\mathcal{A}_0$ and the iteration step is performed.

- Iteration step. Let $\mathcal{A}_{i-1}$ be the acceptance region of the previous step. We show how to derive $\mathcal{A}_i$. Consider all $\mathbf{d} \in \Omega$ such that $\mathcal{A}_i := \mathcal{A}_{i-1} \backslash \{\mathbf{d}\}$ and $\mathcal{R}_i = \Omega - \mathcal{A}_i$ define a one-sided test. Sort these observations by probability under $\mathbf{P}_{\hat{\mathbf{p}}}$. Let $\underline{\mathbf{d}}$ be the observation with the lowest probability. If it holds
$$\mathbf{P}_{\hat{\mathbf{p}}}(\mathcal{A}_{i-1} \backslash \{\underline{\mathbf{d}}\}) \geqslant 1 - \alpha,$$
then $\underline{\mathbf{d}}$ is excluded from the acceptance region, i.e. $\mathcal{A}_i := \mathcal{A}_{i-1} \backslash \{\underline{\mathbf{d}}\}$. If $\underline{\mathbf{d}}$ is not uniquely defined, choose the one which gives access to more observations with lower probabilities than $\underline{\mathbf{d}}$. If $\underline{\mathbf{d}}$ is still not uniquely defined, choose the one with lower d_1. If on the contrary it holds that $\mathbf{P}_{\hat{\mathbf{p}}}(\mathcal{A}_{i-1} \backslash \{\underline{\mathbf{d}}\}) < 1 - \alpha$, the iteration stops.

- Final step. Set $\mathcal{A}_{1IterSterne} := \mathcal{A}_{i-1}$ and $\mathcal{R}_{1IterSterne} = \Omega - \mathcal{A}_{i-1}$.

Appendix B Specification of alternative hypothesis H_1 for power comparisons

This Annex explains in detail how the alternative hypotheses H_1 are derived for the comparison of power in Section 5.1.

For the first and second specification, H_1A and H_1B, we consider that parameter value for which the power of the benchmark multiple test, equals 50% and 30% respectively for a given sample size. Therefore we parameterise the **first specification** of the alternative hypothesis, H_1A, as follows

$$H_1A: \quad p_c = (1-s)\hat{p}_c + s \quad \forall c = 1, ..., C$$

with the one-dimensional parameter $s \in [0, 1]$. Let ϕ_{mult} be the multiple test, which we use as benchmark test.

Now we solve for the value of s which gives a power of 0.5:

$$\mathrm{E}_{\mathbf{p}(s)}\phi_{mult} = \sum_{\mathbf{d}\in\mathcal{R}(\phi_{mult})} \mathbf{P}_{\mathbf{p}(s)}(\mathbf{d}) = 0.5,$$

or equivalently

$$\sum_{(d_1,\ldots,d_C)\in\mathcal{R}(\phi_{mult})} \prod_{c=1}^{C} \binom{n_c}{d_c} \left((1-s)\hat{p}_c + s\right)^{d_c} \left(1-(1-s)\hat{p}_c - s\right)^{n_c-d_c} = 0.5$$

For one-sided tests the power is continuous and monotonically increasing in s. Thus s is unique. In general, to solve for s, numerical iterative methods are required.

The parameterization of the **second specification** of the alternative hypothesis, H_1B, works similar to H_1A except that underperformance is assumed in exactly one of the C rating classes. Therefor we set for $i = 1, ..., C$ the alternative

$$\begin{aligned} H_1B_i: \quad & p_i = (1-s(i))\hat{p}_i + s(i) \quad \text{and} \\ & p_c = \hat{p}_c \qquad\qquad for \quad c \neq i \end{aligned}$$

Then $s(i)$ is solved such that the benchmark test gives a certain power under H_1B_i, i.e.

$$\mathrm{E}_{H_1B_i}\phi_{mult} = \mathrm{E}_{\mathbf{p}(s(i))}\phi_{mult} = \sum_{\mathbf{d}\in\mathcal{R}(\phi_{mult})} \mathbf{P}_{\mathbf{p}(s(i))}(\mathbf{d}) = 0.5.$$

For some test ϕ power is then computed by the average over power under underperformance in each class i, i.e.

$$Power(\phi) = \frac{1}{C}\sum_{i=1}^{C} \mathrm{E}_{H_1B_i}\phi = \frac{1}{C}\sum_{i=1}^{C} \mathrm{E}_{\mathbf{p}(s(i))}\phi.$$

For the **third specification** of the alternative hypothesis, H_1C, we assume the $C-$dimensional normal distribution for the true parameter $\mathbf{p}$, i.e.

$$\mathbf{p} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

with mean vector $\boldsymbol{\mu}$ given by

$$\mu_c = \hat{p}_c$$

for all c and covariance matrix is diagonal and given by

$$\Sigma_{cc} = \min_{i=1,\ldots,C-1} (\hat{p}_{i+1} - \hat{p}_i)^2$$

for all c and $\Sigma_{cd} = 0$ for all $c \neq d$. We further condition on the event H_1 to ensure a probability mass of one on H_1. Thus the conditioned pdf $f^{|H_1}$ of the true parameter is given by

$$f^{|H_1}_{\boldsymbol{\mu},\Sigma}(\mathbf{p}) = \mathcal{P}(\mathbf{p}|H_1) = \frac{f_{\boldsymbol{\mu},\Sigma}(\mathbf{p}) \cdot \mathbf{1}_{H_1}}{\mathcal{P}(H_1)} = \frac{f_{\boldsymbol{\mu},Sigma}(\mathbf{p}) \cdot \mathbf{1}_{H_1}}{1 - \prod_{c=1}^{C} \Phi_c(\hat{p}_c)},$$

where $f_{\boldsymbol{\mu},\Sigma}$ denote the multivariate normal pdf and Φ_c denotes the univariate normal cdf of rating class c.

For a test ϕ its power is given by

$$\begin{aligned} \mathrm{E}\,\phi &= \int_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{d}\in\mathcal{R}_\phi} \mathbf{P}_{\mathbf{p}}(\mathbf{d})\; df^{|H_1}_{\boldsymbol{\mu},\Sigma}(\mathbf{p}) \\ &= \sum_{(d_1,\ldots,d_C)\in\mathcal{R}_\phi} \int_{\mathbf{p}\in H_1} \prod_{c=1}^{C} \binom{n_c}{d_c}\, p_c^{d_c}\; (1-p_c)^{n_c-d_c}\; df^{|H_1}_{\boldsymbol{\mu},\Sigma}(\mathbf{p}). \end{aligned}$$

In words, power is computed for each $\mathbf{p} \in H_1$ and then integrated over the given distribution on H_1. The implementation in the statistical package for R uses Monte Carlo simulation for the numerical integration.

Bibliography

A. Agresti and B.A. Coull. Approximation is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, 52:119–126, 1998.

W. Aussenegg, F. Resch, and G. Winkler. Pitfalls and remedies in testing the calibration quality of rating systems. *Journal of Banking and Finance*, 35: 698–708, 2011.

D. J. Bartholomew. A test of homogeneity for ordered alternatives. *Biometrika*, 46(1/2):36–48, 1959.

Basel Committee on Banking Supervision. Studies on the validation of internal rating systems. *Working Paper*, 14, 2005.

A. Blöchlinger and M. Leippold. Economic benefit of powerful credit scoring. *Journal of Banking and Finance*, 30:851–873, 2006.

C. Clopper and S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.

F. Coppens, F. González, and G. Winkler. The performance of credit rating systems in the assessment of collateral used in eurosystem monetary policy operations. *ECB Occasional Paper Series*, 65, 2007.

DBRS. Rating CLOs backed by loans to european small and medium-sized enterprises (SMEs). `http://www.dbrs.com/research/261652`, 2013.

M. DeGroot and M. Schervish. *Probability and Statistics*. Addison-Wesley, Boston, San Francisco, New York, third edition, 2002.

A. Dmitrienko and J. Hsu. *Multiple Testing in Clinical Trials*. John Wiley and Sons, Inc., 2004.

S. Döhler. Validation of credit default probabilities using multiple-testing procedures. *Journal of Risk Model Validation*, 4:59–92, 2010.

European Banking Authority (EBA). Implementing technical standards on the mapping of ECAIs credit assessments. `https://www.eba.europa.eu/regulation-and-policy/external-credit-assessment-institutions-ecai/draft-implementing-technical-standards-on-the-mapping-of-ecais-credit-assessments`, 2015.

Lloyd D. Fisher. The use of one-sided tests in drug trials: an FDA advisory committee member's perspective. *Journal of Biopharmaceutical Statistics*, 1 (1):151–156, 1991.

R.A. Fisher. *The design of experiments. 1935*. Oliver and Boyd, Edinburgh, 1935.

Fitch. Criteria for rating granular corporate balance-sheet securitisations (SME CLOs). `https://www.fitchratings.com/creditdesk/reports/report_frame.cfm?rpt_id=737615`, 2014.

R. Frey and A. McNeil. Dependent defaults in models of portfolio credit risk. *Journal of Risk*, 6:59–92, 2003.

K. Hornik, R. Jankowitsch, M. Lingo, S. Pichler, and G. Winkler. Validation of credit rating systems using multi-rater information. *Journal of Credit Risk*, 3(4):3–29, 2007.

D.W. Hosmer and S. Lemeshow. A goodness-of-fit test for the multiple logistic regression model. *Communication in Statisitics - Theory and Methods*, 10: 1043–1069, 1980.

R. Jankowitsch, S. Pichler, and W. Schwaiger. Modelling the economic value of credit rating systems. *Journal of Banking and Finance*, 31:181–198, 2007.

J.P. Krahnen and M. Weber. Generally accepted rating principles: A primer. *Journal of Banking and Finance*, 25(1):3–23, 2001.

E. Lehmann and J. Romano. *Testing Statistical Hypotheses*. Springer, 2006.

M. Lingo and G. Winkler. Discriminatory power - an obsolete validation criterion? *Journal of Risk Model Validation*, 4:4–24, 2008.

Moody's. Moody's global approach to rating SME balance sheet securitizations. `https://www.moodys.com/researchdocumentcontentpage.aspx?docid=PBS_SF393152`, 2015.

P. C. O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4):1079–1087, 1984.

M. D. Perlman. One-sided testing problems in multivariate analysis. *Ann. Math. Statist.*, 40(2):549–567, 1969.

J. Reiczigel. Confidence intervals for the binomial parameter: Some new considerations. *Statistical Methods in Medical Research*, 22:611–621, 2003.

J. Reiczigel, Z. Abonyi-Tóth, and J. Singer. An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio of proportions. *Computational Statistics & Data Analysis*, 52(11):5046–5053, 2008.

Standard & Poor's. Annual global corporate default study and rating transitions. `http://www.nact.org/resources/NACT_2012_Global_Corporate_Default.pdf`, 2012.

Standard & Poor's. European SME CLO methodology and assumptions. `https://www.globalcreditportal.com/ratingsdirect/showArticlePage.do?articleId=1287443`, 2013.

T.E. Sterne. Some remarks on confidence or fiducial limits. *Biometrika*, 41: 275–278, 1954.

David A. van Dyk. The role of statistics in the discovery of a higgs boson. *Annual Review of Statistics and Its Application*, 1(1):41–59, 2014.

S.E. Vollset. Confidence intervals for a binomial proportion. *Statistics in Medicine*, 12:809–824, 1993.

P.H. Westfall and R.D. Wolfinger. Multiple tests with discrete distributions. *The American Statistician*, 51:3–8, 1997.

Acknowledgements
We would like to thank Fernando González, Manuel Lingo, Andrés Manzanares, Mario Vackier and Gerhard Winkler and two anonymous referees for their helpful comments. Klaas Schulze worked on this paper mainly during his secondment from Deutsche Bundesbank to the European Central Bank.

François Coppens
Nationale Bank van België / Banque nationale de Belgique, Brussels, Belgium.
e-mail: francois.coppens@nbb.be

Manuel Mayer
Oesterreichische Nationalbank, Wien, Austria;
e-mail: manuel.mayer@oenb.at

Laurent Millischer (corresponding author)
European Central Bank, Frankfurt am Main, Germany;
e-mail: laurent.millischer@ecb.europa.eu

Florian Resch
Oesterreichische Nationalbank, Wien, Austria;
e-mail: florian.resch@oenb.at

Stephan Sauer
European Central Bank, Frankfurt am Main, Germany;
e-mail: stephan.sauer@ecb.europa.eu

Klaas Schulze
Deutsche Bundesbank, Frankfurt am Main, Germany;
e-mail: klaas.schulze@bundesbank.de

Postal address	60640 Frankfurt am Main, Germany
Telephone	+49 69 1344 0
Website	www.ecb.europa.eu

ISSN	1725-2806 (online)
ISBN	978-92-899-2001-8
DOI	10.2866/544500
EU catalogue No	QB-AR-16-002-EN-N