Karlsruher Institut für Technologie

Model Diagnostics meets Forecast Evaluation: Goodness-of-Fit, Calibration, and Related Topics

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
Johannes Resin, M. Sc.
aus
Lörrach

Tag der mündlichen Prüfung: 15. Februar 2023
1. Referent: Prof. Dr. Tilmann Gneiting
2. Referent: Prof. Dr. Norbert Henze

Abstract

Principled forecast evaluation and model diagnostics are vital in fitting probabilistic models and forecasting outcomes of interest. A common principle is that fitted or predicted distributions ought to be calibrated, ideally in the sense that the outcome is indistinguishable from a random draw from the posited distribution. In much of this thesis, I deal with questions centered on calibration properties of various types of forecasts.

In the first part of this thesis, I propose a simple algorithm for exact multinomial goodness-of-fit tests, which computes exact p-values based on various test statistics, such as the log-likelihood ratio and Pearson's chi-square. The algorithm is accompanied by a thorough analysis illustrating improvement on extant methods. However, the use of the algorithm is limited to multinomial distributions with a small number of categories as the runtime grows exponentially in said number.

For real-valued outcomes, a general theory of calibration has been elusive, despite a recent surge of interest in distributional regression and machine learning. In the second part of this thesis, a framework rooted in probability theory is developed, which gives rise to hierarchies of calibration, and applies to both predictive distributions and stand-alone point forecasts. A prediction is conditionally T-calibrated if it can be taken at face value in terms of an identifiable functional T. Based on this general notion of calibration, the thesis introduces population versions of T-reliability diagrams and revisits a score decomposition into measures of miscalibration, discrimination, and uncertainty. In empirical settings, stable and efficient estimators of T-reliability diagrams and score components arise via nonparametric isotonic regression and the pool-adjacent-violators algorithm. For in-sample model diagnostics, the thesis introduces a universal coefficient of determination that nests and reinterprets the classical R^2 in least squares regression and its natural analog R^1 in quantile regression, yet applies to T-regression in general.

In the face of uncertainty, the need for probabilistic assessments has long been recognized in the literature on forecasting. In classification, however, comparative evaluation of classifiers often focuses on predictions specifying a single class through the use of simple accuracy measures, which disregard any probabilistic uncertainty quantification. I propose probabilistic top lists as a novel type of prediction in classification, which bridges the gap between single-class predictions and predictive distributions. The probabilistic top list functional is elicitable through the use of strictly consistent evaluation metrics. The proposed evaluation metrics are based on symmetric proper scoring rules and admit comparison of various types of predictions ranging from single-class point predictions to fully specified predictive distributions. The Brier score yields a metric that is particularly well suited for this kind of comparison.

Zusammenfassung

Die fundierte Bewertung von Modellen und Vorhersagen ist entscheidender Bestandteil der Anpassung probabilistischer Modelle und der Vorhersage ungewisser Größen und Ereignisse. Sowohl aus Sicht der Modelldiagnose als auch bei der Vorhersageauswertung sollten postulierte Verteilungen dem Kalibrationsprinzip entsprechen, d. h. im besten Fall sollte sich die Beobachtung nicht von einer zufälligen Ziehung aus der behaupteten Verteilung unterscheiden lassen Ein beträchtlicher Teil dieser Arbeit widmet sich Fragestellungen, welche sich mit der Kalibration verschiedener Arten von Vorhersagen befassen.

Der erste Teil dieser Arbeit schlägt einen neuen Algorithmus zur exakten Berechnung multinomialer Anpassungstests vor. Der Algorithmus berechnet exakte p-Werte zu verschiedenen Teststatistiken wie dem Likelihood-Quotienten oder Pearsons Chi-Quadrat-Statistik. Eine eingehende Untersuchung belegt eine verbesserte Laufzeit des Algorithmus im Vergleich zu bestehenden Methoden. Die Anwendbarkeit des Algorithmus beschränkt sich jedoch auf Multinomialverteilungen in wenigen Ausprägungen, da die Laufzeit exponentiell in deren Anzahl wächst.

Trotz eines wachsenden Interesses an Verteilungsregression und maschinellem Lernen in jüngster Zeit ist eine allgemeine Kalibrationstheorie für reellwertige Größen schwer zu fassen. Im zweiten Teil dieser Arbeit wird ein wahrscheinlichkeitstheoretisches Modell erarbeitet, in welchem hierarchische Zusammenhänge zwischen diversen Kalibrationsbegriffen aufgedeckt werden. Das Modell berücksichtigt sowohl Vorhersagen in Form von Wahrscheinlichkeitsverteilungen als auch einzelne Punktvorhersagen. Eine Vorhersage ist bedingt T-kalibriert, wenn sie bezüglich eines identifizierbaren Funktionals T beim Wort genommen werden kann. Basierend auf diesem allgemeinen Kalibrationsbegriff werden theoretische T-Kalibrationskurven vorgestellt und die Zerlegung von Vorhersagebewertungen in Maße für die Misskalibration und Trennschärfe einer Vorhersage sowie ein Maß für die zugrundeliegende Unsicherheit aufgegriffen. Für die empirische Anwendung ergeben sich mithilfe der nichtparametrischen isotonen Regression eindeutige, effiziente Schätzer der T-Kalibrationskurven und Bewertungszerlegungen. Zur Modelldiagnose bei allgemeinen T-Regressionen wird in dieser Arbeit ein universelles Bestimmtheitsmaß eingeführt, welches sowohl das klassische R^2 aus der Kleinste-Quadrate-Schätzung als auch sein natürliches Äquivalent R^1 aus der Quantilsregression umfasst.

In der Literatur zum Thema Vorhersagen ist die Notwendigkeit der probabilistischen Quantifizierung unvermeidbarer Ungewissheiten längst bekannt. Zur vergleichenden Bewertung statistischer Klassifikationsverfahren wird dahingegen oft nur die Treffgenauigkeit einer vorhergesagten Klassenausprägung herangezogen, wodurch jegliche Quantifizierung bestehender Unsicherheiten unberücksichtigt bleibt. In dieser Arbeit werden probabilistische Toplisten als neuartiger

Klassifikationstyp eingeführt. Diese lassen sich als Kompromiss zwischen Vorhersagen einer einzelnen Klassenausprägung und Vorhersagen ganzer Verteilungen über alle Klassenausprägungen verstehen. Das zugehörige Funktional lässt sich durch den Gebrauch von streng konsistenten Verlustfunktionen elizitieren. Die zu diesem Zweck in dieser Arbeit eingeführten Verlustfunktionen basieren auf symmetrischen korrekten Bewertungsregeln und ermöglichen es, verschiedene Vorhersagetypen zu vergleichen. Die Vorhersagen können sich dabei von Vorhersagen einzelner Klassenausprägungen bis hin zu Vorhersagen ganzer Verteilungen erstrecken. Die Brier-Bewertung führt zu einer Metrik, welche besonders geeignet für derlei Vergleiche ist.

Acknowledgements

I am deeply grateful to Tilmann Gneiting, under whose supervision I had the privilege of working for the last four and half years. Despite his many duties, he always took time to listen to my questions and concerns, and discuss matters at hand. It was a pleasure to collaborate with him on our joint research, and I would like to express my gratitude for his guidance and support. I would like to thank Norbert Henze for his willingness to review this thesis.
During my doctoral studies, I have benefited from the help, advice, and expertise of many people. I am thankful to the Computational Statistics Group at the Heidelberg Institute for Theoretical Studies and the members of the Institute of Stochastics at the Karlsruhe Institute of Technology for contributing to a stimulating work environment. At the risk of providing an incomplete list, I would like to thank Johannes Bracher, Jonas Brehmer, Timo Dimitriadis, Kira Feldmann, Tobias Fissler, Alexander Jordan, Nils Koster, Fabian Krüger, Sebastian Lerch, Mark-Oliver Pohle, Ghulam Qadir, Patrick Schmidt, Benedikt Schulz, Peter Vogel, Eva-Maria Walz, Daniel Wolffram, and Johanna Ziegel for interesting discussions, helpful comments and advice, and other contributions to an enjoyable time as a Ph.D. student. Valuable feedback from many people, including colleagues, reviewers, and editors, has helped to improve the research presented in this thesis.
I am grateful to Satoshi Kuriki for his willingness to host a research stay at the Institute of Statistical Mathematics, which unfortunately had to be canceled due to the COVID-19 pandemic.
I had the privilege of working at a wonderful research institute and I am thankful to all the people whose hard work provides such an enjoyable work environment, including the administration, the communications team, IT support, and the kitchen and cleaning staff. I gratefully acknowledge financial and infrastructural support from the Heidelberg Institute for Theoretical Studies and the Klaus Tschira Foundation.
On a personal note, I am grateful to my family and friends, who have shaped me as a person. In particular, I would like to thank my parents for their lifelong unconditional support and my wife for her enduring support and companionship.

Contents

1 | Introduction

Typically, we think of forecasts as statements about future events or outcomes. In the face of uncertainty, such statements should accurately reflect the level of confidence placed in them. Reliable forecasts are vital in decision-making. Loosely speaking, reliability or calibration refers to the agreement between forecasts and realizations. Ideally, forecasts are probabilistic (Gneiting and Katzfuss, 2014), that is, they take the form of predictive probability distributions that reflect the outcome's behavior conditional on a forecaster's present knowledge. A probabilistic forecast may be of value to multiple forecast users faced with all kinds of decision problems. Nonetheless, point forecasts that predict a specific value or outcome are frequently encountered in practice (Gneiting, 2011a). Meaningful point forecasts are related to the conditional law of the variable of interest through a statistical functional or as the Bayes act minimizing an expected loss. Typically, point forecasts (and statistical functionals) relate to specific types of decision problems. For example, quantiles naturally arise in the solution to the classical newsvendor problem (Gneiting, 2011b).

In this thesis, I focus on forecast evaluation. A thorough review of forecasting methods and applications is given by Petropoulos et al. (2022). Much of the early literature on forecast evaluation has been driven by meteorological problems (e.g., Brier, 1950; Murphy, 1977). Modern weather forecasts are the result of a great collaborative effort and a pinnacle of human achievement (e.g., Bauer et al., 2015; Benjamin et al., 2018). Weather forecasts are widely used (e.g., in aviation, agriculture, and, of course, to ensure public safety) and have great social and economic impact. Much of the progress in weather prediction has been driven by human understanding of the physical processes governing the weather and advancements in computing capabilities, yet principled forecast verification remains an important topic in the meteorological community and has certainly contributed to forecast quality.

Forecast evaluation is not limited to forecasts of future events but also applies to various kinds of predictions encountered, for example, in statistical classification and supervised learning. In an uncertain world, even the present state may not be without doubt. Nowcasts provide estimates of summary statistics resulting from complex systems, for example, in disease monitoring (`https://covid19nowcasthub.de`) or when it comes to macroeconomic measures such as gross domestic product (GDP; Giannone et al., 2008).

In modern times, decision-making is not exempt from automation. Yet, many decisions entail severe consequences for individual people, and decisions based on misguided forecasts may disproportionately affect minority groups. Popular examples of such delicate endeavors include the prediction of recidivism in pre-

trial defendants (Corbett-Davies et al., 2017) and automation of hiring decisions (Raghavan et al., 2020). Such issues have fueled the debate on ethical automated decision-making — often under terms such as algorithmic or machine (learning) bias and fairness — in the academic community (Mitchell et al., 2021). While careful and critical evaluation of forecasts is crucial in avoiding adverse decisions, reliable prediction systems may benefit human decision-making (Kleinberg et al., 2017).

Model diagnostics and forecast evaluation are closely related and often employ similar means in different contexts. To start with, model fitting is related to forecast evaluation as one often optimizes the parameters of a model in such a way as to optimize a given loss function. I broadly speak of model diagnostics as in-sample validation, where the same data is used for model fitting and verification. In contrast, I refer to out-of-sample validation, where new data is used for verification purposes, as forecast evaluation in a broad sense.

Subject to calibration, a sharper predictive distribution (i.e., a probabilistic forecast with less uncertainty) typically leads to better decisions incurring smaller losses (Gneiting et al., 2007). Among calibrated point forecasts, high variability is preferable (Krüger and Ziegel, 2021). Scoring rules and scoring or loss functions provide summary measures of overall forecast performance and are widely used to evaluate and compare forecasts. Many scoring functions admit interpretation as a loss in certain decision problems. Score decompositions offer further insights, typically providing measures of (mis-)calibration and discriminative ability. Yet empirical estimation (and interpretation) of these decompositions can be challenging. Reliability diagrams and related graphical tools complement in-depth forecast assessment through insightful visualizations of calibration properties.

Chapter 2 collects key notation and terminology in a general framework used throughout this thesis and introduces some fundamental tools used in forecast evaluation. To make for an easy individual read, the main chapters in this thesis are self-contained, (re-)introducing relevant notation and terminology throughout.

Chapter 3 revisits a classical test from the statistical literature, the simple multinomial test. The chapter starts with a brief review that focuses on various popular test statistics used to conduct multinomial testing. As the main contribution of this chapter, a simple algorithm to speed up the computation of exact multinomial tests is proposed. The theoretical validation of the algorithm is complemented by a detailed analysis in a simulation study illustrating the good performance of the algorithm, while also highlighting differences between exact and asymptotic tests based on three popular test statistics. Finally, an application linking multinomial tests to the evaluation of ternary probability forecasts is outlined.

Chapter 4 investigates various notions of calibration for forecasts of real-valued outcomes. As a result, hierarchies of calibration highlighting the intricate connections between the various notions are presented. The chapter focuses on a general notion of conditional T-calibration in terms of a statistical functional, T, which applies to probabilistic forecasts as well as point forecasts. After a theoretical treatment in the prediction space setting of Gneiting and Ranjan (2013),

the chapter turns to empirical forecast evaluation. A powerful generalization of the CORP approach by Dimitriadis et al. (2021) to identifiable statistical functionals is introduced, which results in empirical T-reliability diagrams and score decompositions. Finally, a connection to model diagnostics is drawn through a general coefficient of determination, which is closely linked to skill scores in forecast evaluation. The chapter is complemented by a brief treatment of quantile forecasts and a comprehensive appendix providing further insights.
Chapter 5 introduces probabilistic top list functionals, which specify a number of most likely classes along with predicted individual class probabilities, to statistical classification. The top list functional bridges the gap between simply predicting the mode, i.e., a single most likely class, and a full probability distribution. Probabilistic top lists appear as a promising middle ground, especially in settings where the confidence placed on a single predicted class is underwhelming, while the outcome can be narrowed down to a few classes with high confidence. In particular, prediction practice in multi-label classification, where the specification of predictive distributions is hindered by a large number of classes, may benefit from such a probabilistic approach. The chapter introduces padded symmetric scores based on symmetric proper scoring functions as a consistent way of evaluating top list predictions. In particular, it is shown that the probabilistic top list functional is elicitable. The proposed scores admit a balanced comparison of various types of predictions while encouraging truthful probabilistic assessments.
The thesis concludes with Chapter 6, which collects and discusses key results and avenues for future research.

1.1 Declaration: Previous and published work

Parts of this thesis are adapted from the following research articles with significant contributions by myself (in order of first appearance):

Resin, J. (2022). A simple algorithm for exact multinomial tests. *Journal of Computational and Graphical Statistics*. In press, DOI:10.1080/10618600.2022.2102026.

Gneiting, T. and Resin, J. (2021). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams and coefficient of determination. Preprint, arXiv:2108.03210v3.

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., and Schienle, M. (2023). Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and Its Application*, 10. In press, DOI:10.1146/annurev-statistics-032921-020240.

Parts of the abstract are adapted from the abstract of Gneiting and Resin (2021). Chapter 3 is a modified version of Resin (2022). Chapter 4 is an extended version

of Gneiting and Resin (2021). The chapter contains some additional content from Gneiting et al. (2023). In particular, Section 4.4.3 reproduces Section 4.1 of Gneiting et al. (2023), which is subject to copyright held by the Annual Reviews (`https://www.annualreviews.org`).
As part of my thesis, I worked on the following `R` packages:

Resin, J. (2021a). *CalSim: The calibration simplex*. R package version 0.5.2 at `https://CRAN.R-project.org/package=CalSim`.

Resin, J. (2020). *ExactMultinom: Multinomial Goodness-of-Fit Tests*. R package version 0.1.2 at `https://CRAN.R-project.org/package=ExactMultinom`.

The `ExactMultinom` package implements ideas from Resin (2022), whereas the `CalSim` package implements the *calibration simplex*, which was proposed by Wilks (2013).
Replication material for Gneiting and Resin (2021) and Gneiting et al. (2023) is available through the following resources:

Resin, J. (2021b). *Replication code for Gneiting and Resin (2021)*. `https://github.com/resinj/replication_GR21`.

Wolffram, D., Resin, J., Kraus, K., and Jordan, A. I. (2022). *Replication package for "Model Diagnostics and Forecast Evaluation for Quantiles"*. DOI:10.5281/zenodo.6546490.

2 | Preliminaries on Forecast Evaluation

This chapter introduces key notation and terminology adapted from the pertinent literature on forecast evaluation and probabilistic forecasting. The touched-upon concepts feature prominently throughout this thesis.

The theory on forecast evaluation is centered on the question of what constitutes a good forecast. A forecast should provide an honest and insightful assessment of an uncertain actuality while reflecting its own limitations. Such behavior is encouraged by the use of suitable evaluation metrics, which quantify the value of a given forecast and admit comparison of competing forecasts (Gneiting and Katzfuss, 2014).

From a theoretical viewpoint, forecast evaluation is concerned with the joint distribution of the outcome Y of interest and a forecast F on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. This approach to forecast evaluation dates back to the seminal paper by Murphy and Winkler (1987). The framework and adaptations thereof are often referred to as prediction space (Gneiting and Ranjan, 2013), which is introduced in detail in Chapter 4. The outcome or variable of interest is a random variable $Y\colon \Omega \to \mathcal{Y}$, which takes values in some sample space $\mathcal{Y}$. Throughout this thesis, outcomes are assumed to be real-valued, i.e., $\mathcal{Y} \subseteq \mathbb{R}$, or categorical. Categorical variables, as encountered in statistical classification, map to a finite set of m categories or classes, which are frequently identified with the labels $1, \dots, m$, i.e., $\mathcal{Y} = \{1, \dots, m\}$. A probabilistic forecast $F\colon \Omega \to \mathcal{F}$ is a random probability distribution from a set $\mathcal{F}$ of probability measures on the sample space $\mathcal{Y}$, while a point forecast is a random variable $X\colon \Omega \to \mathcal{Y}$ providing a single-valued prediction from the sample space $\mathcal{Y}$. In the case of a real-valued outcome, probability distributions are typically identified with their cumulative distribution function (CDF). In the categorical case, the set of all categorical distributions is usually considered, and distributions are identified with probability vectors, thereby identifying the set $\mathcal{F}$ with the probability simplex

$$\Delta_{m-1} = \{p = (p_1, \dots, p_m) \in [0,1]^m \mid p_1 + \cdots + p_m = 1\}.$$

Binary variables constitute an important special case of categorical variables. In the case of a binary variable, probabilistic forecasts are completely specified by the probability assigned to one of the classes, thereby identifying the set $\mathcal{F}$ with the unit interval $[0, 1]$.[1]

[1] In binary classification, the specified event probability typically refers to the class labeled 1, while the other class is labeled 0.

Ideally, a probabilistic forecast matches the conditional distribution of Y given the information that is available at the time of forecasting. On the other hand, point forecasts require guidance in the form of a statistical functional or a loss function, which relates them to the (conditional) distribution of Y, as discussed in Section 2.2.

2.1 Evaluation of Probabilistic Forecasts

On account of providing a distribution on possible outcomes, a probabilistic forecast naturally quantifies the remaining uncertainty, thereby reflecting its limitations. The theoretical framework lends itself naturally to the study of eligible properties of probabilistic forecasts in the population. In contrast, empirical practice is complicated by the fact that the underlying distributions are unknown and need to be inferred from a sample of forecasts and respective outcomes. This discrepancy becomes apparent in the following section, which briefly introduces calibration as a theoretical concept, before presenting some practical tools for calibration checks.

2.1.1 Calibration

In general, calibration refers to the agreement between the forecast F and the observed outcome Y. Ideally, a probabilistic forecast matches a conditional distribution of Y, as formalized by the concept of auto-calibration (Tsyplakov, 2013). The forecast F is *(auto-)calibrated* if the conditional distribution of Y given F matches F, i.e.,

$$Y \mid F \sim F.$$

Auto-calibration is a strong requirement that is difficult to verify in practice. Therefore, extant practice typically relies on weaker notions of calibration such as probabilistic calibration (Dawid, 1984; Diebold et al., 1998), which gives rise to simple calibration checks. If distributions in $\mathcal{F}$ are assumed to be continuous, the probability integral transform (PIT) is defined simply as $Z_F = F(Y)$. Points of discontinuity require special treatment (see Chapter 4). The PIT is uniformly distributed if the forecast F is calibrated. In light of this necessary condition, the forecast F is called probabilistically calibrated if the PIT is uniformly distributed, i.e.,

$$Z_F \sim \mathcal{U}([0, 1]).$$

However, probabilistic calibration is not sufficient for calibration (Gneiting and Ranjan, 2013).

2.1.1.1 PIT histograms

PIT histograms (Diebold et al., 1998) are a popular tool to assess (probabilistic) calibration. As the name suggests, PIT histograms approximate the empirical distribution of the realized PIT values by means of a simple histogram. Figure

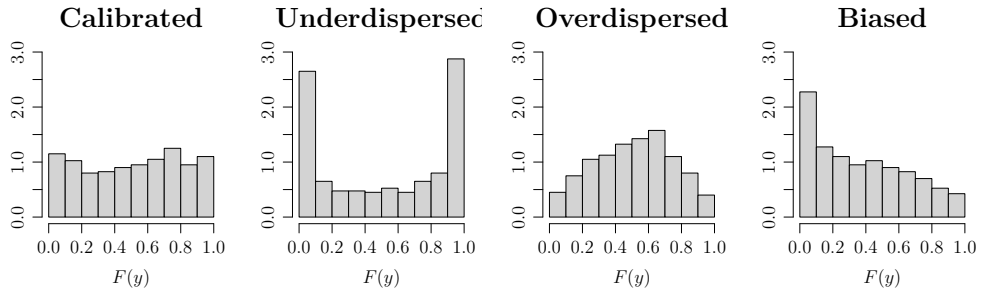


Figure 2.1 PIT histograms based on samples of size 400 for the calibrated forecast F_1, the underdispersed forecast $F_{1/2}$, the overdispersed forecast $F_{3/2}$, and the biased forecast F_{biased} from Example 2.1.

2.1 shows some exemplary PIT histograms based on the forecasts in Example 2.1. A forecast is underdispersed if the variance of the PIT exceeds the variance of a uniform random variable. It is overdispersed if the variance is too low. Underdispersion often results in a characteristic $\cup$-shape in the PIT histogram, whereas overdispersion often yields a $\cap$-shaped histogram. In Chapter 4, an alternative PIT reliability diagram is discussed, which visualizes the empirical distribution without approximation through arbitrary binning.

Example 2.1 (Gneiting and Ranjan (2013))**.** Let $\mu \sim \mathcal{N}(0, 1)$ be standard normal, and $Y \mid \mu \sim \mathcal{N}(\mu, 1)$ be conditionally normal with mean μ and variance 1. The forecasts $F_\sigma = \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$ is calibrated if $\sigma = 1$. It is underdispersed if $\sigma < 1$ and overdispersed if $\sigma > 1$ (Gneiting and Ranjan, 2013). The forecast $F_{\text{biased}} = \mathcal{N}(\mu + \frac{1}{2}, 1)$ is unconditionally biased. Figure 2.1 shows simulated PIT histograms for the different types of forecasts in this example.

2.1.1.2 Traditional reliability diagrams

In binary classification, auto-calibration is the uncontested gold standard of calibration. Calibration of a binary probability forecast is frequently assessed in reliability diagrams (e.g., Bröcker and Smith, 2007). The traditional reliability diagram plots empirical event frequencies against average forecast probabilities conditional on the forecast falling within given subintervals of the unit interval, which are usually referred to as bins. Inset histograms are typically used to indicate the number of forecasts per bin. Figure 2.2 shows some exemplary reliability diagrams based on the forecasts in Example 2.2. The figure contains consistency bars, as suggested by Bröcker and Smith (2007), which visualize the spread of resampled event frequencies. Resamples are obtained from the forecast distributions themselves, thereby demonstrating how much the empirical event frequencies may deviate from calibrated probabilities by mere chance. Here, consistency bars range from the 5th percentile to the 95th percentile of the resampled event frequencies. A forecast is overconfident if the probability assigned to the most likely class is systematically overstated, while it is underconfident if said probability is systematically understated.

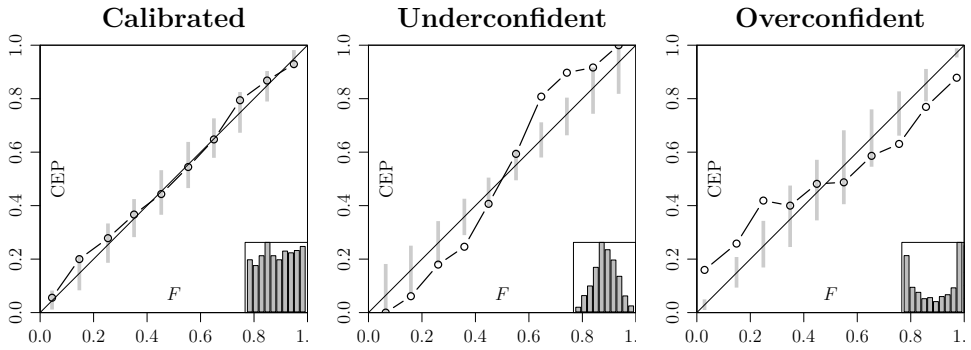


Figure 2.2 Traditional reliability diagrams based on samples of size 1000 for the calibrated forecast F_1, the underconfident forecast $F_{1/2}$, and the overconfident forecast F_2 from Example 2.2 with 90% consistency bars.

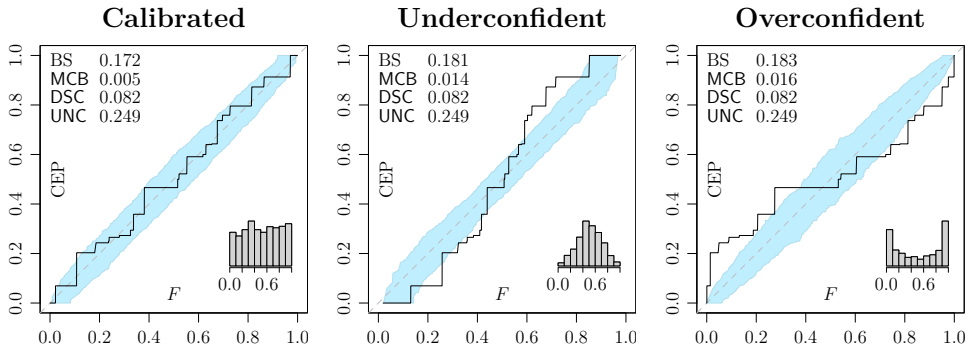


Figure 2.3 CORP reliability diagrams based on the same samples as Figure 2.2 with 90% consistency bands and CORP score components of the Brier score (BS) as described in Section 2.1.2.1.

Example 2.2. Let $p \sim \mathcal{U}([0,1])$ be uniformly distributed on the unit interval, and $Y \mid p \sim \text{Bin}(1,p)$ be a Bernoulli random variable with event probability p. The probability forecast $F_a = (1 + (\frac{1-p}{p})^a)^{-1}$ with $a > 0$ is calibrated if $a = 1$. It is underconfident if $a < 1$ and overconfident if $a > 1$. Figures 2.2 and 2.3 show simulated reliability diagrams for the different types of forecasts in this example.

2.1.1.3 CORP reliability diagrams

Dimitriadis et al. (2021) argue that the traditional reliability diagrams are compromised by the fact that they depend on arbitrary binning choices, which may distort the diagrams (as illustrated in Figure 2 of Dimitriadis et al. (2021)) and propose the CORP reliability diagram as a stable alternative. The CORP reliability diagram shows an isotonic regression fit, which estimates the calibration curve $p \mapsto \mathbb{P}(Y = 1 \mid F = p)$ mapping probability forecasts to conditional event probabilities (CEP). Isotonic regression fits are computed using the pool-adjacent violators (PAV) algorithm. The estimated calibration curves are

Consistent (if the calibration curve is monotonically increasing),

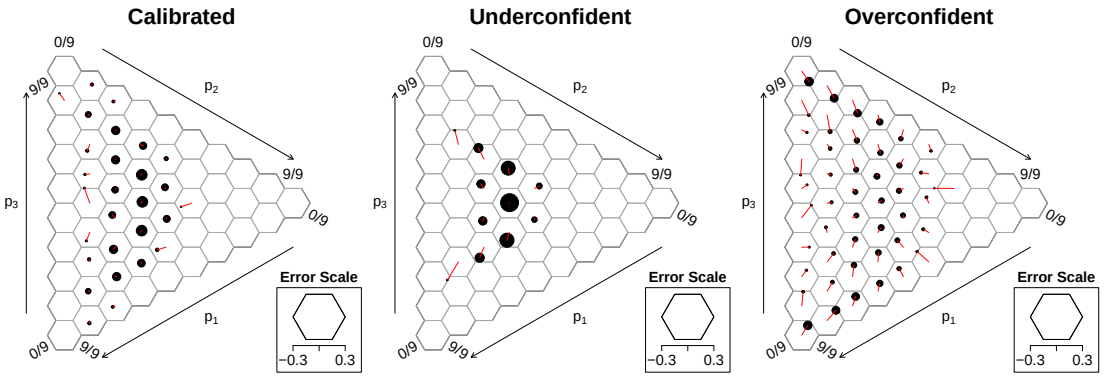


Figure 2.4 Calibration simplexes based on 5000 samples for the calibrated forecast F_{cal}, the underconfident forecast F_{under}, and the overconfident forecast F_{over} from Example 2.3.

Optimal (in the sense that recalibrated values optimize the empirical score under any proper scoring rule subject to monotonicity),
Reproducible, and
PAV based,

hence the acronym CORP. Inset histograms are used to show the empirical distribution of the forecasts. Figure 2.3 shows CORP reliability diagrams for the forecasts from Example 2.2. Consistency bands are obtained from resampled calibration curves. In large samples, asymptotic theory gives rise to confidence bands. Dimitriadis et al. (2021) hint at a generalization to point forecasts of statistical functionals, which is investigated in Chapter 4.

2.1.1.4 The calibration simplex

Wilks (2013) proposed a generalization of the reliability diagram to probabilistic forecasts of a ternary outcome, i.e., in the case of $m = 3$ classes. In this case, the probability simplex can be visualized in the plane. The simplex is subdivided into hexagonal bins by a regular tessellation, i.e., using a grid made up of hexagons. The resulting reliability diagram is called a *calibration simplex*. Figure 2.4 shows calibration simplexes for the forecasts from Example 2.3. The solid circles represent the forecasts in each bin. The area of each circle is proportional to the number of forecasts in its respective bin. If the average forecast probabilities in a bin do not match the empirical frequencies precisely, the circle is shifted towards underforecast outcomes. The magnitude of the shift, as alluded to by an inset error scale referring to a single bin, is proportional to the difference between predicted probabilities and observed frequencies.

Example 2.3 (Wilks (2013))**.** The construction in this example is similar to Wilks (2013), with the difference that Wilks uses the parameter $\bar{p}$ as forecast and uses outcomes distributed according to the different forecast distributions. Let

$p = (p_1, p_2, p_3)$ be a random vector, such that $p_2 = \max\{0, 1 - p_1 - p_3\}$ and the vector

$$(\log(\tfrac{p_1}{1-p_1}), \log(\tfrac{p_3}{1-p_3})) \sim \mathcal{N}\left((\log(\tfrac{1}{2}), \log(\tfrac{1}{2})), \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix}\right)$$

is bivariate normal with variances $\sigma^2 = 1$ and correlation coefficient $\rho = -0.95$. Let $Y \mid p \sim \bar{p}$ follow a categorical distribution with parameter $\bar{p} = \frac{p}{p_1+p_2+p_3}$. The forecast $F_{\text{cal}} = \bar{p}$ is calibrated. Let

$$q = (\max(0, 2\bar{p}_1 - \tfrac{1}{3}), \max(0, 2\bar{p}_2 - \tfrac{1}{3}), \max(0, 2\bar{p}_3 - \tfrac{1}{3})).$$

The overconfident forecast $F_{\text{over}} = \frac{q}{q_1+q_2+q_3}$ inflates probabilities larger than $\frac{1}{3}$ at the expense of small probabilities, whereas the underconfident forecast $F_{\text{under}} = ((\bar{p}_1 + \frac{1}{3})/2, (p_2 + \frac{1}{3})/2, (\bar{p}_3 + \frac{1}{3})/2)$ deflates large probabilities to the benefit of small probabilities. Figure 2.4 shows simulated calibration simplexes for the three forecasts in this example.

As with the binary reliability diagrams, it may be difficult to judge the gravity of the shifts observed in the simplex. To this end, I propose the use of color-coded multinomial p-values in Chapter 3. The calibration simplex is implemented in the `R` package `CalSim` (Resin, 2021a).

2.1.2 Proper scoring rules

Perfect calibration may be difficult to achieve in practice, and a calibrated forecast does not have to be especially insightful. Scoring rules provide summary measures that quantify the value of a probabilistic prediction. This section provides a short introduction to proper scoring rules by introducing a characterization result from Gneiting and Raftery (2007).

A *scoring rule* $\mathrm{S}\colon \mathcal{F} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is a function that assigns a score $\mathrm{S}(F, y)$ from the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cap \{-\infty, \infty\}$ to a predicted probability distribution F if the outcome y is observed. A scoring rule S is *regular* if the set $\mathcal{F}$ is convex and for all $F, P \in \mathcal{F}$ and $Y \sim P$ the expected score $\mathbb{E}[\mathrm{S}(F, Y)]$ exists and is finite with the sole exception of allowing an infinite expected score, i.e., $\mathbb{E}[\mathrm{S}(F, Y)] = \infty$, if $F \neq P$. A scoring rule S is *proper* if the true distribution minimizes the expected score, i.e,

$$\mathbb{E}[\mathrm{S}(P, Y)] \leq \mathbb{E}[\mathrm{S}(F, Y)] \quad \text{for } Y \sim P \text{ and } F, P \in \mathcal{F}.$$

A proper scoring rule S is *strictly proper* if the expected score is minimized only by the true distribution, i.e., $\mathbb{E}[\mathrm{S}(P, Y)] = \mathbb{E}[\mathrm{S}(F, Y)]$ for $Y \sim P$ and $F, P \in \mathcal{F}$ implies $F = P$.

Gneiting and Raftery (2007, Theorem 1) show that a regular scoring rule is (strictly) proper if, and only if, it admits a representation in terms of a (strictly) concave function and a supertangent of said function, as follows. Any regular proper scoring rule can be written as

$$\mathrm{S}(F, y) = G(F) - \mathbb{E}[G^*(F, Y_F)] + G^*(F, y) \quad \text{for } Y_F \sim F \text{ and } F \in \mathcal{F}, y \in \mathcal{Y}, \tag{2.1}$$

where $G\colon \mathcal{F} \to \mathbb{R}$ is a concave function and $G^*\colon \mathcal{F} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is a supertangent of G, i.e., a function such that for all $F, P \in \mathcal{F}$ and $Y \sim P, Y_F \sim F$ the expectation $\mathbb{E}[G^*(P, Y)]$ exists and is finite, the expectation $\mathbb{E}[G^*(F, Y)]$ exists, and

$$G(P) \leq G(F) + \mathbb{E}[G^*(F, Y) - G^*(F, Y_F)]$$

holds. Vice versa, any regular scoring rule S of the form (2.1) is proper. A regular scoring rule of the form (2.1) is strictly proper if, and only if, G is strictly concave. The continuous ranked probability score (CRPS) given by

$$\mathrm{S}_{\mathrm{CRPS}}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{z \geq y\})^2 \, \mathrm{d}z$$

is a prominent example of a scoring rule for real-valued outcomes. The CRPS is a proper scoring rule if $\mathcal{F}$ is the set of Borel probability measures on $\mathbb{R}$. It is strictly proper if the set $\mathcal{F}$ is restricted to the Borel probability measures with finite first moment (Gneiting and Raftery, 2007).
Further examples are provided in Chapter 5, which reintroduces proper scoring rules in the case of a categorical outcome. In the classification setting, the characterization (2.1) gives rise to the Savage representation introduced in Chapter 5.

2.1.2.1 CORP score decompositions

In practice, a sample $(F_1, y_1), \ldots, (F_n, y_n)$ of forecasts F_i and respective observations y_i $(i = 1, \ldots, n)$ is used to compute the average empirical score

$$\widehat{\mathrm{S}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(F_i, y_i)$$

via a proper scoring rule S, which can be used to compare competing forecasts. In the case of a binary variable, the CORP approach by Dimitriadis et al. (2021) described briefly in Section 2.1.1.3 yields a score decomposition into measures of miscalibration (MCB), discrimination (DSC) and uncertainty (UNC). If $\widehat{F}_1, \ldots, \widehat{F}_n$ denote the recalibrated forecast values obtained from the PAV algorithm, i.e., the isotonic regression fits of the observations y_i against the original forecasts F_i, the empirical score

$$\widehat{\mathrm{S}}_{\mathrm{rc}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(\widehat{F}_i, y_i)$$

of the recalibrated values $\widehat{F}_i$ is optimal among monotonically increasing fits. On the other hand, the unconditional distribution of Y is typically regarded as a reference forecast requiring hardly any skill as it can simply be estimated from the observations, which yields the empirical unconditional event probability $\widehat{F}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i$. The given forecasts are seen as having no skill if the empirical score

$$\widehat{\mathrm{S}}_{\mathrm{ref}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(\widehat{F}_0, y_i)$$

of the reference forecast outperforms the forecasts at hand, which yields a measure of inherent uncertainty $\mathsf{UNC} = \widehat{\mathrm{S}}_{\mathrm{ref}}$. The score difference $\mathsf{MCB} = \widehat{\mathrm{S}} - \widehat{\mathrm{S}}_{\mathrm{rc}}$ yields a measure of miscalibration, which estimates how much the forecasts might be improved through recalibration. The score difference $\mathsf{DSC} = \widehat{\mathrm{S}}_{\mathrm{ref}} - \widehat{\mathrm{S}}_{\mathrm{rc}}$ indicates how well the forecasts discriminate between the two outcomes 0 and 1, and hence serves as a measure of discrimination. With these terms, the original mean score can be written as

$$\widehat{\mathrm{S}} = \mathsf{MCB} - \mathsf{DSC} + \mathsf{UNC}.$$

The CORP reliability diagrams in Figure 2.3 are complemented by empirical score decompositions of the popular Brier score

$$\mathrm{S}_{\mathrm{Brier}}(F, y) = (F - y)^2,$$

which is a strictly proper scoring rule for probability forecasts of binary outcomes. Chapter 4 discusses score decompositions of scoring functions in the case of point forecasts linked to various statistical functionals.

2.2 Point Forecast Evaluation: Statistical Functionals and Consistent Scoring Functions

This section provides a short introduction to point forecast evaluation using consistent scoring functions based on Gneiting (2011a). For simplicity, the exposition is restricted to real-valued point forecasts and outcomes.

For various reasons, single-valued point forecasts may be preferred over probabilistic forecasts in practice. Point forecasts are linked to the distribution of Y by means of a statistical functional or as the Bayes rule, i.e., the minimizer of an expected loss, under a scoring or loss function. Naturally, a point forecast cannot be expected to match the outcome precisely. Communicating the true nature of a point forecast in terms of the targeted functional or loss function is key in addressing its limitations. A *statistical functional* is a map $\mathrm{T}\colon \mathcal{F} \to 2^{\mathbb{R}}$ disclosing a certain aspect of a probability distribution. The most prominent example is the mean functional given by $\mathrm{T}_{\mathbb{E}}(F) = \{\mathbb{E}Y \mid Y \sim F\}$ for sets $\mathcal{F}$ of probability measures $F \in \mathcal{F}$ with finite first moment. As the disclosed aspect may not be uniquely determined, set-valued functionals are needed. For example, the median or any other quantile of a discrete distribution may not be unique. In the case of a categorical outcome, the mode functional is typically the functional of choice, as discussed in Chapter 5.

The value of a point forecast is quantified by means of suitable scoring functions. A *scoring function* $\mathrm{S}\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ assigns a score $\mathrm{S}(x, y)$ to a point forecast x based on an observation y. A scoring function S is *consistent* for a functional T if the expected score under any distribution is minimized by values satisfying the functional relationship, i.e.,

$$\mathbb{E}[\mathrm{S}(t, Y)] \leq \mathbb{E}[\mathrm{S}(x, Y)] \quad \text{for } Y \sim F, t \in T(F), F \in \mathcal{F}, \text{ and } x \in \mathbb{R}.$$

A consistent scoring function is *strictly consistent* if the expected score is minimized only by values satisfying the functional relationship, i.e., if $\mathbb{E}[\mathrm{S}(t, Y)] = \mathbb{E}[\mathrm{S}(x, Y)]$ for $Y \sim F, F \in \mathcal{F}$ and $t \in \mathrm{T}(F)$ implies $x \in \mathrm{T}(F)$. A functional T is *elicitable* if a strictly consistent scoring function S for T exists. For example, the mean functional is elicited by the popular squared error scoring function

$$\mathrm{S}_{\mathrm{SE}}(x, y) = (x - y)^2$$

while the median is elicited by the absolute error scoring function

$$\mathrm{S}_{\mathrm{AE}}(x, y) = |x - y|.$$

Point forecasts and related concepts feature prominently in this thesis with Chapters 4 and 5 providing further insights into the cases of real-valued outcomes and categorical variables, respectively.

3 | A Simple Algorithm for Exact Multinomial Tests

This chapter is a slightly modified version of Resin (2022), where I propose a new method for computing acceptance regions of exact multinomial tests. From this method, an algorithm is derived, which finds exact p-values for tests of simple multinomial hypotheses. Using concepts from discrete convex analysis, the method is proven to be exact for various popular test statistics, including Pearson's chi-square and the log-likelihood ratio. The proposed algorithm improves greatly on the naive approach using full enumeration of the sample space. However, its use is limited to multinomial distributions with a small number of categories as the runtime grows exponentially with the number of possible outcomes. The method is applied in a simulation study. Applications of multinomial tests in forecast evaluation are outlined in Sections 3.4.2 and 4.4.2. Additionally, properties of a test statistic using probability ordering, referred to as the "exact multinomial test" by some authors, are investigated and discussed. The algorithm is implemented in the accompanying `R` package `ExactMultinom`.

3.1 Introduction

Multinomial goodness-of-fit tests feature prominently in the statistical literature and a wide range of applications. Tests relying on asymptotics have been available for a long time and have been rigorously studied all through the 20$^{\text{th}}$ century. The use of various test statistics has been investigated with Pearson's chi-square and the log-likelihood ratio statistic being vital examples. These statistics are members of the general family of power divergence statistics (Cressie and Read, 1984). With the widespread availability of computing power, Monte Carlo simulations and exact methods have also gained popularity.
Tate and Hyer (1973) and Kotze and Gokhale (1980) used the "exact multinomial test", which orders samples by probability, to assess the accuracy of asymptotic tests of a simple null hypothesis against an unspecified alternative. In the words of Cressie and Read (1989), this approach "has provided much confusion and contention in the literature". In accordance with Gibbons and Pratt (1975) and Radlow and Alf (1975), they conclude that the asymptotic fit of a test should be assessed using the appropriate exact test based on the test statistic in question. Nevertheless, the exact multinomial test is intuitively appealing, and, as Kotze and Gokhale (1980) put it, "[i]n the absence of [...] a specific alternative, it is reasonable to assume that outcomes with smaller probabilities under the null

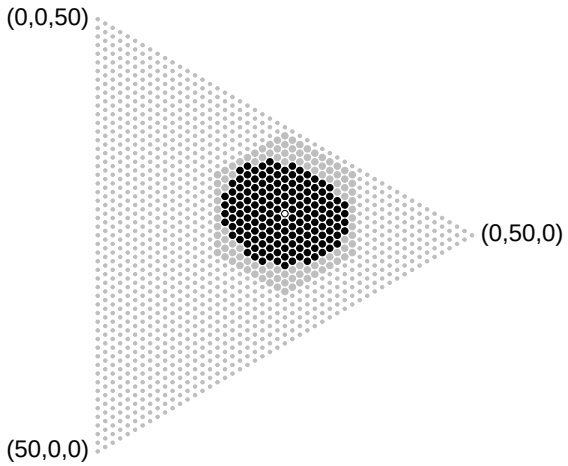


Figure 3.1 An acceptance region (black dots) at level $\alpha = 0.05$ for the null $\pi = (\frac{2}{10}, \frac{5}{10}, \frac{3}{10})$ and samples of size $n = 50$ with $m = 3$ categories. Only points within the ball (big dots) around the expectation (hollow dot) have to be considered to find this region.

hypothesis offer a stronger evidence for its rejection and should belong to the critical region". In Section 3.2, an asymptotic chi-square approximation to the exact multinomial test is derived, and an exemplary comparison of popular test statistics in terms of power is provided.

Regardless of the test statistic used, computing an exact p-value by fully enumerating the sample space is computationally challenging as the test statistic and the probability mass function have to be evaluated at every possible sample of which there are $\binom{n+m-1}{m-1} = \mathcal{O}(n^{m-1})$ for samples of size n with m categories. An improvement on this method has been proposed by Bejerano et al. (2004) for the family of power divergence statistics. Other approaches aimed at exact Pearson's chi-square and log-likelihood ratio tests exist (e.g., Baglivo et al., 1992; Hirji, 1997; Rahmann, 2003; Keich and Nagarajan, 2006). In this chapter, a new approach to exact multinomial tests is investigated.

The key observation underlying the proposed algorithm is that acceptance regions at arbitrary levels contain relatively few points, which are located in a neighborhood of the expected value under the null hypothesis as illustrated in Figure 3.1. An acceptance region can be found by iteratively evaluating points within a ball of increasing radius around the expected value (w.r.t. the Manhattan distance). The algorithm uses this approach to compute an exact p-value from the probability mass of the largest acceptance region that does not contain the observation. If p-values below an arbitrary threshold are not computed exactly, the runtime of the algorithm is guaranteed to be asymptotically faster than the approach using full enumeration as the diameter of any acceptance region essentially grows at a rate proportional to the square root of the sample size. The algorithm is detailed and proven to work for various popular test statistics in Section 3.3.

Furthermore, the algorithm is illustrated to work well in applications detailed

in Section 3.4. In particular, the algorithm's runtime is compared to the full enumeration method in a simulation study, and the resulting p-values are used to assess the fit of asymptotic chi-square approximations and investigate differences between several test statistics. As an application in forecast evaluation, the use of multinomial tests for uncertainty quantification within the so-called calibration simplex (Wilks, 2013) is outlined and justified.
The `R` programming language (`R` Core Team, 2022) has been used for all computations throughout this chapter. An implementation of the proposed method is provided within the `R` package `ExactMultinom` (Resin, 2020).

3.2 A Brief Review on Testing a Simple Multinomial Hypothesis

Consider a multinomial experiment $X = (X_1, \ldots, X_m)$ summarizing $n \in \mathbb{N}$ i.i.d. trials with $m \in \mathbb{N}$ possible outcomes. Let

$$\Delta_{m-1} := \{p \in [0,1]^m \mid p_1 + \ldots + p_m = 1\}$$

denote the *unit* $(m-1)$*-simplex* or *probability simplex* and

$$\Omega_{m,n} = \{x \in \mathbb{N}_0^m \mid x_1 + \ldots + x_m = n\}$$

the sample space, which is a *regular discrete* $(m-1)$*-simplex*. The distribution of X is characterized by a parameter $p = (p_1, \ldots, p_m) \in \Delta_{m-1}$ encoding the occurrence probabilities of the outcomes on any trial, or $X \sim \mathcal{M}_m(n, p)$ for short. The multinomial distribution $\mathcal{M}_m(n, p)$ is fully described by the probability mass function (pmf)

$$f_{n,p} \colon \Omega_{m,n} \to [0,1], x \mapsto n! \prod_{j=1}^{m} \frac{p_j^{x_j}}{x_j!}.$$

Suppose that the true parameter p is unknown. Consider the simple null hypothesis $p = \pi$ for some $\pi \in \Delta_{m-1}$. The agreement of a realization $x \in \Omega_{m,n}$ of X with the null hypothesis is typically quantified by means of a test statistic $T \colon \Omega_{m,n} \times \Delta_{m-1} \to \mathbb{R}$. Given such a test statistic T and presuming from now on that w.l.o.g. high values of $T(x, \pi)$ indicate 'extreme' observations under the null distribution $\mathbb{P}_\pi$, the p-value of x is defined as the probability

$$p_T(x, \pi) := \mathbb{P}_\pi(T(X, \pi) \geq T(x, \pi)) \tag{3.1}$$

of observing an observation that is at least as extreme under the null hypothesis. The *family of power divergence statistics* introduced by Cressie and Read (1984) offers a variety of test statistics for multinomial goodness-of-fit tests. It is defined as

$$T^\lambda(x, \pi) := \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} x_j \left(\left(\frac{x_j}{n\pi_j} \right)^\lambda - 1 \right) \text{ for } \lambda \in \mathbb{R} \setminus \{-1, 0\} \tag{3.2}$$

and as the pointwise limit in (3.2) for $\lambda \in \{-1, 0\}$. Notably, this family includes *Pearson's chi-square* statistic

$$T^{\chi^2}(x, \pi) := \sum_{j=1}^{m} \frac{(x_j - n\pi_j)^2}{n\pi_j} = \sum_{j=1}^{m} \frac{x_j^2}{n\pi_j} - n = T^1(x, \pi)$$

as well as the *log-likelihood ratio* (or G-test) statistic

$$T^G(x, \pi) := 2 \log \frac{f_{n,\frac{x}{n}}(x)}{f_{n,\pi}(x)} = 2 \sum_{j=1}^{m} x_j \log \frac{x_j}{n\pi_j} = T^0(x, \pi).$$

Under a null hypothesis with $\pi_i > 0$ for all $i = 1, \ldots, m$, every power divergence statistic is asymptotically chi-square distributed with $m - 1$ degrees of freedom. A natural test statistic arises if an 'extreme' observation is simply understood to mean an unlikely one, that is, if the pmf itself is used as a test statistic. In what follows, a strictly decreasing transformation of the pmf is used instead, which ensures that large values of the test statistic indicate extreme observations. Furthermore, this strictly decreasing transformation is chosen such that the resulting test statistic is asymptotically chi-square distributed. To this end, let Γ denote the Gamma function and

$$\bar{f}_{n,p} \colon \{x \in \mathbb{R}^m_{\geq 0} \mid x_1 + \ldots + x_m = n\} \to \mathbb{R}, x \mapsto \Gamma(n+1) \prod_{j=1}^{m} \frac{p_j^{x_j}}{\Gamma(x_j + 1)}$$

the continuous extension of the pmf $f_{n,p}$ to the convex hull of the discrete simplex $\Omega_{m,n}$. The *probability mass test statistic* is defined as

$$T^{\mathbb{P}}(x, \pi) := -2 \log \frac{f_{n,\pi}(x)}{\bar{f}_{n,\pi}(n\pi)}.$$

Obviously, the choice of strictly decreasing transformation does not affect the (exact) p-value given by (3.1) for $T = T^{\mathbb{P}}$. The following theorem gives rise to an asymptotic approximation of p-values derived from the probability mass test statistic, which has not been studied previously. In the simulation study of Section 3.4.1, the fit of this approximation is assessed empirically using exact p-values computed with the proposed method for samples of size $n = 100$ with $m = 5$ categories.

Theorem 3.1. *If $X \sim \mathcal{M}_m(n, \pi)$ follows a multinomial distribution with $n \in \mathbb{N}$ and $\pi \in \Delta_{m-1}$ such that $\pi_j > 0$ for $j = 1, \ldots, m$, then $T^{\mathbb{P}}(X, \pi)$ converges in distribution to a chi-square distribution χ^2_{m-1} with $m - 1$ degrees of freedom as $n \to \infty$.*

Proof. By Lemma 3.8 (in Appendix 3.A.1.1), the difference between the log-likelihood ratio and the probability mass statistic is

$$T^{\mathbb{P}}(X, \pi) - T^G(X, \pi) = \sum_{j=1}^{m} \left(\log \frac{X_j}{n\pi_j} + \mathcal{O}(1/X_j) - \mathcal{O}(1/n) \right).$$

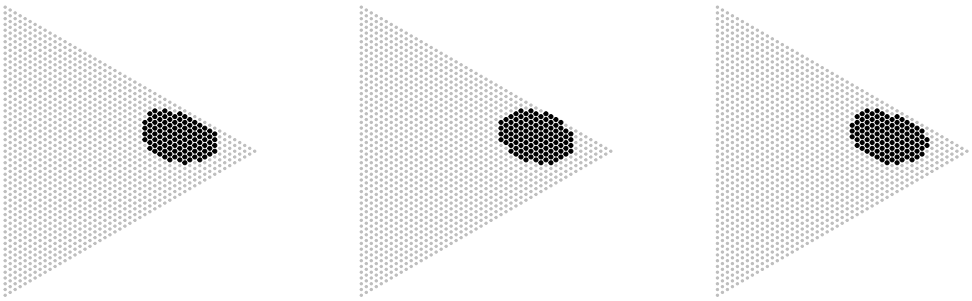


Figure 3.2 Acceptance regions (black) of probability mass (left), chi-square (center), and log-likelihood ratio (right) statistics at level $\alpha = 0.05$ for $n = 50$ and $\pi = (\frac{1}{10}, \frac{7}{10}, \frac{2}{10})$. The regions contain 108, 111, and 111 points, respectively (left to right). The tests are of size $0.0495, 0.0492$, and 0.0481, respectively.

Clearly, the bounded terms converge to zero in probability, and the $\log \frac{X_j}{n\pi_j}$ terms converge to zero in probability by the continuous mapping theorem. Hence, the probability mass statistic has the same asymptotic distribution as the log-likelihood ratio statistic. □

In what follows, the focus is on the chi-square, log-likelihood ratio, and probability mass statistics.

3.2.1 Acceptance regions

As outlined in Section 3.1, acceptance regions are of major importance to the idea pursued in this chapter. Given a test statistic T, the *acceptance region at level* $\alpha > 0$ is defined using p-values given by (3.1) as

$$A^T_{n,\pi}(\alpha) := \{x \in \Omega_{m,n} \mid p_T(x, \pi) > \alpha\}.$$

Equivalently,[1] the acceptance region can be written as the *sublevel set* of $T(\cdot, \pi)$ at the $(1-\alpha)$-quantile $t_{1-\alpha} = \min\{t \in \mathbb{R} \mid \mathbb{P}_\pi(T(X, \pi) \leq t) \geq 1 - \alpha\}$ of $T(X, \pi)$ under the null hypothesis $X \sim \mathcal{M}_m(n, \pi)$, i.e.,

$$A^T_{n,\pi}(\alpha) = \{x \in \Omega_{m,n} \mid T(x, \pi) \leq t_{1-\alpha}\}. \tag{3.3}$$

As illustrated in Figure 3.2, the probability mass test statistic typically yields acceptance regions that contain relatively few points because the regions contain the samples with the largest null probabilities. However, as samples with equal

[1] Since $T(X, \pi)$ follows a discrete distribution, the quantile $t_{1-\alpha}$ satisfies $\mathbb{P}_\pi(T(X, \pi) < t_{1-\alpha}) < 1 - \alpha$ by construction as the minimum value for which the cumulative non-exceedance probability is at least $1 - \alpha$. Hence, the defining inequalities are equivalent for $x \in \Omega_{m,n}$:

$$\begin{aligned} T(x, \pi) \leq t_{1-\alpha} \quad &\Leftrightarrow \quad \mathbb{P}_\pi(T(X, \pi) < T(x, \pi)) < 1 - \alpha \\ &\Leftrightarrow \quad \mathbb{P}_\pi(T(X, \pi) \geq T(x, \pi)) > \alpha \quad \Leftrightarrow \quad p_T(x, \pi) > \alpha. \end{aligned}$$

null probabilities are either all included or all excluded, smaller acceptance regions might be feasible at some levels α. If tests are randomized to ensure an equal level and size of the test, this property can be refined to yield an optimality property of the probability mass test's critical function.
In Section 3.3, it is shown that acceptance regions of the chi-square, log-likelihood ratio, and probability mass test statistic all grow at a rate $\mathcal{O}(n^{\frac{m-1}{2}})$ as their diameter grows at a rate $\mathcal{O}(\sqrt{n})$ if $\alpha > 0$ is fixed, see Proposition 3.7.

3.2.2 Power and bias

The *power function* of a test T of the null hypothesis $p = \pi$ at level α is

$$\Delta_{m-1} \to [0,1], p \mapsto 1 - \mathbb{P}_p(T(X) \in A^T_{n,\pi}(\alpha)),$$

which is the probability of rejecting the null hypothesis at level α if the true parameter is p. The *size* of a test is its power at $p = \pi$. A test T is said to be *unbiased* (for the null $p = \pi$ at level α) if its power is minimized at $p = \pi$.
In the case of the uniform null hypothesis $\pi = (\frac{1}{m}, \ldots, \frac{1}{m})$, Cohen and Sackrowitz (1975, Theorem 2.1) proved that the power function increases away from $p = \pi$ for test statistics of the form

$$T(x) = \sum_{j=1}^{m} h(x_j)$$

if h is a convex function. They concluded that tests based on the chi-square and the log-likelihood ratio test statistic are unbiased for the uniform null hypothesis. As a corollary to their theorem, it shall be noted that this result also applies to the probability mass test statistic.

Corollary 3.2 (to Cohen and Sackrowitz, 1975, Theorem 2.1)**.** *The probability mass test is unbiased for the uniform null hypothesis* $p = \pi = (\frac{1}{m}, \ldots, \frac{1}{m})$.

Proof. Since the probability mass statistic can be written as

$$T^{\mathbb{P}}(x, \pi) = 2 \sum_{j=1}^{m} \log \Gamma(x_j + 1) - x_j \log \pi_j - \log \frac{\Gamma(n\pi_j + 1)}{\pi_j^{n\pi_j}},$$

convexity of the summands as a function of x is an immediate consequence of the fact that the Gamma function is logarithmically convex on the positive real numbers, which is part of a characterization given by the Bohr-Mollerup theorem (Beals and Wong, 2010, Theorem 2.4.2). □

Many authors (e.g., West and Kempthorne, 1972; Cressie and Read, 1984; Wakimoto et al., 1987; Pérez and Pardo, 2003) have conducted small sample studies to investigate the power of chi-square, log-likelihood ratio and other tests. When conducting such studies, π, n, and α need to be chosen, all of which influence the resulting power function. Furthermore, it is frequently infeasible to assess the

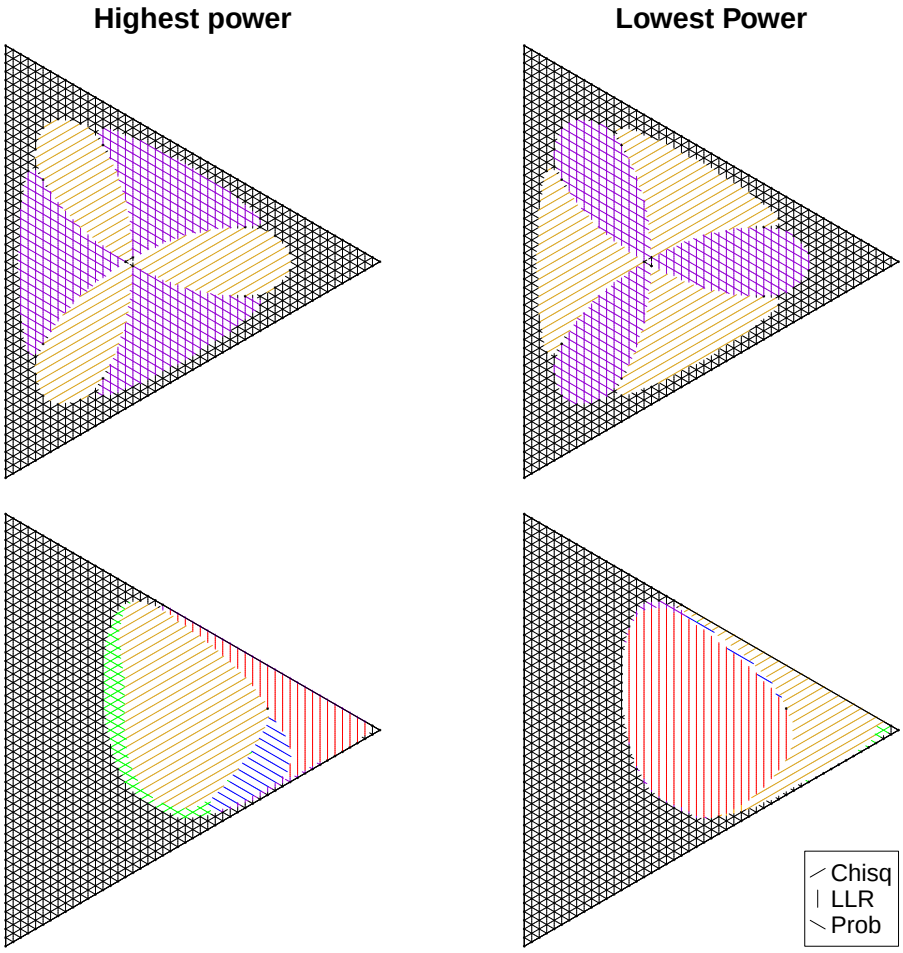


Figure 3.3 Ternary plots indicating which randomized tests of size $\alpha = 0.05$ yield the highest (left) and lowest (right) power for the uniform null hypotheses $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ (top) and $\pi = (\frac{1}{10}, \frac{7}{10}, \frac{2}{10})$ (bottom) for $n = 50$ among chi-square (Chisq), log-likelihood ratio (LLR), and probability mass (Prob) test. Overlapping lines indicate nearly equal powers (difference $< 10^{-5}$).

power function across all alternatives, and so alternatives of interest need to be picked. Therefore, most of these studies focused on the case of the uniform null hypothesis. In this case, the chi-square test has greater power for alternatives that assign a large proportion of the probability mass to relatively few categories, whereas the log-likelihood ratio test has greater power for alternatives that assign considerable probability mass to many categories (see Koehler and Larntz, 1980). In the ternary case, that is, if $m = 3$, comparisons on the full probability simplex are visually accessible. Figure 3.3 illustrates which of the three test statistics yields the highest and lowest power across the full ternary probability simplex. As the actual test size, which is frequently smaller than the level α, depends on the test statistic, the resulting power functions are difficult to compare directly. To account for this behavior, tests are randomized to ensure that test sizes match the specified level. For a test T and level α, let $s_{n,\pi}(T, \alpha) = 1 - \mathbb{P}_\pi(T(X) \in A^T_{n,\pi}(\alpha))$

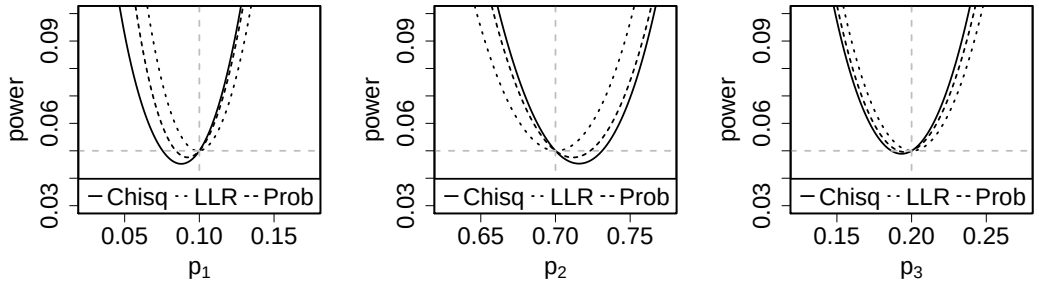


Figure 3.4 Power functions of randomized chi-square (Chisq), log-likelihood ratio (LLR), and probability mass (Prob) tests of size $\alpha = 0.05$ along alternatives given by $p(p_i, i), i = 1, 2, 3$ with null hypothesis $\pi = (\frac{1}{10}, \frac{7}{10}, \frac{2}{10})$ and sample size $n = 50$.

denote the actual size of the test. The *critical function*

$$\phi\colon \Omega_{m,n} \to [0,1], x \mapsto \begin{cases} 0, & \text{if } T(x,\pi) < t_{1-\alpha}, \\ \frac{\alpha - s_{n,\pi}(T,\alpha)}{\mathbb{P}_\pi(T(X) = t_{1-\alpha})}, & \text{if } T(x,\pi) = t_{1-\alpha}, \\ 1, & \text{if } T(x,\pi) > t_{1-\alpha}, \end{cases}$$

defines a randomized test[2] for the null hypothesis $p = \pi$ at level α, which rejects the null hypothesis with probability $\phi(x)$ if x is observed. The power function of a randomized test given by a critical function ϕ is

$$p \mapsto \sum_{x \in \Omega_{m,n}} \phi(x) \mathbb{P}_p(X = x) = 1 - \sum_{x \in A^T_{n,\pi}(\alpha)} (1 - \phi(x)) \mathbb{P}_p(X = x).$$

The power function shows that the probability mass test minimizes the acceptance region in the sense that it minimizes the sum

$$\sum_{x \in \Omega_{m,n}} (1 - \phi(x))$$

across all randomized tests for the null hypothesis $p = \pi$ at level α.
Figure 3.3 suggests that the probability mass test and the log-likelihood ratio test for the uniform null hypothesis at level $\alpha = 0.05$ are the same for $n = 50$. This behavior is a coincidence, and for other choices of α (e.g., $\alpha = 0.13$, for which coincidentally the probability mass statistic yields the same acceptance region as the chi-square statistic) the acceptance regions differ, and so do the power functions.
Figure 3.4 quantitatively compares power along alternatives of the form

$$p(q,i) = (\widetilde{q}\pi_1, \ldots, \widetilde{q}\pi_{i-1}, q, \widetilde{q}\pi_{i+1}, \ldots, \widetilde{q}\pi_m) \in \Delta_{m-1} \quad \text{with} \quad \widetilde{q} = \frac{1-q}{1-\pi_i}$$

[2]Randomized tests like this traditionally arise in the theory of uniformly most powerful tests, see for example Lehmann and Romano (2005, Chapter 3).

for $i = 1, \ldots, m$ and $q \in [0, 1]$, which yields parametrizations of the lines through π and the corners of the probability simplex. Arguably, the log-likelihood ratio test does not show any visible bias for $n = 50, \pi = (\frac{1}{10}, \frac{7}{10}, \frac{2}{10})$, and $\alpha = 0.05$, whereas the chi-square test shows the largest bias. The power function of the probability mass test lies in between the other power functions across most of the probability simplex, and so the probability mass test might serve as a good compromise in terms of power.

3.3 Exact p-Values via Acceptance Regions

Throughout this section, T is a test statistic, and $m, n \in \mathbb{N}$ and $\pi \in \Delta_{m-1}$ are fixed. To ease notation, the subscripts in the pmf of the null distribution are omitted, i.e., I write $f = f_{n,\pi}$, and the test statistic T is considered as a function on the sample space only, i.e., $T(\cdot) = T(\cdot, \pi)$. Let

$$d \colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}, (x, y) \mapsto \frac{1}{2}\|x - y\|_1 = \frac{1}{2}\sum_j |x_j - y_j|$$

be a rescaled version of the *Manhattan distance* and

$$B_r(y) = \{x \in \Omega_{m,n} \mid d(x, y) \leq r\}$$

the discrete ball with radius $r \in \mathbb{N}$ and center $y \in \Omega_{m,n}$. Furthermore, let $e_i = (\delta_{ij})_{j=1}^m$ denote the i-th vector of the standard basis of $\mathbb{R}^m$, where δ_{ij} is the Kronecker delta.

3.3.1 Finding acceptance regions using discrete convex analysis

As alluded to in Section 3.1, for many statistics an acceptance region $A = A_{n,\pi}^T(\alpha)$ for $\alpha \in (0, 1)$ can be found without enumerating the entire sample space $\Omega_{m,n}$ by considering only points in some ball around the expected value. The following theorem, which is proven at the end of this subsection, formalizes this approach for weakly quasi M-convex test statistics. A test statistic T is *weakly quasi M-convex* (Murota, 2003, Section 6.14, Property (QM_w)) if for all $x, y \in \Omega_{m,n}$ with $x \neq y$ there exist indices $i, j \in \{1, \ldots, m\}$ such that $x_i > y_i, x_j < y_j$ and

$$T(x - e_i + e_j) \leq T(x) \quad \text{or} \quad T(y + e_i - e_j) \leq T(y).$$

Theorem 3.3. *Let T be weakly quasi M-convex, and suppose $y \in \Omega_{m,n}$, $r \in \mathbb{N}$ and $\alpha \in (0, 1)$ are such that $\sum_{x \in B_r(y)} f(x) \geq 1 - \alpha$. Let $t \in \mathbb{R}$ be the smallest level such that the sublevel set $A = \{x \in B_r(y) \mid T(x) \leq t\}$ satisfies $\sum_{x \in A} f(x) \geq 1 - \alpha$. If $A \subseteq B_{r-1}(y)$, then A is the acceptance region $A_{n,\pi}^T(\alpha)$.*

Hence, an acceptance region can be found by iteratively enumerating a ball of increasing radius with arbitrary center until a sublevel set with enough probability

mass is found, and this sublevel set remains unchanged upon further increasing the ball, as illustrated in Figure 3.1 of Section 3.1 for an acceptance region of the probability mass statistic. The following proposition ensures that this approach can be applied to the chi-square, log-likelihood ratio, and probability mass test statistics.

Proposition 3.4.

(a) The probability mass test statistic $T^{\mathbb{P}}$ is weakly quasi M-convex.

(b) The power divergence test statistic T^{λ} is weakly quasi M-convex if $\lambda \geq 0$.

Proof. Throughout the proof, let $x, y \in \Omega_{m,n}$ be such that $x \neq y$, and define the index sets

$$S^+ := \{i \mid x_i > y_i\} \quad \text{and} \quad S^- := \{j \mid x_j < y_j\}.$$

(a) Let $T = T^{\mathbb{P}}$ and assume w.l.o.g. $T(x) \geq T(y)$. Then

$$\begin{aligned} T(y) - T(x) = -2 \log \frac{f(y)}{f(x)} &= -2 \log \left(\prod_{i \in S^+} \frac{x_i!}{y_i!} \pi_i^{y_i - x_i} \cdot \prod_{j \in S^-} \frac{x_j!}{y_j!} \pi_j^{y_j - x_j} \right) \\ &= -2 \log \left(\prod_{i \in S^+} \prod_{k=1}^{x_i - y_i} \frac{y_i + k}{\pi_i} \cdot \prod_{j \in S^-} \prod_{k=1}^{y_j - x_j} \frac{\pi_j}{x_j + k} \right) \leq 0. \end{aligned}$$

Both double products contain an equal number of multiplicands (since $\sum_j x_j = \sum_j y_j = n$) and are nonempty (since $x \neq y$). As the entire product is at least 1, there exist indices $i \in S^+$ and $j \in S^-$ and natural numbers $k^+ \leq x_i - y_i$ and $k^- \leq y_j - x_j$ such that the second inequality holds in

$$\frac{\pi_j}{x_j + 1} \geq \frac{\pi_j}{x_j + k^-} \geq \frac{\pi_i}{y_i + k^+} \geq \frac{\pi_i}{x_i}.$$

Therefore, the inequality

$$T(x - e_i + e_j) = T(x) - 2 \log \left(\frac{x_i}{\pi_i} \cdot \frac{\pi_j}{x_j + 1} \right) \leq T(x)$$

holds.

(b) See Appendix 3.A.1.2.

□

The rest of this section is devoted to the proof of Theorem 3.3, which utilizes the existence of certain sequences in the sublevel sets of weakly quasi M-convex functions given by the first part of the following lemma.

Lemma 3.5. *Let T be a weakly quasi M-convex function and $L = \{x \in \Omega_{m,n} \mid T(x) \leq t\}$ be the sublevel set of T at $t \in \mathbb{R}$.*

(a) *If $x, y \in L$ and $d = d(x, y)$, then there exists a sequence $x_0, x_1, \ldots, x_d \in L$ with $x_0 = x$, $x_d = y$ and $d(x_i, x_{i+1}) = 1$ for all $i = 0, 1, \ldots, d-1$.*

(b) *Suppose $y \in \Omega_{m,n}$ and $r \in \mathbb{N}$ are such that $A = \{x \in B_r(y) \mid T(x) \leq t\}$ is not empty. If $A \subseteq B_{r-1}(y)$, then $A = L$ is the sublevel set of T at t.*

Proof. (a) Proof by induction on d: Let $x, y \in L$ and $d = d(x, y)$. If $d = 0$, then $x = x_0 = y$ satisfies the condition. If $d > 0$, there exist i, j such that $x_i > y_i, x_j < y_j$ and $x_{d-1} = y + e_i - e_j \in L$ (or $x_{d-1} = x - e_i + e_j \in L$, in which case interchanging x and y and i and j yields the former formula for x_{d-1}) by weak quasi M-convexity of T. Then $d(x_{d-1}, y) = 1$ and

$$\begin{aligned} d(x, x_{d-1}) &= \frac{1}{2}\left(\sum_{k \neq i,j} |x_k - y_k| + \underbrace{|x_i - (y_i + 1)|}_{=|x_i - y_i| - 1} + \underbrace{|x_j - (y_j - 1)|}_{=|x_j - y_j| - 1} \right) \\ &= \frac{1}{2}(\|x - y\|_1 - 2) = d - 1. \end{aligned}$$

By induction hypothesis, there exists a sequence $x_0, x_1, \ldots, x_{d-1} \in L$, such that $x = x_0, x_1, \ldots, x_{d-1}, x_d = y \in L$ is the sought-after sequence.

(b) Assume there exists some $b \in L \setminus A$ and fix $a \in A$. By part (a), the sublevel set L contains a sequence $a = x_0, x_1, \ldots, x_d = b \in L$ with $d = d(a, b)$ and $d(x_i, x_{i+1}) = 1$ for $i = 0, 1, \ldots, d-1$. By the reverse triangle inequality $|d(x_{i+1}, y) - d(x_i, y)| \leq 1$, and, since $d(a, y) < r < d(b, y)$, there is an x_j such that $d(x_j, y) = r$, which yields $x_j \in A$, a contradiction (as $A \subseteq B_{r-1}(y)$). Therefore, $L \subseteq A$, and hence $A = L$. □

It can be shown that the existence of sequences as in part (a) of the previous lemma characterizes a weakly quasi M-convex set.[3] For further details on weak quasi M-convexity and discrete convex analysis in general, see Murota (2003). Finally, the theorem is readily proven as follows.

[3] The sublevel sets of weakly quasi M-convex functions are weakly quasi M-convex sets (Murota and Shioura, 2003, Theorem 3.10). A subset $M \subset \Omega_{m,n}$ is *weakly quasi M-convex* (Murota and Shioura, 2003, Property (Q-EXC$_w$)) if for all $x, y \in M$ with $x \neq y$ there exist indices $i, j \in \{1, \ldots, m\}$ such that $x_i > y_i, x_j < y_j$ and

$$x - e_i + e_j \in M \quad \text{or} \quad y + e_i - e_j \in M.$$

The proof of Lemma 3.5(a) is easily adapted to show that weakly quasi M-convex sets admit sequences as in part (a) of the lemma. To show the reverse implication, i.e., a set M is weakly quasi M-convex only if it admits sequences as in Lemma 3.5(a), let $x, y \in M, x \neq y$, $d = d(x, y)$, and $x_0, x_1, \ldots, x_d \in M$ be such a sequence. As $d(x, x_1) = 1$, there exist i, j such that $x_1 = x - e_i + e_j$. Furthermore, the inequalities $x_i > y_i$ and $x_j < y_j$ hold since

$$d - 1 = \sum_{l=1}^{d-1} d(x_l, x_{l+1}) \geq d(x_1, y) = \frac{1}{2}\left(\sum_{k \neq i,j} |x_k - y_k| + |x_i - 1 - y_i| + |x_j + 1 - y_j| \right)$$

yields a contradiction otherwise.

Algorithm 3.1. Compute exact p-value above some threshold.

Input: Observation $x \in \Omega_{m,n}$, hypothesis $\pi \in \Delta_{m-1}$, threshold $0 < \theta \ll 1$

Output: Exact p-value $p \in [\theta, 1]$ or 0 if the p-value is less than θ

compute $y \in \Omega_{m,n}$ minimizing $d(y, \mathbb{E}_\pi X)$
if $T(x) \leq T(y)$ **then** set $y = x$
initialize $r = 0$, $s = 0$
repeat
  for $z \in B_r(y) \setminus B_{r-1}(y)$ **do**
    if $T(z) < T(x)$ **then** set $s = s + f(z)$
  end
  increment $r = r + 1$
  set $t_{\min} = \min\{T(z) \mid d(y, z) = r\}$
until $(T(x) \leq t_{\min}$ and $T(y) < t_{\min})$ or $s > 1 - \theta$
if $s \leq 1 - \theta$ **then return** $1 - s$
else return 0

Proof of Theorem 3.3. Let $t \in \mathbb{R}$ be minimal such that $A = \{x \in B_r(y) \mid T(x) \leq t\}$ has probability mass $\sum_{x \in A} f(x) \geq 1 - \alpha$ and $A \subseteq B_{r-1}(y)$. Recall that the acceptance region $A^T_{n,\pi}(\alpha)$ is the sublevel set (3.3) at $t_{1-\alpha}$, and note that $t_{1-\alpha} \leq t$ holds as $\mathbb{P}_\pi(T(X) \leq t) \geq \sum_{x \in A} f(x) \geq 1 - \alpha$. By Lemma 3.5(b), A is the sublevel set at t, and hence $A \supseteq A^T_{n,\pi}(\alpha)$. Since t is minimal, it follows that $t = t_{1-\alpha}$ and $A = A^T_{n,\pi}(\alpha)$. □

3.3.2 Computing a p-value

As described in the previous subsection, an acceptance region can be determined by taking an arbitrary point and increasing the radius of a ball around this center point until the acceptance region is found using the criterion provided by Theorem 3.3. Obviously, the center of the ball should lie within the acceptance region, ideally at its center, to minimize the necessary iterations and the number of points for which to evaluate the pmf and the test statistic. The expected value $\mathbb{E}_\pi X = n \cdot \pi$ of the multinomial distribution, which is the center of mass of all probability weighted points in the discrete simplex, is close to the center of mass of the acceptance region as the region contains most of the mass. Therefore, a point close to the expected value is a suitable center for the ball.
The p-value of an observation x can be found by computing the total probability of the largest acceptance region not containing the observation, as formalized by Algorithm 3.1 and the following theorem.

Theorem 3.6. *Let T be weakly quasi M-convex and $r \in \mathbb{N}$. Suppose $x, y \in \Omega_{m,n}$ are such that $T(y) < T(x)$. If $A = \{z \in B_r(y) \mid T(z) < T(x)\}$ satisfies $A \subseteq B_{r-1}(y)$, then $p_T(x, \pi) = 1 - \sum_{z \in A} f(z)$.*

Proof. By Lemma 3.5(b), the set A is the sublevel set at $t = \max\{T(z) \mid z \in$

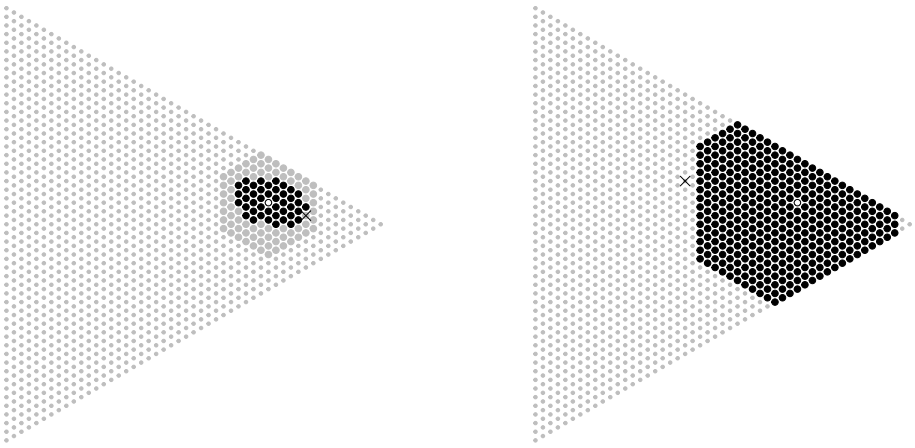


Figure 3.5 Points (big dots) in $\Omega_{3,50}$ for which the probability mass and test statistic are evaluated given the marked observations $x = (4, 40, 6)$ (left) and $x = (10, 20, 20)$ (right) under the null hypothesis $\pi = (\frac{1}{10}, \frac{7}{10}, \frac{2}{10})$ and $T = T^{\mathbb{P}}$. The p-values are 0.3049 (left) and less than $\theta = 0.0001$ (right). The black region on the left is the largest acceptance region not containing the observation x.

$\Omega_{m,n}, T(z) < T(x)\}$, and hence $p_T(x, \pi) = \mathbb{P}_\pi(T(X) \geq T(x)) = 1 - \mathbb{P}_\pi(T(X) \leq t) = 1 - \sum_{z \in A} f(z)$. □

The condition $T(y) < T(x)$ in Theorem 3.6 ensures that the sublevel set A is not empty, as otherwise the empty set may falsely be identified as the largest acceptance region not containing x. The case where no point y with $T(x) > T(y)$ is known requires special care. In this case, Algorithm 3.1 enumerates an acceptance region containing the observation itself to avoid premature termination.
To avoid enumerating unreasonably large balls, Algorithm 3.1 only determines exact p-values above a threshold θ and otherwise indicates that the p-value is smaller than the threshold θ by returning a value of 0. Figure 3.5 shows the points evaluated by Algorithm 3.1 for an observation with p-value greater, respectively, smaller than some threshold θ.

3.3.3 Implementation

Enumeration of the full sample space can be implemented using a simple recursion, as in the `R` packages `EMT` (Menzel, 2013) and `XNomial` (Engels, 2015). Whereas `EMT` is written purely in `R`, the function `xmulti` of the `XNomial` package uses an efficient `C++` subroutine for the recursion. To enumerate the samples at a given radius r in the repeat-loop of Algorithm 3.1, a similar, more complicated recursive scheme is implemented in the `R` package `ExactMultinom` using a `C++` subroutine to allow for fast recursions.
As an alternative, Bejerano et al. (2004) proposed a *branch and bound* approach to compute exact multinomial p-values, as implemented by Bejerano (2006). However, the branch and bound approach does not consider the probability mass

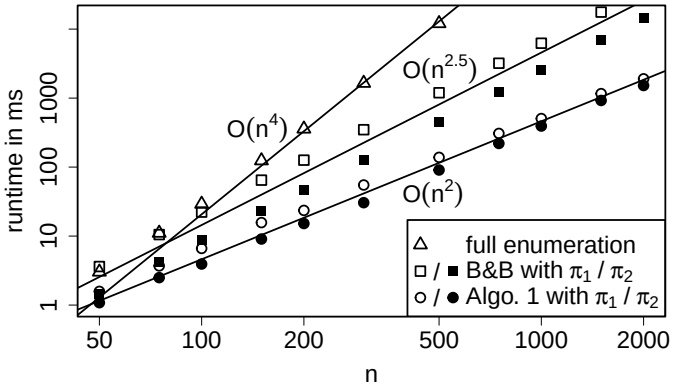


Figure 3.6 Mean runtime across 10 samples with p-values of about 0.001 under null hypotheses $\pi_1 = (0.2, 0.2, 0.2, 0.2, 0.2)$ and $\pi_2 = (0.01, 0.19, 0.2, 0.3, 0.3)$, respectively, using full enumeration, the branch and bound (B&B) approach and Algorithm 3.1.

statistic, and its implementation is limited to the log-likelihood ratio test. In contrast, the implementation of Algorithm 3.1 simultaneously computes p-values for the chi-square, log-likelihood ratio, and probability mass test statistics, as does `xmulti`. Further discussion of the branch and bound approach and other methods is deferred to Appendix 3.A.2 as none of these methods have been tailored to the probability mass test and other approaches do not produce "strictly exact" p-values (Keich and Nagarajan, 2006).

The current implementation of Algorithm 3.1 accurately finds p-values of order roughly as small as 10^{-10}. Smaller p-values often lead to negative output because of limited computational precision in the addition of many floating point numbers. To ensure accurate results, I recommend choosing θ no less than 10^{-8} with the current implementation.

During early runs of the simulation study described in Section 3.4, it was noticed that the runtime of Algorithm 3.1 tends to increase drastically if the null distribution contains a very small probability $\pi_i \ll n^{-1}$ for some $i \leq m$. In this case, the acceptance region is very flat, containing mostly points within a lower dimensional face of the discrete simplex, as hits in category i are improbable under the null. Hence, the asymptotic advantage of Algorithm 3.1 discussed in the next subsection requires a large sample size n to take effect under sparse null hypotheses. As a heuristic, which turned out to be an effective remedy, the implementation does not enumerate entire balls if $n \cdot \pi_i < \frac{1}{2}$ but only considers points $z \in \Omega_{m,n}$ with small z_i, by skipping all points z for which $\mathbb{P}_\pi(X_i \geq z_i) < \theta \cdot 10^{-8}$.

3.3.4 Runtime complexity

The discrete simplex $\Omega_{m,n}$ contains $|\Omega_{m,n}| = \binom{n+m-1}{m-1}$ points, and so the full enumeration takes $\mathcal{O}(n^{m-1})$ operations to compute a p-value. In comparison, the acceptance regions at a fixed level $\alpha > 0$ only contain $\mathcal{O}(n^{\frac{m-1}{2}})$ points, which

continues to hold for the smallest ball around the acceptance region centered at the expected value, as proven by Proposition 3.7 below. Therefore, Algorithm 3.1 only takes $\mathcal{O}(n^{\frac{m-1}{2}})$ operations to determine a p-value above the threshold θ. Figure 3.6 shows runtime as a function of n for $m = 5$. Whereas the runtime of the full enumeration method depends only on the parameters m and n, the runtime of the implementation of Algorithm 3.1 described in Section 3.3.3 depends on both the parameter π and the observation x. As with the branch and bound approach, the uniform null hypothesis results in a longer runtime than sparse null hypotheses, but the difference is less pronounced. Furthermore, the runtime of Algorithm 3.1 increases if the p-value of x is small, which is further investigated in the simulation study of Section 3.4.1. As the runtime increases exponentially with m, Algorithm 3.1 is only feasible if the number of categories m is small.

Proposition 3.7. *Let $T \in \{T^{\chi^2}, T^G, T^{\mathbb{P}}\}$, $\alpha \in (0,1)$ and $\pi \in \Delta_{m-1}$. Then there exists $c = c(\alpha, \pi)$ such that $A^T_{n,\pi}(\alpha) \subset B_{\sqrt{n}c}(n\pi)$ for sufficiently large n.*

Proof. Consider the canonical extension $\bar{T}$ of T to $\bar{\Omega}_{m,n} = \{x \in \mathbb{R}^m_{\geq 0} \mid x_1 + \ldots + x_m = n\}$ and let $\bar{B}_{n,r}(y) = \{x \in \bar{\Omega}_{m,n} \mid d(x,y) \leq r\}$ denote a ball in $\bar{\Omega}_{m,n}$ with boundary $\partial\bar{B}_{n,r}(y) = \{x \in \bar{\Omega}_{m,n} \mid d(x,y) = r\}$. Let $r_0 = \min_j \pi_j > 0$ and $n_0 \in \mathbb{N}$. If $n \geq n_0$, then every $x \in \partial\bar{B}_{n,\sqrt{nn_0}r_0}(n\pi)$ can be written as $x = x(n, x_0) := n\pi + \sqrt{nn_0}(x_0 - \pi)$ for some $x_0 \in \partial\bar{B}_{1,r_0}(\pi)$.
Let $t_{n,1-\alpha} = \min\{t \in \mathbb{R} \mid \mathbb{P}_\pi(T_n \leq t) \geq 1 - \alpha\}$ be the $(1-\alpha)$-quantile of $T_n = T(X_n)$, $X_n \sim \mathcal{M}_m(n, \pi)$ for $n \in \mathbb{N}$. As T_n converges to χ^2_{m-1} in distribution, the sequence $(t_{n,1-\alpha})$ of quantiles converges to the $(1-\alpha)$-quantile $\chi^2_{m-1,1-\alpha}$ (Van der Vaart, 1998, Lemma 21.2). Consequently, the maximum $t = \max_n t_{n,1-\alpha}$ exists, and the set $A_n = \{x \in \bar{\Omega}_{m,n} \mid \bar{T}(x) \leq t\}$ contains the acceptance region $A^T_{n,\pi}(\alpha)$ for every n.
As $\bar{T}$ is convex (by Lemma 3.9 in Appendix 3.A.1.3) and thus has convex sub-level sets, it suffices to show that n_0 can be chosen such that $\min\{\bar{T}(x) \mid x \in \partial\bar{B}_{n,\sqrt{nn_0}r_0}(n\pi)\}$ converges to a value greater than t to ensure that $A^T_{n,\pi}(\alpha) \subset A_n \subset \bar{B}_{n,\sqrt{n}(\sqrt{n_0}r_0)}(n\pi)$ for sufficiently large n.
In the case of the chi-square statistic, $T = T^{\chi^2}$, observe that

$$\bar{T}(x(n, x_0)) = \sum_j \frac{(x_j(n, x_0) - n\pi_j)^2}{n\pi_j} = \sum_j \frac{n_0(x_{0,j} - \pi_j)^2}{\pi_j}$$

does not depend on n, and so the canonical extension $\bar{T}$ of the chi-square statistic at radius $\sqrt{nn_0}r_0$ is bounded from below by $b(n_0) = \min\{\bar{T}(x) \mid x \in \partial\bar{B}_{n_0,r_0}(n_0\pi)\}$. This bound becomes arbitrarily large as n_0 is increased.
In case $T = T^G$ or $T = T^{\mathbb{P}}$, if n_0 is fixed, $\bar{T}(x(n, x_0))$ converges uniformly to $\bar{T}^{\chi^2}(x(n, x_0))$ for $x_0 \in \partial\bar{B}_{1,r_0}(\pi)$ (by Lemma 3.10 in Appendix 3.A.1.3). Hence, $\min\{\bar{T}(x) \mid x \in \partial\bar{B}_{n,\sqrt{nn_0}r_0}(n\pi)\}$ converges to $b(n_0)$. □

3.4 Application

In this section, the new method is applied in a simulation study. On the one hand, this study serves to show the improvements in runtime in comparison to some other methods. On the other hand, this study sheds some light on the fit of the asymptotic approximation to the probability mass test provided by Theorem 3.1 for a moderate sample size ($n = 100$). As a practical application in forecast evaluation, the usage of exact multinomial tests to increase the information conveyed by the *calibration simplex* (Wilks, 2013), a graphical tool used to assess ternary probability forecasts, is outlined.

3.4.1 Simulation study

For the simulation study, pairs $(\pi^{(1)}, x^{(1)}), \ldots, (\pi^{(N)}, x^{(N)})$ of null hypothesis parameters and samples were generated as i.i.d. realizations of the random quantity (P, X), where $X \mid P \sim \mathcal{M}_m(n, P)$ follows a multinomial distribution with random parameter $P \sim \mathcal{U}(\Delta_{m-1})$ drawn from a uniform distribution on the unit simplex. For each pair, p-values were computed using various test statistics and algorithms. Thereby, no specific null hypothesis had to be chosen and instead a wide variety was considered. By drawing samples from the null hypotheses, p-values follow a uniform distribution on $[0, 1]$. Various aspects of the tests and algorithms in question can be examined using the resulting rich data set and subsets thereof.

The following results were obtained using $N = 10^6$ such pairs with samples of size $n = 100$ drawn from multinomial distributions with $m = 5$ categories. Exact p-values were computed using the implementation of Algorithm 3.1 provided by the accompanying `R` package. To illustrate the speedup achieved by the new method in this study, the full enumeration method provided by the `xmulti` function of the `XNomial` package (Engels, 2015) and the branch and bound approach (Bejerano et al., 2004) were applied to the first 10^4 pairs. Essentially, the computational cost of the full enumeration is constant, independent of the null hypothesis at hand and the resulting p-value, whereas the cost of Algorithm 3.1 increases as the p-value decreases and also varies with the null hypothesis similar to the cost of the branch and bound approach.

The implementation of Algorithm 3.1 took an average of 0.59 ms to compute a p-value, improving on the branch and bound approach (1.78 ms), even though the latter only computes p-values for the log-likelihood ratio test, and full enumeration (29.76 ms). Perhaps surprisingly, Monte Carlo estimation (using `xmonte` from `XNomial`, which simulates 10000 samples by default) took almost twice as long (53.49 ms) as the full enumeration. Figure 3.7 illustrates the connection between runtime and size of the resulting p-values for the new method. As there are other factors influencing the runtime, and, since the implementation computes p-values for multiple statistics simultaneously, samples were ordered by their mean p-value $\bar{p}_T = \frac{1}{3}(p_{T^{\mathbb{P}}} + p_{T^{\chi^2}} + p_{T^G})$ and put in groups of 1000 samples with similar mean p-value (in particular, the groups contain samples with p-values in between

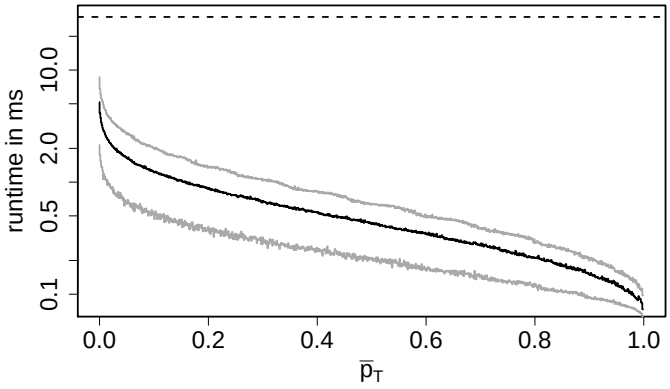


Figure 3.7 Runtime against mean p-value in groups of 1000 samples with similar mean p-value. The black line shows the mean runtime per group, whereas the gray lines show the 5th and 95th percentile. The dashed line shows the mean runtime using full enumeration.

the empirical $(\frac{a}{1000})$- and $(\frac{a+1}{1000})$-quantile for $a = 0, \ldots, 999$). The figure shows the mean runtime in each group as well as the 5th and 95th percentile.
To illustrate the fit of the classical chi-square approximation, the probability of a chi-square distribution with $m - 1$ degrees of freedom exceeding the values of the test statistics for each pair were computed. Figure 3.8 shows relative errors of the asymptotic approximations to the p-values for the three test statistics of interest. Given a test statistic T and asymptotic approximation $\widetilde{p}_T = \widetilde{p}_T(x, \pi)$ to the exact p-value $p_T = p_T(x, \pi)$, the relative error is the deviation from the exact value in parts of said value, $\frac{\widetilde{p}_T - p_T}{p_T}$. The asymptotic approximation to the chi-square statistic is quite accurate in most cases but tends to underestimate small p-values (< 0.1). The asymptotic approximation to the log-likelihood ratio statistic tends to slightly underestimate p-values on average. While the exact p-values are *valid* in that $\mathbb{P}_\pi(p_T(X, \pi) \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$, underestimation may result in invalid p-values. Asymptotic approximations of Pearson's chi-square and the log-likelihood ratio have been studied well, and the classical chi-square approximations can be improved through moment corrections (see Cressie and Read, 1989, and references therein). Furthermore, the errors typically increase if some category has a small expectation under the null hypothesis. The approximation to the probability mass p-values provided by Theorem 3.1 produces somewhat larger errors, especially for large p-values, and it clearly overestimates the p-values. This overestimation is emphasized by the fact that within the simulation data only a vanishingly small number of p-values was (slightly) underestimated, all of which were larger than 0.9. Figure 3.9 illustrates how estimation errors influence the distribution of the resulting p-values. Whereas the exact p-values appear to follow a uniform distribution, the asymptotic p-values clearly deviate from uniformity. For the probability mass statistic, the asymptotic test yields a conservative test, whereas the asymptotic log-likelihood ratio test (and also the asymptotic chi-square test at small significance levels) is slightly

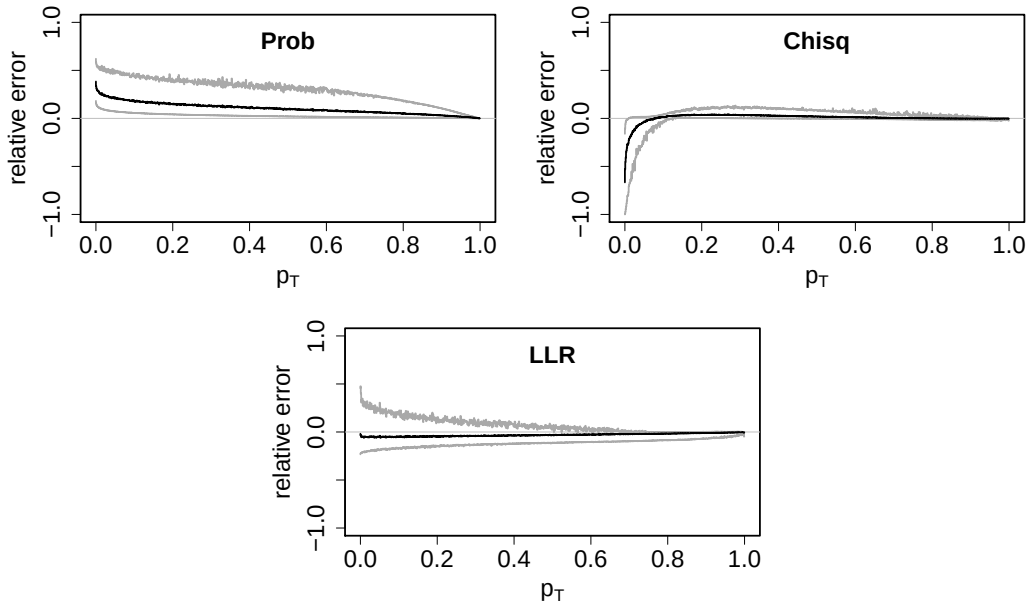


Figure 3.8 Relative errors of asymptotic approximations to p-values for probability mass (Prob), chi-square (Chisq) and log-likelihood ratio (LLR) test statistic. The plots were obtained using the same grouping scheme as in Figure 3.7.

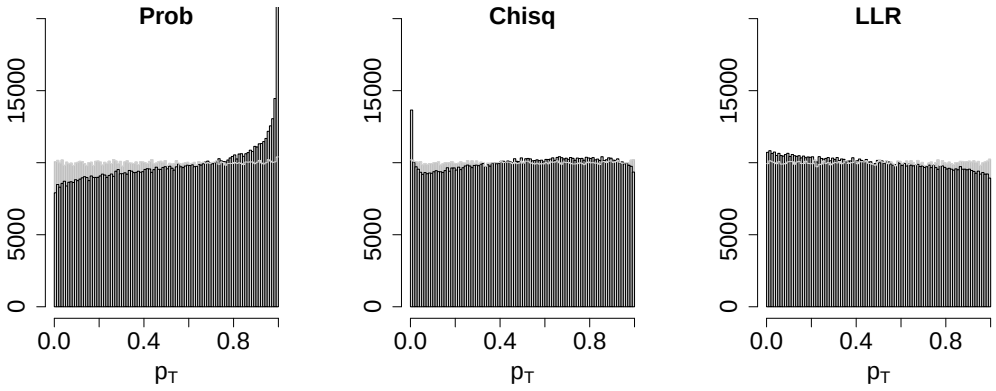


Figure 3.9 Histograms of asymptotic approximations to p-values for probability mass (Prob), chi-square (Chisq) and log-likelihood ratio (LLR) test statistic in black. The gray histograms show respective exact p-values. The rightmost bar within the left histogram is not fully shown and extends further up to over 30000 counts.

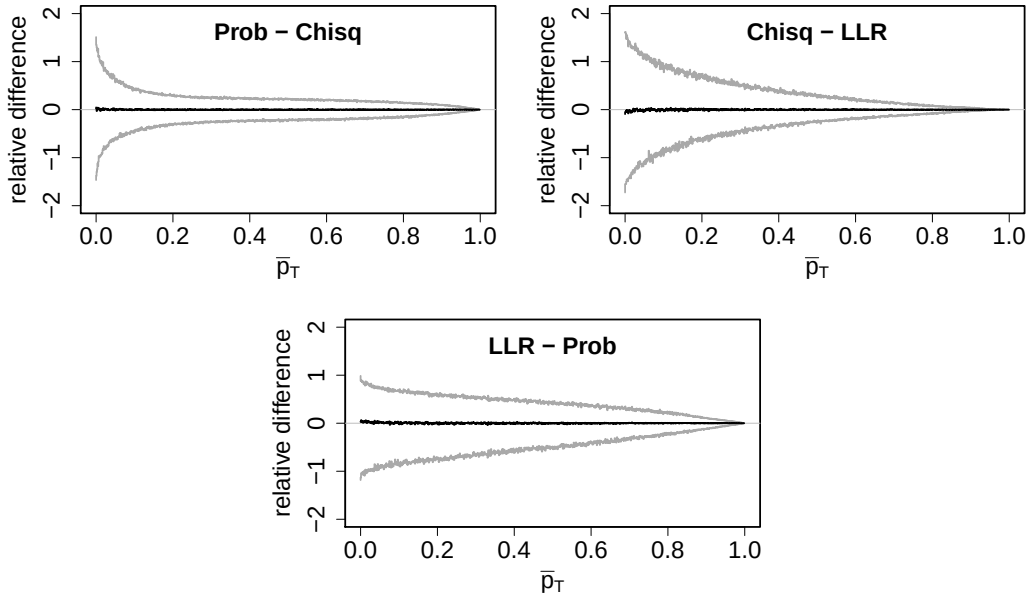


Figure 3.10 Relative differences between exact p-values of probability mass (Prob), chi-square (Chisq) and log-likelihood ratio (LLR) test statistic against the mean of compared p-values. The plots were obtained using the same grouping scheme as in Figure 3.7.

Table 3.1 Exact p-values p_T and asymptotic p-values $\widetilde{p}_T$ of five randomly selected pairs (x, π) with $0.01 < p_{T^G}(x, \pi) < 0.1$.

π	$p_{T^{\mathbb{P}}}$	$\widetilde{p}_{T^{\mathbb{P}}}$	$p_{T^{X^2}}$	$\widetilde{p}_{T^{X^2}}$	p_{T^G}	$\widetilde{p}_{T^G}$
$(0.116, 0.225, 0.259, 0.002, 0.398)$	0.0068	0.0092	0.0190	0.0073	0.0126	0.0172
$(0.038, 0.079, 0.224, 0.387, 0.272)$	0.1150	0.1268	0.1437	0.1469	0.0361	0.0307
$(0.595, 0.129, 0.093, 0.064, 0.118)$	0.0447	0.0495	0.0477	0.0482	0.0719	0.0665
$(0.497, 0.217, 0.223, 0.057, 0.007)$	0.0761	0.0994	0.0803	0.0741	0.0461	0.0498
$(0.243, 0.022, 0.237, 0.373, 0.125)$	0.0474	0.0566	0.0508	0.0507	0.0628	0.0568

anti-conservative.

Figure 3.10 shows relative differences between exact p-values obtained with the three test statistics. Given test statistics T and T', the relative difference between p-values $p_T = p_T(x, \pi)$ and $p_{T'} = p_T(x, \pi)$ is $\frac{p_T - p_{T'}}{\overline{p}_T}$, where $\overline{p}_T = \frac{p_T + p_{T'}}{2}$. It can be seen that the choice of test statistic can make quite a difference. A closer look at the simulation data revealed that these differences tend to be smaller if expectations for all categories are large under the null. To provide some numerical insights, Table 3.1 lists exact and asymptotic p-values.

3.4.2 The calibration simplex

Turning to an application in forecast verification, consider a random variable Y and a probabilistic forecast F for Y. For an introduction to probabilistic forecasting in general, see Gneiting and Katzfuss (2014). A probabilistic forecast

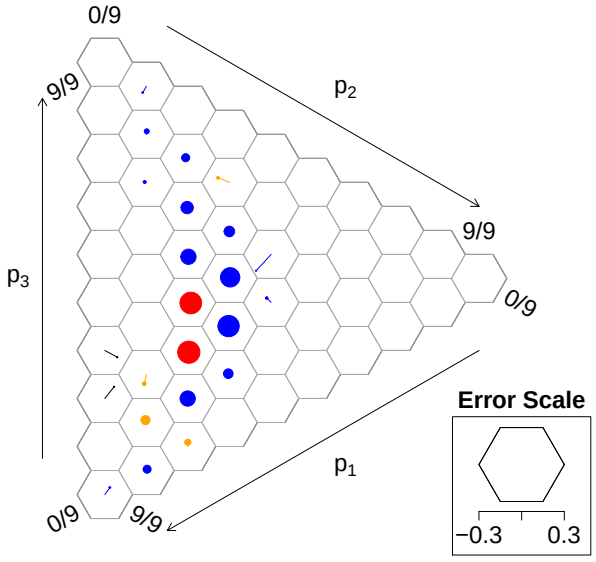


Figure 3.11 Calibration simplex with color-coded p-values from the log-likelihood ratio statistic evaluating a total of 21,240 club soccer predictions by FiveThirtyEight (`https://projects.fivethirtyeight.com/soccer-predictions/`) for matches from September 2016 until April 2019. Outcomes are encoded as 1 = "home win", 2 = "draw" and 3 = "away win". Only groups containing at least ten forecasts are shown. Blue indicates a p-value $p_{TG} > 0.1$, orange $0.1 > p_{TG} \geq 0.01$, red $p_{TG} < 0.01$ and black $p_{TG} = 0$.

is said to be *calibrated* if the conditional distribution of the quantity of interest given a forecast coincides with the forecast distribution, that is,

$$Y \mid F \sim F \tag{3.4}$$

holds almost surely. Suppose now that Y maps to one of three different outcomes only. Then, a probabilistic forecast is fully described by the probabilities it assigns to each outcome. In this case, the calibration simplex (Wilks, 2013) can be used to graphically identify discrepancies between predicted probabilities and conditional outcome frequencies. Given i.i.d. realizations $(f_1, y_1), \ldots, (f_N, y_N)$ consisting of forecast probabilities (vectors within the unit 2-simplex) and observed outcomes encoded 1, 2, and 3, forecast-outcome pairs with similar forecast probabilities are grouped according to a tessellation of the probability simplex. Thereafter, calibration is assessed by comparing average forecast and actual outcome frequencies within each group.
Figure 3.11 shows a calibration simplex, a graphical tool used to conduct this comparison visually. The groups are determined by overlaying the probability simplex with a hexagonal grid. The circular dots correspond to nonempty groups of forecasts given by a hexagon. The dots' areas are proportional to the number of forecasts per group. A dot is shifted away from the center of the respective hexagon by a scaled version of the difference in average forecast probabilities and outcome frequencies. This graphical display provides valuable insight into the

forecast distribution and the conditional distribution of the quantity of interest. However, it is not apparent how big the differences may be merely by chance.
If the forecast is calibrated, then, by (3.4), the outcome frequencies $\bar{y}$ within a group of size n with mean forecast $\bar{f}$ follow a generalized multinomial distribution (the multinomial analog of the Poisson binomial distribution), that is, a convolution of multinomial distributions $\mathcal{M}(1, f_i)$ with parameters $f_1, \ldots, f_n \in \Delta_{m-1}$. If these parameters only deviate little from their mean $\bar{f} = \frac{1}{n}\sum_i f_i$, then, presumably, the generalized multinomial distribution should not deviate much from a multinomial distribution with parameter $\bar{f}$. Under this presumption, multinomial tests can be applied to quantify the discrepancy within each group through a p-value. As the number of outcomes $m = 3$ is small, exact p-values are efficiently computed by Algorithm 3.1 even for large sample sizes n.
In Figure 3.11, p-values obtained from the log-likelihood ratio statistic are conveyed through a coloring scheme. Note that a p-value is exactly zero only if an outcome is forecast to have zero probability and said outcome still realizes.[4] Figure 3.11 was generated using the `R` package `CalSim` (Resin, 2021a).
The calibration simplex can be seen as a generalization of the popular reliability diagram. In light of this analogy, the use of multinomial tests to assess the statistical significance of differences in predicted probabilities and observed outcome frequencies serves the same purpose as consistency bars in reliability diagrams introduced by Bröcker and Smith (2007). Consistency bars are constructed using Monte Carlo simulation. To justify the above presumption, the multinomial p-values used to construct Figure 3.11 were compared to p-values computed from 10000 Monte Carlo samples obtained from the generalized multinomial distributions. To this end, the standard deviation of the Monte Carlo p-values was estimated using the estimated p-value in place of the true generalized multinomial p-value. Most of the multinomial p-values were quite close to the Monte Carlo estimates with an absolute difference less than two standard deviations, whereas two of them deviated on the order of 6 to 8 standard deviations from the Monte Carlo estimates, which nonetheless resulted in a relatively small absolute error. In particular, using the Monte Carlo estimated p-values did not change Figure 3.11. As the computation of Monte Carlo estimates from the generalized multinomial distributions is computationally expensive, the multinomial p-values serve as a fast and adequate alternative. Further improving uncertainty quantification within the calibration simplex is a subject for future work.

3.5 Concluding Remarks

In this chapter, a new method for computing exact p-values was investigated. It has been illustrated that the new method works well when the number m of categories is small, which results in a concrete speedup in practical applications,

[4] The p-values of exactly zero observed in Figure 3.11 appear to be due to mislabeled data as some matches that resulted in a penalty shoot-out are encoded as draws in the data by FiveThirtyEight.

as illustrated through a simulation study. As a further application not discussed in this chapter, the new method appears to be well suited to determine level set confidence regions discussed by Chafai and Concordet (2009) and Malloy et al. (2021). When m is too large for exact methods to be feasible, other methods may be used to approximate exact p-values as hinted at in Appendix 3.A.2. Such an approach may be added to the `ExactMultinom` package in a future version.
Regarding the choice of test statistic, the "exact multinomial test" was treated as a test statistic, and the asymptotic distribution of the resulting probability mass statistic was derived. Like most prominent test statistics, the probability mass statistic yields unbiased tests for the uniform null hypothesis. It was shown that a randomized test based on the probability mass statistic can be characterized in that it minimizes the respective (weighted) acceptance region.
Although asymptotic approximations work well in many use cases, there are cases, where these approximations are not adequate, for example, when dealing with small sample sizes or small expectations. On the other hand, there is nothing to be said against the use of exact tests whenever feasible, and it is recommended in the applied literature (McDonald, 2009, p. 83) for samples of moderate size up to 1000. As the available implementations of exact multinomial tests in `R` use full enumeration, the new implementation increases the scope of exact multinomial tests for practitioners.

3.A Appendix

3.A.1 Mathematical details

3.A.1.1 Difference Between Log-Likelihood Ratio and Probability Mass Statistic

The following lemma completes the proof of Theorem 3.1.

Lemma 3.8. *Let $\pi \in \Delta_{m-1}$ with $\pi_j > 0$ for all $j = 1, \dots, m$ and $x \in \Omega_{m,n}$. Then*

$$T^{\mathbb{P}}(x, \pi) - T^{G}(x, \pi) = \sum_{j=1}^{m} \left(\log(x_j) + 2r(x_j) - \log(n\pi_j) - 2r(n\pi_j)\right)$$

holds for a function r on the positive real numbers for which $0 < r(x) < \frac{1}{12x}$ for $x > 0$. In case $x_j = 0$ for some $j = 1, \dots, m$, the above equality holds if $\log(0) + 2r(0)$ is understood to be 0.

Proof. The logarithm of the Gamma function can be written as

$$\log \Gamma(x+1) = \log x\Gamma(x) = x \log(x) - x + \frac{1}{2}\log(2\widetilde{\pi}x) + r(x)$$

for a function r on the positive real numbers for which $0 < r(x) < \frac{1}{12x}$ holds for all $x > 0$ (see Abramowitz and Stegun, 1972, 6.1.41 and 6.1.42; here $\widetilde{\pi}$ denotes Archimedes' constant). This formula yields

$$\begin{aligned}
\log \bar{f}_{n,\frac{y}{n}}(y) &= \log \Gamma(n+1) + \sum_j \left(y_j \log \frac{y_j}{n} - \log \Gamma(y_j + 1)\right) \\
&= \log \Gamma(n+1) + \sum_j \left(y_j \log \frac{y_j}{n} - y_j \log(y_j) + y_j - \frac{1}{2}\log(2\widetilde{\pi}y_j) - r(y_j)\right) \\
&= \log \Gamma(n+1) + n(1 - \log n) - \sum_j \left(\frac{1}{2}\log(2\widetilde{\pi}y_j) + r(y_j)\right)
\end{aligned}$$

for $y \in \mathbb{R}^m_{>0}$ such that $\sum_j y_j = n$, and hence

$$\begin{aligned}
T^{\mathbb{P}}(x, \pi) - T^{G}(x, \pi) &= 2(\log \bar{f}_{n,\pi}(n\pi) - \log f_{n,\frac{x}{n}}(x)) \\
&= 2\sum_j \left(\frac{1}{2}\log \frac{x_j}{n\pi_j} + r(x_j) - r(n\pi_j)\right).
\end{aligned}$$

□

3.A.1.2 Proof of Proposition 3.4(b)

Proof of Proposition 3.4(b). Throughout the proof, let $x, y \in \Omega_{m,n}$ be such that $x \neq y$, and define the index sets

$$S^+ := \{i \mid x_i > y_i\} \quad \text{and} \quad S^- := \{j \mid x_j < y_j\}.$$

Let $T = T^\lambda$ and assume w.l.o.g. $T(x) \geq T(y)$. First, consider the case $\lambda > 0$. Note that

$$T(x) - T(y) = \frac{2}{\lambda(\lambda+1)} \left(\sum_{i \in S^+} \frac{x_i^{\lambda+1} - y_i^{\lambda+1}}{(n\pi_i)^\lambda} - \sum_{j \in S^-} \frac{y_j^{\lambda+1} - x_j^{\lambda+1}}{(n\pi_j)^\lambda} \right) \geq 0 \tag{3.5}$$

and

$$\begin{aligned} T(x - e_{i^*} + e_{j^*}) = T(x) &- \frac{2}{\lambda(\lambda+1)} \left(\frac{x_{i^*}^{\lambda+1} - (x_{i^*} - 1)^{\lambda+1}}{(n\pi_{i^*})^\lambda} \right) \\ &+ \frac{2}{\lambda(\lambda+1)} \left(\frac{(x_{j^*} + 1)^{\lambda+1} - x_{j^*}^{\lambda+1}}{(n\pi_{j^*})^\lambda} \right) \end{aligned} \tag{3.6}$$

for $i^* \in S^+, j^* \in S^-$. If

$$i^* = \arg\max_{i \in S^+} \frac{x_i^{\lambda+1} - (x_i - 1)^{\lambda+1}}{(n\pi_i)^\lambda}, \quad j^* = \arg\min_{j \in S^-} \frac{(x_j + 1)^{\lambda+1} - x_j^{\lambda+1}}{(n\pi_j)^\lambda}$$

and $d = d(x, y)$, then

$$\begin{aligned} \frac{x_{i^*}^{\lambda+1} - (x_{i^*} - 1)^{\lambda+1}}{(n\pi_{i^*})^\lambda} &= \frac{1}{d} \sum_{i \in S^+} \sum_{k=1}^{x_i - y_i} \frac{x_{i^*}^{\lambda+1} - (x_{i^*} - 1)^{\lambda+1}}{(n\pi_{i^*})^\lambda} \\ &\geq \frac{1}{d} \sum_{i \in S^+} \sum_{k=1}^{x_i - y_i} \frac{x_i^{\lambda+1} - (x_i - 1)^{\lambda+1}}{(n\pi_i)^\lambda} \\ &\geq \frac{1}{d} \sum_{i \in S^+} \sum_{k=1}^{x_i - y_i} \frac{(x_i + 1 - k)^{\lambda+1} - (x_i - k)^{\lambda+1}}{(n\pi_i)^\lambda} \\ &= \frac{1}{d} \sum_{i \in S^+} \frac{x_i^{\lambda+1} - y_i^{\lambda+1}}{(n\pi_i)^\lambda} \\ &\geq \frac{1}{d} \sum_{j \in S^-} \frac{y_j^{\lambda+1} - x_j^{\lambda+1}}{(n\pi_j)^\lambda} \\ &= \frac{1}{d} \sum_{j \in S^-} \sum_{k=1}^{y_j - x_j} \frac{(x_j + k)^{\lambda+1} - (x_j - 1 + k)^{\lambda+1}}{(n\pi_j)^\lambda} \\ &\geq \frac{1}{d} \sum_{j \in S^-} \sum_{k=1}^{y_j - x_j} \frac{(x_j + 1)^{\lambda+1} - x_j^{\lambda+1}}{(n\pi_j)^\lambda} \\ &\geq \frac{1}{d} \sum_{j \in S^-} \sum_{k=1}^{y_j - x_j} \frac{(x_{j^*} + 1)^{\lambda+1} - x_{j^*}^{\lambda+1}}{(n\pi_{j^*})^\lambda} \\ &= \frac{(x_{j^*} + 1)^{\lambda+1} - x_{j^*}^{\lambda+1}}{(n\pi_{j^*})^\lambda}, \end{aligned} \tag{3.7}$$

where the third inequality is due to inequality (3.5). Hence, $T(x) \geq T(x - e_{i^*} + e_{j^*})$ by equation (3.6).
For $\lambda = 0$, simply taking the limit (as $\lambda \to 0$) in the above equations with

$$i^* = \arg\max_{i \in S^+} \quad 2x_i \log\left(\frac{x_i}{n\pi_i}\right) - 2(x_i - 1)\log\left(\frac{x_i - 1}{n\pi_i}\right),$$
$$j^* = \arg\min_{j \in S^-} \quad 2(x_j + 1)\log\left(\frac{x_j + 1}{n\pi_j}\right) - 2x_j \log\left(\frac{x_j}{n\pi_j}\right)$$

yields the desired inequality since

$$\begin{aligned}
&2x_{i^*} \log\left(\frac{x_{i^*}}{n\pi_{i^*}}\right) - 2(x_{i^*} - 1)\log\left(\frac{x_{i^*} - 1}{n\pi_{i^*}}\right) \\
&= \lim_{\lambda \to 0} \frac{2}{\lambda(\lambda+1)} x_{i^*} \left(\left(\frac{x_{i^*}}{n\pi_{i^*}}\right)^{\lambda} - 1\right) - \lim_{\lambda \to 0} \frac{2}{\lambda(\lambda+1)} (x_{i^*} - 1)\left(\left(\frac{x_{i^*} - 1}{n\pi_{i^*}}\right)^{\lambda} - 1\right) \\
&= \lim_{\lambda \to 0} \frac{2}{\lambda(\lambda+1)} \left(\frac{x_{i^*}^{\lambda+1} - (x_{i^*} - 1)^{\lambda+1}}{(n\pi_{i^*})^{\lambda}} - 1\right) \\
&\geq \lim_{\lambda \to 0} \frac{2}{\lambda(\lambda+1)} \left(\frac{(x_{j^*} + 1)^{\lambda+1} - x_{j^*}^{\lambda+1}}{(n\pi_{j^*})^{\lambda}} - 1\right) \\
&= 2(x_{j^*} + 1)\log\left(\frac{x_{j^*} + 1}{n\pi_{j^*}}\right) - 2x_{j^*} \log\left(\frac{x_{j^*}}{n\pi_{j^*}}\right),
\end{aligned}$$

where the inequality is due to inequality (3.7). □

3.A.1.3 Details for the Proof of Proposition 3.7

The following two lemmas provide further details not contained in the proof of Proposition 3.7 itself.

Lemma 3.9. *Using notation as in the proof of Proposition 3.7, $x \mapsto \bar{T}(x)$ is convex.*

Proof. The function $x \mapsto \bar{T}^{\chi^2}(x) = \sum_j \frac{x_j^2}{n\pi_j} - n$ is clearly convex as it is a sum of convex functions. The function $x \mapsto \bar{T}^G(x) = 2\sum_j x_j \log(x_j) - x_j \log(n\pi_j)$ is convex since it is easy to show that $x \mapsto x\log(x)$ is convex (using either the inequality of the arithmetic and geometric means or the second derivative). The function $x \mapsto \bar{T}^{\mathbb{P}}(x) = 2(\log(\bar{f}_{n,\pi}(n\pi)) - \log(\Gamma(n+1)) + \sum_j \log(\Gamma(x_j + 1)) - \sum_j x_j \log(p_j))$ is convex as the Gamma function is logarithmically convex by the Bohr-Mollerup theorem (Beals and Wong, 2010, Theorem 2.4.2). □

Lemma 3.10. *Using notation as in the proof of Proposition 3.7, the function $\partial\bar{B}_{1,r_0}(\pi) \to \mathbb{R}, x_0 \mapsto \bar{T}(x(n, x_0))$ converges uniformly to $\bar{T}^{\chi^2}(x(n, x_0))$ as $n \to \infty$ if $T = T^G$ or $T = T^{\mathbb{P}}$.*

Proof. Let $x_0 \in \partial \bar{B}_{1,r_0}(\pi)$, and define $c = c(x_0) := \sqrt{n_0}(x_0 - \pi)$. Hence $|c_j| \leq \sqrt{n_0} r_0 < \sqrt{n_0}$ for all $j = 1, \ldots, m$.
In the case $T = T^G$, the Taylor expansion $\log(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}$ yields

$$\begin{aligned}
\bar{T}(x(n,x_0)) &= 2\sum_{j=1}^{m} x(n,x_0)_j \log \frac{x(n,x_0)_j}{n\pi_j} \\
&= 2\sum_{j} (n\pi_j + \sqrt{n}c_j) \log \frac{n\pi_j + \sqrt{n}c_j}{n\pi_j} \\
&= 2\sum_{j} (n\pi_j + \sqrt{n}c_j) \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{c_j}{\sqrt{n}\pi_j} \right)^k \\
&= 2\sum_{j} \left(\sqrt{n}c_j + \frac{c_j^2}{2\pi_j} - \frac{c_j^3}{2\sqrt{n}\pi_j^2} + \frac{n\pi_j + \sqrt{n}c_j}{\sqrt{n}^3} \sum_{k=3}^{\infty} \frac{(-1)^{k+1} c_j^k}{k\sqrt{n}^{k-3}\pi_j^k} \right).
\end{aligned}$$

As $\sum_j c_j = 0$ and $2\sum_j \frac{c_j^2}{2\pi_j} = T^{\chi^2}(x(n,x_0))$, the inequalities

$$\begin{aligned}
&|\bar{T}^{\chi^2}(x(n,x_0)) - \bar{T}(x(n,x_0))| \\
&< \sum_{j} \left(\frac{|c_j|^3}{2\sqrt{n}\pi_j^2} + \frac{n\pi_j + \sqrt{n}|c_j|}{\sqrt{n}^3} \sum_{k=3}^{\infty} \frac{|c_j|^k}{k\sqrt{n}^{k-3}\pi_j^k} \right) \\
&< \sum_{j} \left(\frac{\sqrt{n_0}^3}{2\sqrt{n}\pi_j^2} + \frac{n\pi_j + \sqrt{n}\sqrt{n_0}}{\sqrt{n}^3} \sum_{k=3}^{\infty} \frac{\sqrt{n_0}^k}{k\sqrt{n}^{k-3}\pi_j^k} \right) \\
&< \frac{1}{\sqrt{n}} \sum_{j} \left(\frac{\sqrt{n_0}^3}{2\pi_j^2} + (\pi_j + \sqrt{n_0}) C(n) \right)
\end{aligned}$$

hold, where the series converges to some $C(n)$ for sufficiently large n by the ratio test and $C(n)$ decreases as n increases. The inequalities yield an upper bound that is independent of the choice of x_0, thereby ensuring uniform convergence.
In the case $T = T^{\mathbb{P}}$, Lemma 3.8 yields the inequality

$$\begin{aligned}
&|\bar{T}^G(x(n,x_0)) - \bar{T}(x(n,x_0))| \\
&= \left| \sum_{j=1}^{m} \left(\log \frac{x(n,x_0)_j}{n\pi_j} + 2r(x(n,x_0)_j) - 2r(n\pi_j) \right) \right| \\
&= \left| \sum_{j} \left(\log \frac{n\pi_j + \sqrt{n}c_j}{n\pi_j} + 2r(n\pi_j + \sqrt{n}c_j) - 2r(n\pi_j) \right) \right| \\
&< \sum_{j} \left(\left| \log \left(1 - \frac{\sqrt{n_0}r_0}{\sqrt{n}\pi_j} \right) \right| + \frac{2}{12(n\pi_j - \sqrt{nn_0}r_0)} \right),
\end{aligned}$$

which results in an upper bound that converges to zero independent of the choice of x_0. Hence

$$\bar{T}^{\chi^2} - \bar{T} = (\bar{T}^{\chi^2} - \bar{T}^G) + (\bar{T}^G - \bar{T})$$

Table 3.2 Runtime and p-values obtained by different methods for the five pairs from Table 3.1 in Section 3.4.1. Results from the full enumeration implemented by `xmulti` were included to show the agreement between the p-values produced by the exact methods. *Branch & Bound* refers to the implementation by Bejerano (2006) and *Dynamic* refers to the dynamic programming approach by Rahmann (2003) as implemented by myself with lattice size q. Times are in milliseconds.

Algorithm 3.1		Branch & Bound		`xmulti`		Dynamic ($q = 1000$)		Dynamic ($q = 10000$)	
p_{T^G}	time	p_{T^G}	time	p_{T^G}	time	p_{T^G}	time	p_{T^G}	time
0.0126	1.6	0.0126	2.7	0.0126	29.8	0.0141	22.2	0.0135	240.2
0.0361	3.5	0.0361	6.7	0.0361	29.1	0.0339	22.0	0.0359	237.2
0.0719	1.6	0.0719	5.8	0.0719	28.9	0.0675	21.2	0.0721	224.4
0.0461	0.9	0.0461	2.3	0.0461	29.3	0.0758	22.2	0.0460	241.4
0.0628	1.7	0.0628	5.0	0.0628	29.2	0.0967	21.8	0.0625	235.5

converges uniformly to zero as a function on $\partial\bar{B}_{1,r_0}(\pi)$ in the sense of the lemma. □

3.A.2 Comparison with other methods

As mentioned in Sections 3.1 and 3.3.3, approaches for computing exact multinomial p-values other than the full enumeration method exist. However, none of these methods have considered the probability mass statistic but have focused on the log-likelihood ratio statistic (Rahmann, 2003; Keich and Nagarajan, 2006) and other statistics from the family of power divergence statistics (Baglivo et al., 1992; Hirji, 1997; Bejerano et al., 2004). Adaptions of these methods to the probability mass statistic are beyond the scope of the present work.
Except for the branch and bound approach by Bejerano et al. (2004), these methods are not "strictly exact" but compute the distribution of a discretized test statistic under the null hypothesis (Keich and Nagarajan, 2006), thereby reducing the complexity of the resulting algorithms to polynomial time regardless of the number of categories m. While this approach seems to result in good approximations of very small p-values, which are of interest in some bioinformatics applications, the approximations are not exact and may differ quite strongly from the exact p-values of moderate size depending on the granularity of the discretization (see Table 3.2). This effect seems to be amplified by the fact that test statistic values span quite a large range but most of the probability mass is concentrated in a small part of this range. Of course, using finer discretizations improves these approximations, however, increasing the lattice size (i.e., the number of discretized values of the test statistic) increases the runtime (and memory usage) in practice. An instructive mathematical formulation of the idea as a dynamic programming problem is given by Rahmann (2003), which I implemented to obtain the results in Table 3.2. This approach has a complexity of $\mathcal{O}(mqn^2)$, where the lattice size $q \in \mathbb{N}$ needs to grow linearly with n to preserve

the accuracy of the approximation. The approach by Keich and Nagarajan (2006) reduces the complexity to $\mathcal{O}(mqn\log(n))$ (for the log-likelihood ratio statistic) by using a discrete Fourier transform to obtain the distribution of the discretized test statistic. As these approaches allow to approximate exact p-values when m is too large for exact algorithms to be feasible, such an approach may be added to the `ExactMultinom` package in a future version.
The branch and bound approach proposed by Bejerano et al. (2004) and implemented by Bejerano (2006) improves on the full enumeration method, while also suffering from exponential runtime in m. The implementation by Bejerano (2006) computes exact p-values for the log-likelihood ratio statistic and can be adapted to any statistic in the family of power divergence statistics. It provides optional speedups (one of which might be used to speed up the implementation of Algorithm 3.1 as well), which however may result in precision loss and were therefore not used for the computations in this study. Similar to the method proposed in this chapter, the runtime of the branch and bound approach depends on the null hypothesis parameter π and increases as the p-value decreases (Bejerano et al., 2004, Figure 5). The implementation of Algorithm 3.1 discussed in Section 3.3.3 outperformed the implementation by Bejerano (2006) in the experiments of this study (see Figure 3.6, Section 3.4.1 and Table 3.2), even though the former computes p-values for multiple test statistics simultaneously. Figure 3.6 discussed in Section 3.3.4 suggests that the branch and bound approach may have a complexity of $\mathcal{O}(n^{\frac{m}{2}})$ (in agreement with Figure 4 in Bejerano et al. (2004)). Adapting the branch and bound approach to the probability mass statistic is left as a subject for future research.

4 | Conditional Calibration, Reliability Diagrams, and Coefficient of Determination

This chapter is an extended version of Gneiting and Resin (2021) including content from Gneiting et al. (2023) and the discussion of a calibration test typically used to validate expert opinions.

Model diagnostics and forecast evaluation are two sides of the same coin. A common principle is that fitted or predicted distributions ought to be calibrated or reliable, ideally in the sense of auto-calibration, where the outcome is a random draw from the posited distribution. For binary responses, auto-calibration is the universal concept of reliability. For real-valued outcomes, a general theory of calibration has been elusive, despite a recent surge of interest in distributional regression and machine learning. We develop a framework rooted in probability theory, which gives rise to hierarchies of calibration, and applies to both predictive distributions and stand-alone point forecasts. In a nutshell, a prediction — distributional or single-valued — is conditionally T-calibrated if it can be taken at face value in terms of the functional T. Whenever T is defined via an identification function — as in the cases of threshold (non) exceedance probabilities, quantiles, expectiles, and moments — auto-calibration implies T-calibration. We introduce population versions of T-reliability diagrams and revisit a score decomposition into measures of miscalibration (MCB), discrimination (DSC), and uncertainty (UNC). In empirical settings, stable and efficient estimators of T-reliability diagrams and score components arise via nonparametric isotonic regression and the pool-adjacent-violators algorithm. For in-sample model diagnostics, we propose a universal coefficient of determination,

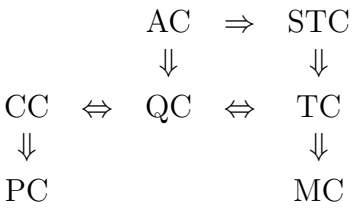
$$\mathrm{R}^* = \frac{\mathrm{DSC} - \mathsf{MCB}}{\mathsf{UNC}},$$

that nests and reinterprets the classical R^2 in least squares (mean) regression and its natural analog R^1 in quantile regression, yet applies to T-regression in general, with $\mathsf{MCB} \geq 0$, $\mathsf{DSC} \geq 0$, and $\mathrm{R}^* \in [0, 1]$ under modest conditions.

4.1 Introduction

Predictive distributions ought to be calibrated or reliable (Dawid, 1984; Gneiting and Katzfuss, 2014). More generally, statistical models ought to provide

(a) Under Assumption 4.15:



(b) Under Assumption 4.6:

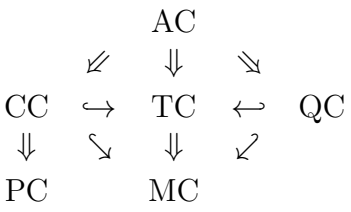


Figure 4.1 Preview of key findings in Section 4.2.3: Hierarchies of calibration (a) for continuous, strictly increasing cumulative distribution functions (CDFs) with common support and (b) under minimal conditions, with auto-calibration (AC) being the strongest notion. Conditional exceedance probability calibration (CC) is a conditional version of probabilistic calibration (PC), whereas threshold calibration (TC) is a conditional version of marginal calibration (MC). Quantile calibration (QC) differs from CC and TC in subtle ways. Strong threshold calibration (STC) is a stronger notion of threshold calibration introduced by Sahoo et al. (2021) for continuous CDFs. Hook arrows show conjectured implications.

plausible probabilistic explanations of observations, be it in-sample or out-of-sample, ideally in the sense of auto-calibration, meaning that the outcomes are indistinguishable from random draws from the posited distributions. For binary outcomes, auto-calibration is the universal standard of reliability. In the general case of linearly ordered, real-valued outcomes, weaker, typically unconditional facets of calibration have been studied, with probabilistic calibration, which corresponds to the uniformity of the probability integral transform (PIT; Dawid, 1984; Diebold et al., 1998), being the most popular notion. Recently, conditional notions have been proposed (Mason et al., 2007; Bentzien and Friederichs, 2014; Strähl and Ziegel, 2017), and there has been a surge of attention to calibration in the machine learning community, where the full conditional distribution of a response, given a feature vector, is of increasing interest, as exemplified by the work of Guo et al. (2017), Kuleshov et al. (2018), Song et al. (2019), Gupta et al. (2020), Zhao et al. (2020), Sahoo et al. (2021) and Roelofs et al. (2022). While in the literature on forecast evaluation predictive performance is judged out-of-sample, calibration is relevant in regression diagnostics as well, where in-sample goodness-of-fit is assessed via test statistics or criteria of R^2-type. In many ways, the complementary perspectives of model diagnostics and forecast evaluation are two sides of the same coin.

In this chapter, we strive to develop a theory of calibration for real-valued outcomes that complements the aforementioned strands of literature. Starting from measure theoretic and probabilistic foundations, we develop practical tools for visualizing, diagnosing and testing calibration, for both in-sample and out-of-sample settings, and applying both to full distributions and functionals thereof. Section 4.2 develops an overarching, rigorous theoretical framework in a general population setting, where we establish hierarchical relations between notions

of unconditional and conditional calibration, with Figure 4.1 summarizing key results. We reduce a posited distribution to a typically single-valued statistical functional, T, and define conditional calibration in terms of said functional. While in general auto-calibration fails to imply calibration in terms of a functional, we prove this implication for functionals defined via an identification function, such as event probabilities, means, quantiles, and generalized quantiles. We plot recalibrated values of the functional against posited values to obtain T-reliability diagrams and revisit extant score decompositions to define nonnegative measures of miscalibration (MCB), discrimination (DSC) and uncertainty (UNC), for which the mean score satisfies $\bar{S} = \mathsf{MCB} - \mathsf{DSC} + \mathsf{UNC}$. These considerations continue to apply when T-regression is studied as an end in itself, such as in mean (least squares) and quantile regression. In this setting, Theorem 4.26 establishes a general link between unconditional calibration and canonical score optimization, which nests classical results in least squares regression and the partitioning inequalities of quantile regression (Koenker and Bassett, 1978, Theorem 3.4).

In Section 4.3, we turn to empirical settings and statistical inference. We adopt and generalize the approach of Dimitriadis et al. (2021) that uses isotonic regression and the pool-adjacent-violators (PAV) algorithm (Ayer et al., 1955) to obtain consistent, optimally binned, reproducible, and PAV based (CORP) estimates of T-reliability diagrams and score components, along with uncertainty quantification via resampling. As opposed to extant estimators, the CORP approach yields non-decreasing reliability diagrams and guarantees the nonnegativity of the estimated MCB and DSC components. The regularizing constraint of isotonicity avoids artifacts and overfitting. For in-sample model diagnostics, we introduce a generalized coefficient of determination R^* that links to skill scores, and nests both the classical variance explained or R^2 in least squares regression (Kvålseth, 1985), and its natural analogue R^1 in quantile regression (Koenker and Machado, 1999). Subject to modest conditions $R^* \in [0, 1]$, with values of 0 and 1 indicating uninformative and immaculate fits, respectively.

In forecast evaluation, reliability diagrams and score components serve to diagnose and quantify performance on test samples. The most prominent case arises when T is the mean functional and performance is assessed by the mean squared error (MSE). As a preview of the diagnostic tools developed in this chapter, we assess point forecasts by Tredennick et al. (2021) of (log-transformed) butterfly population size from a ridge regression and a null model. The CORP mean reliability diagrams and MSE decompositions in Figure 4.2 show that, while both models are reliable, ridge regression enjoys considerably higher discrimination ability.

Section 4.4 briefly discusses the important case of quantile forecasts. The simultaneous prediction of multiple quantiles poses an appealing alternative to fully specified predictive distributions and single-valued point forecasts. Beyond reliability diagrams and score decompositions, coverage plots visualize deviations from unconditional quantile calibration, while p-values of simple multinomial tests may serve as a summary measure of simultaneous unconditional quantile calibration. The classical food expenditure data by Engel (1857) is used to contrast in-sample

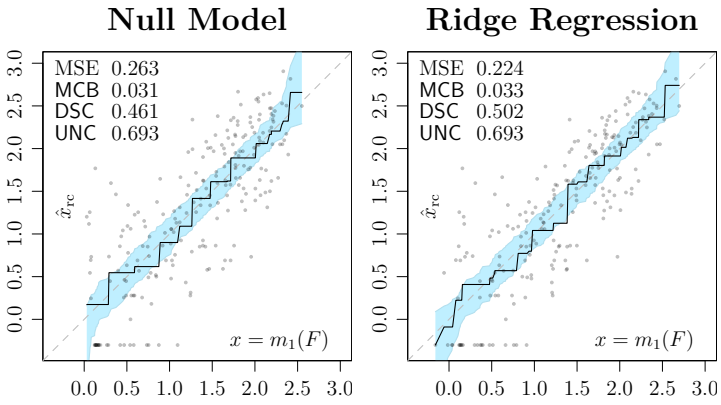


Figure 4.2 CORP mean reliability diagrams for point forecasts of (log-transformed) butterfly population size from the null model (left) and ridge regression (right) of Tredennick et al. (2021), along with 90% consistency bands and miscalibration (MCB), discrimination (DSC) and uncertainty (UNC) components of the mean squared error (MSE).

and out-of-sample evaluation of fitted quantiles at multiple levels.
The chapter closes in Section 4.5, where we discuss our findings and provide a roadmap for follow-up research. While Dimitriadis et al. (2021) introduced the CORP approach in the nested case of probability forecasts for binary outcomes, the setting of real-valued outcomes treated in this chapter is far more complex as it necessitates the consideration of statistical functionals in general. Throughout, we link the traditional case of regression diagnostics and (stand-alone) point forecast evaluation, where functionals such as conditional means, moments, quantiles, or expectiles are modeled and predicted, to model diagnostics and forecast evaluation in the fully distributional setting (Gneiting and Katzfuss, 2014; Hothorn et al., 2014). The Appendix 4.A includes material of more specialized or predominantly technical character.

4.2 Notions of Calibration, Reliability Diagrams, and Score Decompositions

Generally, we use the symbol $\mathcal{L}$ to denote a generic conditional or unconditional law or distribution, and we identify distributions with their cumulative distribution functions (CDFs). We write $\mathcal{N}(m, c^2)$ to denote a normal distribution with mean m and variance c^2, and we let φ and Φ denote the density and the CDF, respectively, of a standard normal variable.

4.2.1 Prediction spaces and prequential principle

We consider the joint law of a posited distribution and the respective outcome in the technical setting of Gneiting and Ranjan (2013). Specifically, let $(\Omega, \mathcal{A}, \mathbb{P})$ be

a *prediction space*, i.e., a probability space where the elementary elements $\omega \in \Omega$ correspond to realizations of the random triple

$$(F, Y, U),$$

where Y is the real-valued outcome, F is a posited distribution for Y in the form of a CDF, and U is uniformly distributed on the unit interval. Statements involving conditional or unconditional distributions, expectations, or probabilities, generally refer to the probability measure $\mathbb{P}$, which specifies the joint distribution of the forecast F and the outcome Y. The uniform random variable U allows for randomization. Throughout, we assume that U is independent of the σ-algebra generated by the random variable Y and the random function F in the technical sense detailed prior to Definition 2.6 in Strähl and Ziegel (2017).
Let $\mathcal{A}_0 \subseteq \mathcal{A}$ denote the forecaster's *information basis*, i.e., a sub-σ-algebra such that F is measurable with respect to $\mathcal{A}_0$. Then F is *ideal* relative to $\mathcal{A}_0$ (Gneiting and Ranjan, 2013) if

$$F(y) = \mathbb{P}\left(Y \leq y \mid \mathcal{A}_0\right) \quad \text{almost surely,} \quad \text{for all} \quad y \in \mathbb{R}.$$

If F is ideal relative to some sub σ-algebra $\mathcal{A}_0$, then it is *auto-calibrated* (Tsyplakov, 2013) in the sense that

$$F(y) = \mathbb{P}\left(Y \leq y \mid F\right) \quad \text{almost surely,} \quad \text{for all} \quad y \in \mathbb{R},$$

which is equivalent to being ideal relative to the information basis $\sigma(F) \subseteq \mathcal{A}_0$. Extensions to prediction spaces with tuples $(Y, F_1, \ldots, F_k, U)$ that allow for multiple CDF-valued forecasts $F_1, \ldots, F_k$ with associated information bases $\mathcal{A}_1, \ldots, \mathcal{A}_k \subset \mathcal{A}$ are straightforward.

Example 4.1 (Gneiting and Ranjan (2013); Pohle (2020))**.** Conditionally on a standard normal variate μ, let the outcome Y be normal with mean μ and variance 1. Then the *perfect* forecast $F_1 = \mathcal{N}(\mu, 1)$ is ideal relative to the information basis $\mathcal{A}_1 = \sigma(\mu)$ generated by μ. The *unconditional* forecast $F_2 = \mathcal{N}(0, 2)$ agrees with the marginal distribution of the outcome Y and is ideal relative to the trivial σ-algebra $\mathcal{A}_2 = \{\emptyset, \Omega\}$.

More elaborate notions of prediction spaces are feasible. In particular, one might include a covariate or feature vector Z and consider random tuples of the form (Z, F, Y, U). Indeed, the transdisciplinary scientific literature has considered reliability relative to covariate information, under labels such as *strong* (Van Calster et al., 2016) or *individual* (Chung et al., 2021; Zhao et al., 2020) calibration. We refrain from doing so as our simple setting adheres to the *prequential principle* posited by Dawid (1984), according to which predictive performance needs to be evaluated on the basis of the tuple (F, Y) only, without consideration of the forecast-generating mechanism. The aforementioned extensions become critical in studies of cross-calibration (Strähl and Ziegel, 2017), stratification (Ehm and Ovcharov, 2017; Ferro et al., 2020; Allen, 2021), sensitivity (Fissler and Pesenti, 2022), and fairness (Pleiss et al., 2017; Mitchell et al., 2021).

4.2.2 Traditional notions of unconditional calibration

Let us recall the classical notions of calibration of predictive distributions for real-valued outcomes. In order to do so, we define the *probability integral transform* (PIT)

$$Z_F = F(Y-) + U\,(F(Y) - F(Y-)) \tag{4.1}$$

of the CDF-valued random quantity F, where $F(y-) = \lim_{x \uparrow y} F(x)$ denotes the left-hand limit of F at $y \in \mathbb{R}$, and the random variable U is standard uniform and independent of F and Y. The PIT of a continuous CDF F is simply $Z_F = F(Y)$. The predictive distribution F is *probabilistically calibrated* or *PIT calibrated* if Z_F is uniformly distributed on the unit interval. The use of the probabilistic calibration criterion was suggested by Dawid (1984) and popularized by Diebold et al. (1998), who proposed the use of PIT histograms as a diagnostic tool. Importantly, in continuous settings probabilistic calibration implies that prediction intervals bracketed by quantiles capture the outcomes at the respective nominal level.
Furthermore, the predictive distribution F is *marginally calibrated* (Gneiting et al., 2007) if

$$\mathbb{E}[F(y)] = \mathbb{P}\,(Y \leq y) \quad \text{for all} \quad y \in \mathbb{R}.$$

Thus, for a marginally calibrated predictive distribution, the frequency of (not) exceeding a threshold value matches the posited unconditional probability.

Example 4.2. In the setting of Example 4.1, let η attain the values $\pm\eta_0$ with equal probability, independently of μ and Y, where $\eta_0 > 0$. Then the *unfocused* forecast with CDF

$$F(y) = \frac{1}{2}\,(\Phi(y-\mu) + \Phi(y-\mu-\eta)) \tag{4.2}$$

is probabilistically calibrated but fails to be marginally calibrated (Gneiting et al., 2007). Similarly, let δ take the values $\pm\delta_0$ with equal probability, independently of μ and Y, where $\delta_0 \in (0,1)$. Then the *lopsided* forecast F with density

$$f(y) = (1-\delta)\varphi(y-\mu)\,\mathbb{1}(y < \mu) + (1+\delta)\varphi(y-\mu)\,\mathbb{1}(y > \mu) \tag{4.3}$$

is marginally calibrated but fails to be probabilistically calibrated. For details see Appendix 4.A.1.2.

It is well known that an ideal forecast is both probabilistically calibrated and marginally calibrated (Gneiting and Ranjan, 2013, Theorem 2.8; Song et al., 2019, Theorem 1). Reformulated in terms of auto-calibration the following holds.

Theorem 4.3. *Auto-calibration implies marginal and probabilistic calibration.*

Auto-calibration thus is a stronger requirement than either marginal or probabilistic calibration, and the latter are logically independent. However, in the special case of a binary outcome, probabilistic calibration and auto-calibration

are equivalent (Gneiting and Ranjan, 2013, Theorem 2.11), and auto-calibration serves as a universal notion of calibration. In the case of three or more distinct outcomes, Gneiting and Ranjan (2013) conjectured that auto-calibration is stronger than simultaneous marginal and probabilistic calibration. We resolve and prove their conjecture within the following example.

Example 4.4. We begin by considering continuous CDFs and then discuss a discrete example with three distinct outcomes only.

(a) Suppose that μ is normal with mean 0 and variance c^2. Conditionally on μ, let the *piecewise uniform* predictive distribution F be a mixture of uniform measures on $[\mu, \mu+1]$, $[\mu+1, \mu+2]$, and $[\mu+2, \mu+3]$ with weights p_1, p_2 and p_3, respectively, and let the outcome Y be drawn from a mixture with weights q_1, q_2 and q_3 on these intervals. Finally, let the tuple $(p_1, p_2, p_3; q_1, q_2, q_3)$ attain each of the values

$$\left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}; \tfrac{5}{10}, \tfrac{1}{10}, \tfrac{4}{10}\right), \quad \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}; \tfrac{1}{10}, \tfrac{8}{10}, \tfrac{1}{10}\right), \quad \text{and} \quad \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}; \tfrac{4}{10}, \tfrac{1}{10}, \tfrac{5}{10}\right),$$

with $\mathbb{P}$-probability $\frac{1}{3}$. Evidently, F fails to be auto-calibrated. However, F is marginally calibrated as, conditionally on μ, it assigns the same mass $\frac{1}{3}$ to each of the intervals, in agreement with the conditional distribution of Y. As for the PIT Z_F, conditionally on μ its CDF is piecewise linear on the partition induced by 0, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, and 1. Thus, in order to establish probabilistic calibration it suffices to verify that $\mathbb{P}(Z_F \leq x) = x$ for $x \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$, as confirmed by elementary calculations. Integration over μ completes the argument.

(b) For a full resolution of the aforementioned conjecture by Gneiting and Ranjan (2013), we fix $\mu = 0$ and replace the intervals by fixed numbers $y_1 < y_2 < y_3$. Thus, F assigns mass p_j to y_j, whereas the event $Y = y_j$ realizes with probability q_j for $j = 1$, 2, and 3. The forecast remains probabilistically and marginally calibrated, and fails to be auto-calibrated.

4.2.3 Conditional calibration

While checks for probabilistic calibration have become a cornerstone of predictive distribution evaluation (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007), both marginal and probabilistic calibration concern unconditional facets of predictive performance, which is increasingly being considered insufficient (e.g., Levi et al., 2022). Stronger conditional notions of calibration, which condition on facets of the predictive distribution, have emerged in various strands of the scientific literature. For example, Mason et al. (2007) used conditional (non) exceedance probabilities (CEP) to assess the calibration of ensemble weather forecasts. These were used by Held et al. (2010) and Strähl and Ziegel (2017) to derive calibration tests, which operate under the hypothesis that the forecast F is *CEP calibrated* in the sense that

$$\mathbb{P}\left(Z_F \leq \alpha \mid q_\alpha^-(F)\right) = \alpha \quad \text{almost surely,} \quad \text{for all} \quad \alpha \in (0,1), \tag{4.4}$$

where $q_\alpha^-(F) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$ denotes the (lower) α-quantile of F. Similarly, Henzi et al. (2021) introduced the notion of a *threshold calibrated* forecast F, which stipulates that

$$\mathbb{P}\left(Y \leq t \mid F(t)\right) = F(t) \quad \text{almost surely,} \quad \text{for all} \quad t \in \mathbb{R}. \tag{4.5}$$

Essentially, CEP calibration is a conditional version of probabilistic calibration, and threshold calibration is conditional marginal calibration.

Theorem 4.5. *CEP calibration implies probabilistic calibration, and threshold calibration implies marginal calibration.*

Proof. Immediate by taking unconditional expectations, as noted by Henzi et al. (2021). □

Variants of these concepts can be found scattered in the literature. Notably, Sahoo et al. (2021) introduce a notion of calibration for continuous predictive distributions, which requires that

$$\mathbb{P}(Z_F \leq \alpha \mid F(t)) = \alpha \quad \text{almost surely,} \quad \text{for all} \quad \alpha \in (0, 1),\ t \in \mathbb{R}. \tag{4.6}$$

As in Figure 4.1, we refer to this property as *strong threshold calibration*. The notion is weaker than auto-calibration, but it implies both CEP calibration and threshold calibration, subject to conditions that we discuss below.
We proceed to the general notion of conditional T-calibration in terms of a statistical functional T as introduced by Arnold (2020) and Ferro et al. (2020). Other authors (Pohle, 2020; Krüger and Ziegel, 2021) refer to this notion or special cases thereof as auto-calibration with respect to T. A statistical functional on some class $\mathcal{F}$ of probability measures is a measurable function $\mathrm{T}\colon \mathcal{F} \to \mathcal{T}$ into a (typically, finite-dimensional) space $\mathcal{T}$ with Borel-σ-algebra $\mathcal{B}(\mathcal{T})$. Technically, we work in the prediction space setting under a natural measurability condition that is not restrictive (Fissler and Holzmann, 2022).

Assumption 4.6. The class $\mathcal{F}$ and the functional T are such that $F \in \mathcal{F}$, the mapping $\mathrm{T}(F)\colon (\Omega, \mathcal{A}) \to (\mathcal{T}, \mathcal{B}(\mathcal{T}))$ is measurable, and $\mathcal{L}\left(Y \mid \mathrm{T}(F)\right) \in \mathcal{F}$ almost surely.

Definition 4.7. Under Assumption 4.6, the predictive distribution F is *conditionally* T-*calibrated*, or simply T-*calibrated*, if

$$\mathrm{T}\left(\mathcal{L}(Y \mid \mathrm{T}(F))\right) = \mathrm{T}(F) \quad \text{almost surely.}$$

Essentially, under a T-calibrated predictive distribution F, we can take $\mathrm{T}(F)$ at face value. Perhaps surprisingly, an auto-calibrated forecast is not necessarily T-calibrated, as noted by Arnold (2020, Section 3.2). For a simple counterexample, consider the perfect forecast from Example 4.1, which fails to be T-calibrated when T is the variance, the standard deviation, the interquartile range, or a related measure of dispersion.

We proceed to show that this startling issue does not occur with *identifiable* functionals, i.e., functionals induced by an identification function (see Theorem 4.11). Similar to the classical procedure in M-estimation (Huber, 1964), an identification function weighs negative values in the case of underprediction against positive values in the case of overprediction, and the corresponding functional maps to the possibly set-valued argument at which an associated expectation changes sign. Following Jordan et al. (2022), a measurable function $V\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is an *identification function* if $V(\cdot, y)$ is increasing and left-continuous for all $y \in \mathbb{R}$. We operate under Assumption 4.6 with the implicit understanding that $V(x, \cdot)$ is quasi-integrable with respect to all $F \in \mathcal{F}$ for all $x \in \mathbb{R}$. Then, for any probability measure F in the class $\mathcal{F}$, the functional $\mathrm{T}(F)$ *induced* by V is defined as

$$\mathrm{T}(F) = [\mathrm{T}^-(F), \mathrm{T}^+(F)] \subseteq [-\infty, +\infty] = \overline{\mathbb{R}},$$

where the lower and upper bounds are given by the random variables

$$\mathrm{T}^-(F) = \sup\left\{x : \int V(x, y)\,\mathrm{d}F(y) < 0\right\} \tag{4.7}$$

and

$$\mathrm{T}^+(F) = \inf\left\{x : \int V(x, y)\,\mathrm{d}F(y) > 0\right\}. \tag{4.8}$$

An identifiable functional T is of *singleton type* if $\mathrm{T}(F)$ is a singleton for every $F \in \mathcal{F}$. Otherwise, T is of *interval type*. Table 4.1 lists key examples, such as threshold-defined event probabilities, quantiles, expectiles, and moments. The definition of the Huber functional involves the clipping function $\kappa_{a,b}(t) = \max(\min(t, b), -a)$ with parameters $a, b > 0$ (Taggart, 2022). In the limiting cases as $a = b \to 0$ and $a = b \to \infty$, the Huber functional recovers the α-quantile (q_α) and the α-expectile (e_α), respectively.
For identifiable functionals, we can define an unconditional notion of T-calibration as well. Note that in contrast to traditional settings, where F is fixed, we work in the prediction space setting, where F is a random CDF. In principle, the subsequent Definitions 4.9 and 4.24 depend on the choice of the identification function. However, as we demonstrate in Appendix 4.A.1.4, the following condition ensures that the identification function is unique, up to a positive constant, so that ambiguities are avoided.

Assumption 4.8. The identification function V induces the functional T on a convex class $\mathcal{F}_0 \supseteq \mathcal{F}$ of probability measures, which contains the Dirac measures δ_y for all $y \in \mathbb{R}$. The identification function V is

(i) of *prediction error form*, i.e., there exists an increasing, left-continuous function $v\colon \mathbb{R} \to \mathbb{R}$ such that $V(x, y) = v(x - y)$ with $v(-r) < 0$ and $v(r) > 0$ for some $r > 0$, *or*

(ii) of the form $V(x, y) = x - \mathrm{T}(\delta_y)$ for a functional T of singleton type.

Table 4.1 Key examples of identifiable functionals with associated parameters, identification function, and generic type. For a similar listing see Table 1 in Jordan et al. (2022).

Functional	Parameters	Identification function	Type
Threshold (non) exceedance	$t \in \mathbb{R}$	$V(x,y) = x - \mathbb{1}\{y \leq t\}$	singleton
Mean		$V(x,y) = x - y$	singleton
Median		$V(x,y) = \mathbb{1}\{y < x\} - \frac{1}{2}$	interval
Moment of order n (m_n)	$n = 1, 2, \ldots$	$V(x,y) = x - y^n$	singleton
α-Expectile (e_α)	$\alpha \in (0,1)$	$V(x,y) = \lvert\mathbb{1}\{y < x\} - \alpha\rvert\,(x-y)$	singleton
α-Quantile (q_α)	$\alpha \in (0,1)$	$V(x,y) = \mathbb{1}\{y < x\} - \alpha$	interval
Huber	$\alpha \in (0,1)$, $a, b > 0$	$V(x,y) = \lvert\mathbb{1}\{y < x\} - \alpha\rvert\,\kappa_{a,b}(x-y)$	interval

The examples in Table 4.1 all satisfy Assumption 4.8.

Definition 4.9. Suppose that the functional T is generated by an identification function V, and let Assumptions 4.6 and 4.8 hold. Then the predictive distribution F is *unconditionally* T*-calibrated* if

$$\mathbb{E}[V(\mathrm{T}^+(F) - \varepsilon, Y)] \leq 0 \quad \text{and} \quad \mathbb{E}[V(\mathrm{T}^-(F) + \varepsilon, Y)] \geq 0 \quad \text{for all} \quad \varepsilon > 0. \tag{4.9}$$

For the interval-valued α-quantile functional, $q_\alpha(F) = [q_\alpha^-(F), q_\alpha^+(F)]$, condition (4.9) reduces to the traditional unconditional coverage condition

$$\mathbb{P}\left(Y \leq q_\alpha^-(F)\right) \geq \alpha \quad \text{and} \quad \mathbb{P}\left(Y \geq q_\alpha^+(F)\right) \geq 1 - \alpha, \tag{4.10}$$

with the latter being equivalent to $\mathbb{P}(Y < q_\alpha^+(F)) \leq \alpha$. Probabilistic calibration implies unconditional α-quantile calibration at every level $\alpha \in (0,1)$, as hinted at by Kuleshov et al. (2018, Section 3.1).[1] Under technical assumptions, condition (4.9) simplifies to

$$\mathbb{E}[V(\mathrm{T}(F), Y)] = 0, \tag{4.11}$$

with the classical unbiasedness condition $\mathbb{E}[m_1(F)] = \mathbb{E}[Y]$ arising in the case of the mean or expectation functional.

Example 4.10. Let T be the mean functional or a quantile. Then the unfocused forecast from Example 4.2 is unconditionally T-calibrated but fails to be conditionally T-calibrated. For details see Figure 4.4 and Appendix 4.A.1.1.

[1] To verify this implication, it suffices to note that if F is probabilistically calibrated, then

$$\alpha = \mathbb{P}(Z_F \leq \alpha) \leq \mathbb{P}(F(Y-) \leq \alpha) = \mathbb{P}(Y \leq q_\alpha^-(F)),$$
$$1 - \alpha = \mathbb{P}(Z_F > \alpha) \leq \mathbb{P}(F(Y) > \alpha) = \mathbb{P}(Y \geq q_\alpha^+(F)).$$

As Example 4.14(b) demonstrates, the reverse implication does not hold in general. However, Assumption 4.15 ensures the equivalence of probabilistic calibration and unconditional α-quantile calibration at every level $\alpha \in (0,1)$.

Importantly, for any identifiable functional auto-calibration implies both unconditional and conditional T-calibration, as we demonstrate now. Note that Assumption 4.6 is a minimal condition as it is required to define conditional T-calibration in the first place.

Theorem 4.11. *Suppose that the functional* T *is generated by an identification function and Assumption 4.6 holds. Then auto-calibration implies conditional* T*-calibration, and, subject to Assumption 4.8, conditional* T*-calibration implies unconditional* T*-calibration.*

Proof. The statements in this proof are understood to hold almost surely. By Theorem 4.34 and Proposition 4.35 of Breiman (1992) in concert with auto-calibration, F is a regular conditional distribution of Y given F, and we conclude that

$$\mathbb{E}\left[V(x, Y) \mid F\right] = \int V(x, y) \, \mathrm{d}F(y).$$

Furthermore, a regular conditional distribution $F_{\mathrm{T}} = \mathcal{L}(Y \mid \mathrm{T}(F))$ of Y given $\mathrm{T}(F)$ exists, and the tower property of conditional expectation implies that

$$\begin{aligned}\int V(x, y) \, \mathrm{d}F_{\mathrm{T}}(y) &= \mathbb{E}\left[V(x, Y) \,|\, \mathrm{T}(F)\right] = \mathbb{E}\left[\mathbb{E}\left[V(x, Y) \,|\, F\right] \,|\, \mathrm{T}(F)\right] \\ &= \mathbb{E}\left[\int V(x, y) \, \mathrm{d}F(y) \,\middle|\, \mathrm{T}(F)\right].\end{aligned}$$

Let $\mathrm{T}(F) = [\mathrm{T}^-(F), \mathrm{T}^+(F)]$ and $\mathrm{T}(F_{\mathrm{T}}) = [\mathrm{T}^-(F_{\mathrm{T}}), \mathrm{T}^+(F_{\mathrm{T}})]$, where the boundaries are random variables. The proof of the first part is complete if we can show that $\mathrm{T}^-(F_{\mathrm{T}}) = \mathrm{T}^-(F)$ and $\mathrm{T}^+(F_{\mathrm{T}}) = \mathrm{T}^+(F)$.
Let $\varepsilon > 0$. By the definition of $\mathrm{T}^+(F)$, we know that $\int V(\mathrm{T}^+(F), y) \, \mathrm{d}F(y) \leq 0$ and $\int V(\mathrm{T}^+(F) + \varepsilon, y) \, \mathrm{d}F(y) > 0$. Using nested conditional expectations as above, the same inequalities hold almost surely when integrating with respect to F_{T}. Hence, by the definition of $\mathrm{T}^+(F_{\mathrm{T}})$, we obtain $\mathrm{T}^+(F) \leq \mathrm{T}^+(F_{\mathrm{T}}) < \mathrm{T}^+(F) + \varepsilon$. An analogous argument shows that $\mathrm{T}^-(F) - \varepsilon \leq \mathrm{T}^-(F_{\mathrm{T}}) \leq \mathrm{T}^-(F) + \varepsilon$, which completes the proof of the first part and shows that F is conditionally T-calibrated.
Finally, if F is conditionally T-calibrated, unconditional T-calibration follows by taking nested expectations in the terms in the defining inequalities. □

An analogous result is easily derived for CEP calibration.

Theorem 4.12. *Under Assumption 4.6 for quantiles, auto-calibration implies CEP calibration.*

Proof. It holds that

$$\mathbb{P}(Z_F \leq \alpha \mid q_\alpha^-(F)) = \mathbb{E}\left[\mathbb{1}\{Z_F \leq \alpha\} \mid q_\alpha^-(F)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{Z_F \leq \alpha\} \mid F\right] \mid q_\alpha^-(F)\right]$$

almost surely for $\alpha \in (0, 1)$. As F is a version of $\mathcal{L}(Y \mid F)$, the nested expectation equals α almost surely by Proposition 2.1 of Rüschendorf (2009), which implies CEP calibration. □

When evaluating full predictive distributions, it is natural to consider families of functionals as in the subsequent definition, where part (a) is compatible with the extant notion in (4.5).

Definition 4.13. A predictive distribution F is

(a) *threshold calibrated* if it is conditionally $F(t)$-calibrated for all $t \in \mathbb{R}$;

(b) *quantile calibrated* if it is conditionally q_α-calibrated for all $\alpha \in (0, 1)$;

(c) *expectile calibrated* if it is conditionally e_α-calibrated for all $\alpha \in (0, 1)$;

(d) *moment calibrated* if it is conditionally n-th moment calibrated for all integers $n = 1, 2, \ldots$

While CEP, quantile, and threshold calibration are closely related notions, they generally are not equivalent. For illustration, we consider predictive CDFs in the spirit of Example 4.4.

Example 4.14.

(a) Let $\mu \sim \mathcal{N}(0, c^2)$. Conditionally on μ, let F be a mixture of uniform distributions on the intervals $[\mu, \mu + 1], [\mu + 1, \mu + 2], [\mu + 2, \mu + 3]$, and $[\mu + 3, \mu + 4]$ with weights p_1, p_2, p_3, and p_4, respectively, and let Y be from a mixture with weights q_1, q_2, q_3, and q_4. Furthermore, let the tuple $(p_1, p_2, p_3, p_4; q_1, q_2, q_3, q_4)$ attain each of the values

$$\begin{aligned} &\left(\tfrac{1}{2}, 0, \tfrac{1}{2}, 0; \tfrac{3}{4}, 0, \tfrac{1}{4}, 0\right), \quad \left(\tfrac{1}{2}, 0, 0, \tfrac{1}{2}; \tfrac{1}{4}, 0, 0, \tfrac{3}{4}\right), \\ &\left(0, \tfrac{1}{2}, \tfrac{1}{2}, 0; 0, \tfrac{1}{4}, \tfrac{3}{4}, 0\right), \quad \left(0, \tfrac{1}{2}, 0, \tfrac{1}{2}; 0, \tfrac{3}{4}, 0, \tfrac{1}{4}\right) \end{aligned}$$

with equal probability. Then the *continuous* forecast F is threshold calibrated and CEP calibrated but fails to be quantile calibrated.

(b) Let the tuple $(p_1, p_2, p_3; q_1, q_2, q_3)$ attain each of the values

$$\left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}; \tfrac{5}{10}, \tfrac{4}{10}, \tfrac{1}{10}\right), \quad \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}; \tfrac{1}{10}, \tfrac{5}{10}, \tfrac{4}{10}\right), \quad \text{and} \quad \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}; \tfrac{4}{10}, \tfrac{1}{10}, \tfrac{5}{10}\right)$$

with equal probability. Let F assign mass p_j to numbers y_j for $j = 1, 2, 3$, where $y_1 < y_2 < y_3$, and let the event $Y = y_j$ realize with probability q_j. The resulting *discrete* forecast is quantile and threshold calibrated. However, it fails to be CEP calibrated or even PIT calibrated.

Under the following conditions CEP, quantile, and threshold calibration coincide.

Assumption 4.15. In addition to Assumption 4.6 for quantiles and threshold (non) exceedances at all levels and thresholds, respectively, let the following hold.

(i) The CDFs in the class $\mathcal{F}$ are continuous and strictly increasing on a common support interval.

(ii) There exists a countable set $\mathcal{G} \subseteq \mathcal{F}$ such that $\mathbb{P}(F \in \mathcal{G}) = 1$.

Theorem 4.16.

(a) Under Assumption 4.15(i) CEP and quantile calibration are equivalent and imply probabilistic calibration.

(b) Under Assumptions 4.15(i)–(ii) CEP, quantile and threshold calibration are equivalent and imply both probabilistic and marginal calibration.

Proof. By Assumption 4.15(i) the CDFs $F \in \mathcal{F}$ are invertible on the common support with the quantile function $\alpha \mapsto q_\alpha^-(F)$ as inverse. Hence, for every $\alpha \in (0,1)$ the functional q_α is of singleton-type and $q_\alpha(F) = \{q_\alpha^-(F)\}$. In this light, the almost-sure identity

$$\mathbb{P}(Z_F \leq \alpha \mid q_\alpha^-(F)) = \mathbb{P}(Y \leq q_\alpha^-(F) \mid q_\alpha(F))$$

implies part (a). To prove part (b), let $\mathcal{G}$ be as in Assumption 4.15(ii) and assume without loss of generality that $\mathbb{P}(F = G) > 0$ for all $G \in \mathcal{G}$. If $\alpha \in (0,1)$ and $t \in \mathbb{R}$ are such that $\mathbb{P}(F(t) = \alpha) > 0$, then

$$\mathbb{P}(Y \leq t \mid F(t) = \alpha) = \mathbb{P}(Y \leq q_\alpha^-(F) \mid q_\alpha^-(F) = t),$$

where Assumption 4.15(ii) ensures that the events conditioned on have positive probability. Hence, quantile and threshold calibration are equivalent. The remaining implications are immediate from Theorem 4.5. □

We conjecture that the statement of part (b) holds under Assumption 4.15(i) alone but are unaware of a measure theoretic argument that serves to generalize the discrete reasoning in our proof. As indicated in panel (b) of Figure 4.1, we also conjecture that CEP or quantile calibration imply threshold calibration in general, though we have not been able to prove this implication, nor can we show that CEP or quantile calibration imply marginal calibration in general. Strong threshold calibration as defined in (4.6) implies both CEP and threshold calibration under Assumption 4.15, by arguments similar to those in the above proof. The following result thus demonstrates that the hierarchies in panel (a) and, with the aforementioned exceptions, in panel (b) of Figure 4.1 are complete, with the caveat that hierarchies may collapse if the class $\mathcal{F}$ is sufficiently small, as exemplified by Theorem 2.11 of Gneiting and Ranjan (2013).

Proposition 4.17. *Under Assumption 4.15(i)–(ii) the following hold:*

(a) Strong threshold calibration does not imply auto-calibration.

(b) Joint CEP, quantile, and threshold calibration does not imply strong threshold calibration.

(c) Joint probabilistic and marginal calibration does not imply threshold calibration.

(d) Probabilistic calibration does not imply marginal calibration.

(e) Marginal calibration does not imply probabilistic calibration.

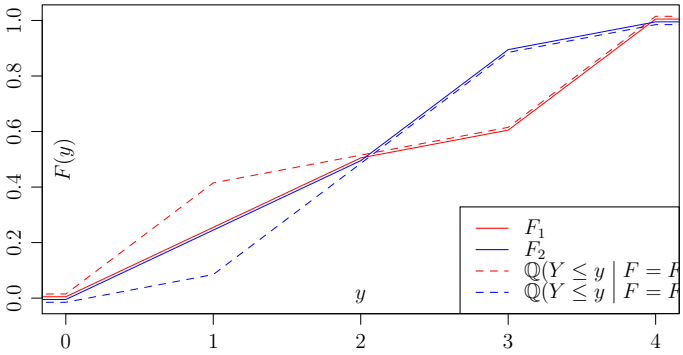


Figure 4.3 The *equiprobable* predictive distribution F picks the piecewise linear, partially (namely, for $y \leq 2$) identical CDFs F_1 and F_2 with equal probability. It is jointly CEP, quantile, and threshold calibrated but fails to be auto-calibrated.

Table 4.2 Properties of the forecasts in our examples. We note whether they are auto-calibrated (AC), CEP calibrated (CC), quantile calibrated (QC), threshold calibrated (TC), probabilistically calibrated (PC), or marginally calibrated (MC), and whether the involved distributions are continuous and strictly increasing on a common support (CSI) as in Assumption 4.15(i). Except for the auto-calibrated cases, the forecasts fail to be moment calibrated.

Source	Forecast Type	CSI	AC	CC	QC	TC	PC	MC
Example 4.1	Perfect	✓	✓	✓	✓	✓	✓	✓
Example 4.1	Unconditional	✓	✓	✓	✓	✓	✓	✓
Figure 4.3	Equiprobable	✓		✓	✓	✓	✓	✓
Example 4.4	Piecewise uniform	as $c \to 0$					✓	✓
Example 4.2	Unfocused	✓					✓	
Example 4.2	Lopsided	✓						✓
Example 4.14	Continuous			✓		✓	✓	✓
Example 4.14	Discrete				✓	✓		✓

Proof. We establish the claims in a series of (counter) examples, starting with part (b), where we present an example based on two equiprobable, partially overlapping CDFs in Figure 4.3. A similar example based on four equiprobable, partially overlapping CDFs in Appendix 4.A.1.5 yields part (a). As for part (c), we return to the piecewise uniform forecast in Example 4.4, where for simplicity we fix $\mu = 0$. This forecast is probabilistically and marginally calibrated, but it fails to be threshold calibrated because

$$\mathbb{P}\left(Y \leq \tfrac{3}{2} \mid F(\tfrac{3}{2}) = \tfrac{5}{8}\right) = \tfrac{5}{10} + \tfrac{1}{2} \cdot \tfrac{1}{10} = \tfrac{11}{20} \neq \tfrac{5}{8}.$$

As for parts (d) and (e), we refer to the unfocused and lopsided forecasts from Example 4.2. □

Clearly, further hierarchical relations are immediate. For example, given that

probabilistic calibration does not imply marginal calibration, it does not imply threshold calibration nor auto-calibration. We leave further discussion to future work but note that moment calibration does not imply probabilistic nor marginal calibration, as follows easily from classical results on the moment problem (e.g., Stoyanov, 2000). For an overview of calibration properties in our examples, see Table 4.2.

4.2.4 Reliability diagrams

As we proceed to define reliability diagrams, it is useful to restrict attention to single-valued functionals. To this end, if an identifiable functional T is of interval type, we instead consider its single-valued lower or upper bound, $\mathrm{T}^-(F)$ or $\mathrm{T}^+(F)$, which we call the *lower* and *upper version* of T, or simply the *lower* and *upper functional*, without explicit reference to the original functional T. The following result demonstrates that T-calibration implies calibration of the upper and lower functional.

Proposition 4.18. *Suppose that the functional* T *is generated by an identification function* V*, and let Assumption 4.6 hold. Then conditional* T*-calibration implies conditional* T^-*- and* T^+*-calibration, and, subject to Assumption 4.8, unconditional* T*-calibration implies unconditional* T^-*- and* T^+*-calibration.*

Proof. Suppose that T^* is the lower or upper version of a functional T generated by the identification function V. As $\sigma(\mathrm{T}^*(F)) \subseteq \sigma(\mathrm{T}(F))$, we find that

$$\mathbb{E}\left[V(x,Y) \mid \mathrm{T}^*(F)\right] = \mathbb{E}\left[\mathbb{E}\left[V(x,Y) \mid \mathrm{T}(F)\right] \mid \mathrm{T}^*(F)\right]$$

is almost surely ≤ 0 if $x < \mathrm{T}^*(F)$, and almost surely ≥ 0 if $x > \mathrm{T}^*(F)$. Hence, $\mathrm{T}^*(F) \in \mathrm{T}(\mathcal{L}(Y \mid \mathrm{T}^*(F)))$. If T^* is the lower functional, the former inequality is strict and hence $\mathrm{T}^-(F) = \min \mathrm{T}(\mathcal{L}(Y \mid \mathrm{T}^-(F)))$, whereas if T^* is the upper functional, the latter is strict and hence $\mathrm{T}^+(F) = \max \mathrm{T}(\mathcal{L}(Y \mid \mathrm{T}^+(F)))$. Unconditional T^*-calibration is an immediate consequence of unconditional T-calibration. □

In this light, we restrict attention to single-valued functionals that are lower or upper versions of identifiable functionals, or identifiable functionals of singleton type. Any such functional can be associated with a random variable $X = \mathrm{T}(F)$, and we call any random variable X_{rc}, for which

$$X_{\mathrm{rc}} = \mathrm{T}\left(\mathcal{L}\left(Y \mid X\right)\right) \tag{4.12}$$

almost surely, a *recalibrated* version of X. Clearly, we can also define X_{rc} for a stand-alone point forecast X, based on conceptualized distributions, by resorting to the joint distribution of the random tuple (X, Y), provided the right-hand side of (4.12) is well defined and finite almost surely. The point forecast X is *conditionally* T*-calibrated*, or simply T*-calibrated*, if $X = X_{\mathrm{rc}}$ almost surely. Subject to Assumption 4.8, X is *unconditionally* T*-calibrated* if

$$\mathbb{E}[V(X-\varepsilon, Y)] \leq 0 \quad \text{and} \quad \mathbb{E}[V(X+\varepsilon, Y)] \geq 0 \quad \text{for all} \quad \varepsilon > 0. \tag{4.13}$$

For recent discussions of the particular cases of the mean or expectation and quantile functionals see, e.g., Nolde and Ziegel (2017, Sections 2.1–2.2). Patton (2020, Proposition 2), Krüger and Ziegel (2021, Definition 3.1) and Satopää (2021, Section 2).
To compare the posited functional X with its recalibrated version X_{rc}, we introduce the T-reliability diagram.

Assumption 4.19. The functional T is a lower or upper version of an identifiable functional, or an identifiable functional of singleton type. The point forecast X is a random variable, and the recalibrated forecast $X_{\mathrm{rc}} = \mathrm{T}(\mathcal{L}(Y \mid X))$ is well defined and finite almost surely.

Definition 4.20. Under Assumption 4.19, the T-*reliability diagram* is the graph of a mapping $x \mapsto \mathrm{T}\left(\mathcal{L}(Y \mid X = x)\right)$ on the support of X.

While technically the T-reliability diagram depends on the choice of a regular conditional distribution for the outcome Y, this issue is not a matter of practical relevance. Evidently, for a T-calibrated forecast the T-reliability diagram is concentrated on the diagonal. Conversely, deviations from the diagonal indicate violations of T-calibration and can be interpreted diagnostically, as illustrated in Figure 4.4 for threshold, quantile, and moment calibration. For a similar display in the specific case of mean calibration see Figure 1 of Pohle (2020).
In the setting of fully specified predictive distributions, the distinction between unconditional and conditional T-calibration is natural. Perhaps surprisingly, the distinction vanishes in the setting of stand-alone point forecasts if the associated identification function is of prediction error form and the forecast and the residual are independent.

Theorem 4.21. *Let Assumption 4.19 hold, and suppose that the underlying identification function V satisfies Assumption 4.8. Suppose furthermore that the point forecast X and the generalized residual $\mathrm{T}(\delta_Y) - X$ are independent. Then X is conditionally* T*-calibrated if, and only if, it is unconditionally* T*-calibrated.*

Proof. Given any constant $c \in \mathbb{R}$ it holds that

$$\mathbb{E}[V(X + c, Y) \mid X] = \mathbb{E}[v(\mathrm{T}(\delta_Y) - X - c) \mid X] = \mathbb{E}[v(\mathrm{T}(\delta_Y) - X - c)].$$

In view of (4.12) and (4.13), conditional and unconditional T-calibration are equivalent. □

For quantiles, expectiles, and Huber functionals, the identification function V is of prediction error form and the generalized residual reduces to the standard residual, $X - Y$. In particular, this observation applies in the case of least squares regression, where T is the mean functional, and the forecast and the residual have typically been assumed to be independent in the literature. We discuss the statistical implications of Theorem 4.21 in Appendix 4.A.2.

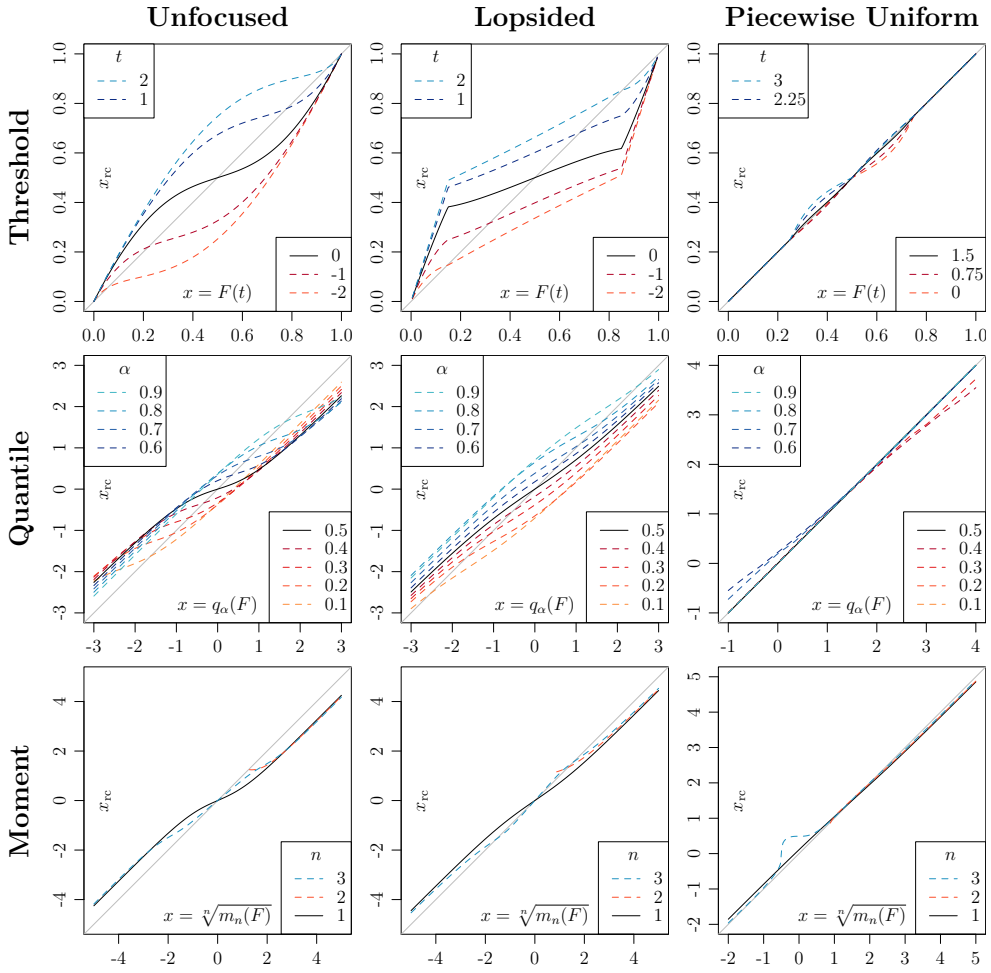


Figure 4.4 Threshold (top), quantile (middle), and moment (bottom) reliability diagrams for point forecasts induced by (left) the unfocused forecast with $\eta_0 = 1.5$ and (middle) the lopsided forecast with $\delta_0 = 0.7$ from Example 4.2, and (right) the piecewise uniform forecast with $c = 0.5$ from Example 4.4. Each display plots recalibrated against original values. Deviations from the diagonal indicate violations of T-calibration. For details see Appendix 4.A.1.

4.2.5 Score decompositions

We now revisit a score decomposition into measures of miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) based on consistent scoring functions. Specifically, suppose that S is a *consistent* loss or scoring function for the functional T on the class $\mathcal{F}$ in the sense that

$$\mathbb{E}_F[\mathrm{S}(t, Y)] \leq \mathbb{E}_F[\mathrm{S}(x, Y)]$$

for all $F \in \mathcal{F}$, all $t \in \mathrm{T}(F) = [\mathrm{T}^-(F), \mathrm{T}^+(F)]$ and all $x \in \mathbb{R}$ (Savage, 1971; Gneiting, 2011a). If the inequality is strict unless $x \in \mathrm{T}(F)$, then S is *strictly consistent*. Consistent scoring functions serve as all-purpose performance measures that elicit fair and honest assessments and reward the utilization of broad

information bases (Holzmann and Eulert, 2014). If the functional T is of interval type, a consistent scoring function S is consistent for both T^- and T^+, but strict consistency is lost when T is replaced by its lower or upper version and S is strictly consistent for T. For prominent examples of consistent scoring functions, see Table 4.3.

A functional is *elicitable* if it admits a strictly consistent scoring function (Gneiting, 2011a). Under general conditions, elicitability is equivalent to identifiability (Steinwart et al., 2014, Theorem 5). The respective functionals allow for both principled relative forecast evaluation through the use of consistent scoring functions, and principled absolute forecast evaluation via T-reliability diagrams and score decompositions, as discussed in what follows.

Let $\mathcal{L}(Y)$ denote the unconditional distribution of the outcome and suppose that $x_0 = \mathrm{T}(\mathcal{L}(Y))$ is well defined. As before, we operate under Assumption 4.19 and work with $X = \mathrm{T}(F)$, its recalibrated version X_{rc}, and the reference forecast x_0. Again, the simplified notation accommodates stand-alone point forecasts, and it suffices to consider the joint distribution of the tuple (X, Y). Following the lead of Dawid (1986) in the case of binary outcomes, and Ehm and Ovcharov (2017) and Pohle (2020) in the setting of point forecasts for real-valued outcomes, we consider the expected scores

$$\bar{\mathrm{S}} = \mathbb{E}[\mathrm{S}(X, Y)], \quad \bar{\mathrm{S}}_{\mathrm{rc}} = \mathbb{E}[\mathrm{S}(X_{\mathrm{rc}}, Y)], \quad \text{and} \quad \bar{\mathrm{S}}_{\mathrm{mg}} = \mathbb{E}[\mathrm{S}(x_0, Y)] \tag{4.14}$$

for the forecast at hand, its recalibrated version, and the marginal reference forecast x_0, respectively.

Definition 4.22. Let Assumption 4.19 hold, and let $x_0 = \mathrm{T}(\mathcal{L}(Y))$ and the expectations $\bar{\mathrm{S}}$, $\bar{\mathrm{S}}_{\mathrm{rc}}$, and $\bar{\mathrm{S}}_{\mathrm{mg}}$ in (4.14) be well defined and finite. Then we refer to

$$\mathsf{MCB}_{\mathrm{S}} = \bar{\mathrm{S}} - \bar{\mathrm{S}}_{\mathrm{rc}}, \quad \mathsf{DSC}_{\mathrm{S}} = \bar{\mathrm{S}}_{\mathrm{mg}} - \bar{\mathrm{S}}_{\mathrm{rc}}, \quad \text{and} \quad \mathsf{UNC}_{\mathrm{S}} = \bar{\mathrm{S}}_{\mathrm{mg}},$$

as *miscalibration*, *discrimination*, and *uncertainty*, respectively.

The following result decomposes the expected score $\bar{\mathrm{S}}$ for the forecast at hand into miscalibration ($\mathsf{MCB}_{\mathrm{S}}$), discrimination ($\mathsf{DSC}_{\mathrm{S}}$), and uncertainty ($\mathsf{UNC}_{\mathrm{S}}$) components.

Theorem 4.23 (Dawid (1986), Pohle (2020))**.** *In the setting of Definition 4.22, suppose that the scoring function* S *is consistent for the functional* T*. Then it holds that*

$$\bar{\mathrm{S}} = \mathsf{MCB}_{\mathrm{S}} - \mathsf{DSC}_{\mathrm{S}} + \mathsf{UNC}_{\mathrm{S}}, \tag{4.15}$$

where $\mathsf{MCB}_{\mathrm{S}} \geq 0$ *with equality if* X *is conditionally* T*-calibrated, and* $\mathsf{DSC}_{\mathrm{S}} \geq 0$ *with equality if* $X_{\mathrm{rc}} = x_0$ *almost surely. If* S *is strictly consistent then* $\mathsf{MCB}_{\mathrm{S}} = 0$ *only if* X *is conditionally* T*-calibrated, and* $\mathsf{DSC}_{\mathrm{S}} = 0$ *only if* $X_{\mathrm{rc}} = x_0$ *almost surely.*

A remaining question is what consistent scoring function S ought to be used in practice. To address this issue, we resort to mixture or Choquet representations

Table 4.3 Canonical loss functions in the sense of Definition 4.24.

Functional	Parameter	Canonical Loss
Moment of order n	$n = 1, 2, \ldots$	$\mathrm{S}(x, y) = (x - y^n)^2$
α-Expectile	$\alpha \in (0, 1)$	$\mathrm{S}(x, y) = 2\,\lvert \mathbb{1}\{x \geq y\} - \alpha \rvert\,(x - y)^2$
α-Quantile	$\alpha \in (0, 1)$	$\mathrm{S}(x, y) = 2\,(\mathbb{1}\{x \geq y\} - \alpha)\,(x - y)$

of consistent loss functions, as introduced by Ehm et al. (2016) for quantiles and expectiles and developed in full generality by Dawid (2016), Ziegel (2016) and Jordan et al. (2022). Specifically, we rely on an obvious generalization of Proposition 2.6 of Jordan et al. (2022), as noted at the start of their Section 2. Let T be identifiable with identification function V satisfying Assumption 4.8, and let $\eta \in \mathbb{R}$. Then the *elementary* loss function S_η, given by

$$\mathrm{S}_\eta(x, y) = (\mathbb{1}\{\eta \leq x\} - \mathbb{1}\{\eta \leq y\})\, V(\eta, y), \tag{4.16}$$

is consistent for T. As an immediate consequence, any well-defined function of the form

$$\mathrm{S}(x, y) = \int_{\mathbb{R}} \mathrm{S}_\eta(x, y)\, \mathrm{d}H(\eta), \tag{4.17}$$

where H is a locally finite measure on $\mathbb{R}$, is consistent for T. If T is a quantile, an expectile, an event probability or a moment, then the construction includes all consistent scoring functions, subject to standard conditions, and agrees with suitably adapted classes of generalized piecewise linear (GPL) and Bregman functions, respectively (Gneiting, 2011a; Ehm et al., 2016).
We now formalize what Ehm and Ovcharov (2017, p. 477) call the "most prominent" choice, namely, scoring functions for which the mixing measure H in the representation (4.17) is uniform.

Definition 4.24. Suppose that the functional T is generated by an identification function V satisfying Assumption 4.8, with elementary loss functions S_η as defined in (4.16). Then a loss function S is *canonical* for T if it is nonnegative and admits a representation of the form

$$\mathrm{S}(x, y) = a \int_{\mathbb{R}} \mathrm{S}_\eta(x, y)\, \mathrm{d}\lambda(\eta) + b(y), \tag{4.18}$$

where λ is the Lebesgue measure, $a > 0$ is a constant, and b is a measurable function.

Clearly, any canonical loss function is a consistent scoring function for T. Furthermore, if the identification function is of the prediction error form, then any canonical loss function has score differentials that are invariant under translation in the sense that $\mathrm{S}(x_1 + c, y + c) - \mathrm{S}(x_2 + c, y + c) = \mathrm{S}(x_1, y) - \mathrm{S}(x_2, y)$. Conversely, we note from Section 5.1 of Savage (1971) that for the mean functional the canonical loss functions are the only consistent scoring functions of this type.

Table 4.4 Components of the decomposition (4.15) for the mean squared error (MSE) under mean-forecasts induced by the predictive distributions in Examples 4.1 and 4.2. Uncertainty (UNC) equals 2 irrespective of the forecast at hand. The term $I(\eta_0)$ is in integral form and can be evaluated numerically. For details see Appendix 4.A.1.

Predictive Distribution	Mean-Forecast	MSE	MCB	DSC
Perfect	μ	1	0	1
Unconditional	0	2	0	0
Unfocused	$\mu + \frac{1}{2}\eta$	$1 + \frac{1}{4}\eta_0^2$	$(\frac{1}{4} - I(\eta_0))\eta_0^2$	$1 - I(\eta_0)\eta_0^2$
Lopsided	$\mu + \frac{\sqrt{2}}{\sqrt{\pi}}\delta$	$1 + \frac{2}{\pi}\delta_0^2$	$(\frac{1}{4} - I(\sqrt{\frac{8}{\pi}}\delta_0))\frac{8}{\pi}\delta_0^2$	$1 - I(\sqrt{\frac{8}{\pi}}\delta_0)\frac{8}{\pi}\delta_0^2$

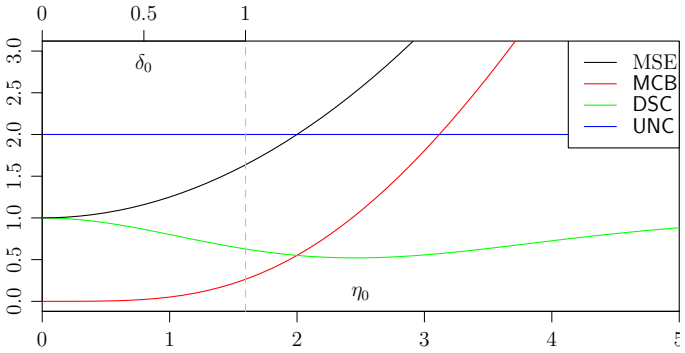


Figure 4.5 Components of the decomposition (4.15) for the mean squared error (MSE) under mean-forecasts induced by the unfocused and lopsided predictive distributions from Example 4.2 and Table 4.4, as functions of $\eta_0 \geq 0$ and $\delta_0 \in (0, 1)$, respectively.

Typically, one chooses the constant $a > 0$ and the measurable function $b(y)$ in (4.18) such that the canonical loss admits a concise closed form, as exemplified in Table 4.3. Since any selection incurs the same point forecast ranking, we refer to the choice in Table 4.3 as *the* canonical loss function. The most prominent example arises when T is the mean functional, where the ubiquitous *quadratic* or *squared error* scoring function,

$$\mathrm{S}(x, y) = (x - y)^2, \tag{4.19}$$

is canonical. In this case, the UNC component equals the unconditional variance of Y as x_0 is simply the marginal mean μ_Y of Y, and the MCB and DSC components of the general score decomposition (4.15) are

$$\mathsf{MCB} = \mathbb{E}\,(X - X_{\mathrm{rc}})^2 \quad \text{and} \quad \mathsf{DSC} = \mathbb{E}\,(X_{\mathrm{rc}} - \mu_Y)^2\,,$$

respectively. Note that here and in the following, we drop the subscript S whenever we use a canonical loss. Table 4.4 and Figure 4.5 provide explicit examples.

In the nested case of a binary outcome Y, where X and X_{rc} specify event probabilities, the quadratic loss function reduces to the Brier score (Gneiting and Raftery, 2007), and we refer to Dimitriadis et al. (2021) and references therein for details on score decompositions. In the case of threshold calibration, the point forecast $x = F(t)$ is induced by a predictive distribution, and the Brier score can be written as

$$\mathrm{S}(x, y) = \left(F(t) - \mathbb{1}\{y \leq t\}\right)^2. \tag{4.20}$$

For both real-valued and binary outcomes, it is often preferable to use the square root of the miscalibration component ($\mathsf{MCB}^{1/2}$) as a measure of calibration error that can be interpreted on natural scales (e.g., Roelofs et al., 2022).
A canonical loss function for the Huber functional (Table 4.1) is given by

$$\mathrm{S}(x, y) = 2\,|\mathbb{1}\{x \geq y\} - \alpha| \begin{cases} 2a|x-y| - a^2, & x - y < -a, \\ (x-y)^2, & -a \leq x - y \leq b, \\ 2b|x-y| - b^2, & x - y > b; \end{cases}$$

cf. Taggart (2022, Definition 4.2). In the limiting case as $a = b \to \infty$, we recover the canonical loss functions for the α-expectile, which include the quadratic loss in (4.19). Similarly, if we rescale suitably and take the limit as $a = b \to 0$, we recover the asymmetric *piecewise linear* or *pinball* loss, as listed in Table 4.3, which lies at the heart of quantile regression.
We move on to a remarkable property of canonical loss functions. In a nutshell, the point forecast X is unconditionally T-calibrated if, and only if, the expected canonical loss deteriorates under translation. This property, which nests classical results in regression theory, as we demonstrate at the end of Section 4.3.3, does not hold under consistent scoring functions in general.[2]

Assumption 4.25. The point forecast X, the functional T and the identification function V satisfy Assumptions 4.8 and 4.19, and S is a canonical loss for T. Furthermore, $\mathbb{E}[\mathrm{S}(X + \eta, Y)]$ and $\mathbb{E}[V(X + \eta, Y)]$ are well defined and locally bounded as functions of $\eta \in \mathbb{R}$.

Theorem 4.26. *Under Assumption 4.25, the point forecast X is unconditionally* T*-calibrated if, and only if,*

$$\mathbb{E}\left[\mathrm{S}(X + c, Y)\right] \geq \mathbb{E}\left[\mathrm{S}(X, Y)\right] \quad \textit{for all} \quad c \in \mathbb{R}.$$

[2]In particular, the statement in Theorem 4.26 does not hold for arbitrary consistent scoring functions. For a counterexample, consider the empirical distribution of $(x_1, y_1), \ldots, (x_{10}, y_{10})$, where $x_i = i$ and $y_i = x_i + \frac{10}{9}$ for $i = 1, \ldots, 9$, and $x_{10} = y_{10} = -10$. The respective mean-forecast X fails to be unconditionally mean calibrated, whereas the shifted version $X + 1$ is unconditionally mean calibrated. Nonetheless, the expected elementary score (4.16) for the mean functional (i.e., $V(x, y) = x - y$) with index $\eta = -\frac{19}{2}$ increases when X gets replaced with $X + 1$.

Proof. If X is unconditionally T-calibrated and $c > 0$, then

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{S}(X + c, Y)\right] - \mathbb{E}\left[\mathrm{S}(X, Y)\right] \\
&= \mathbb{E}\left[\int \left(\mathbb{1}\{\eta \leq X + c\} - \mathbb{1}\{\eta \leq X\}\right) V(\eta, Y)\, \mathrm{d}\eta\right] \\
&= \mathbb{E}\left[\int_{(0,c]} V(X + \eta, Y)\, \mathrm{d}\eta\right] = \int_{(0,c]} \mathbb{E}\left[V(X + \eta, Y)\right]\, \mathrm{d}\eta
\end{aligned} \tag{4.21}$$

is nonnegative by the second part of the unconditional T-calibration criterion (4.13). Conversely, if the score difference in (4.21) is nonnegative for all $c > 0$, then so is

$$\mathbb{E}[V(X + c, Y)] = \frac{1}{c} \int_{(0,c]} \mathbb{E}[V(X + c, Y)]\, \mathrm{d}\eta \geq \frac{1}{c} \int_{(0,c]} \mathbb{E}[V(X + \eta, Y)]\, \mathrm{d}\eta.$$

Hence, the second part of (4.13) is satisfied.
An analogous argument shows that the score difference (4.21) is nonnegative for all $c < 0$ if, and only if, the first inequality in (4.13) is satisfied. □

As a consequence, under a canonical loss function the MCB component in the score decomposition (4.15) of Theorem 4.23 decomposes into nonnegative *unconditional* and *conditional* components $\mathsf{MCB_u}$ and $\mathsf{MCB_c}$, respectively, subject to the mild condition that unconditional recalibration via translation is feasible.

Theorem 4.27. *Let Assumption 4.25 hold, and suppose there is a constant c such that $X + c$ is unconditionally* T*-calibrated. Let $X_{\mathrm{urc}} = X + c$ and $\bar{\mathrm{S}}_{\mathrm{urc}} = \mathbb{E}[\mathrm{S}(X_{\mathrm{urc}}, Y)]$, and define*

$$\mathsf{MCB_u} = \bar{\mathrm{S}} - \bar{\mathrm{S}}_{\mathrm{urc}} \quad \textit{and} \quad \mathsf{MCB_c} = \bar{\mathrm{S}}_{\mathrm{urc}} - \bar{\mathrm{S}}_{\mathrm{rc}}.$$

Then

$$\mathsf{MCB} = \mathsf{MCB_u} + \mathsf{MCB_c},$$

where $\mathsf{MCB_u} \geq 0$ with equality if X is unconditionally T*-calibrated, and $\mathsf{MCB_c} \geq 0$ with equality if $X_{\mathrm{rc}} = X_{\mathrm{urc}}$ almost surely. If* S *is strictly consistent, then $\mathsf{MCB_u} = 0$ only if X is unconditionally* T*-calibrated, and $\mathsf{MCB_c} = 0$ only if $X_{\mathrm{rc}} = X_{\mathrm{urc}}$ almost surely.*

Proof. Immediate from Theorems 4.23 and 4.26, and the fact that conditional recalibration of X and $X + c$ yields the same X_{rc}. □

In settings that are equivariant under translation, such as for expectiles, quantiles, and Huber functionals when both X and Y are supported on the real line, X can always be unconditionally recalibrated by adding a constant. Under any canonical loss function S, the basic decomposition (4.15) then extends to

$$\bar{\mathrm{S}} = \mathsf{MCB_u} + \mathsf{MCB_c} - \mathsf{DSC} + \mathsf{UNC}. \tag{4.22}$$

For instance, when $\mathrm{S}(x, y) = (x-y)^2$ is the canonical loss for the mean functional, $\mathsf{MCB}_\mathsf{u} = c^2$ is the squared unconditional bias. The forecasts in Figure 4.5 and Table 4.4 are free of unconditional bias, so $\mathsf{MCB}_\mathsf{u} = 0$ and $\mathsf{MCB}_\mathsf{c} = \mathsf{MCB}$.
In all cases studied thus far, canonical loss functions are strictly consistent (Ehm et al., 2016), and so $\mathsf{MCB}_\mathsf{u} = 0$ if, and only if, the forecast is unconditionally T-calibrated, and $\mathsf{MCB}_\mathsf{c} = 0$ if, and only if, $X_{\mathrm{urc}} = X_{\mathrm{rc}}$ almost surely. While in other settings, such as when the outcomes are bounded, unconditional recalibration by translation might be counterintuitive (in principle) or impossible (in practice), the statement of Theorem 4.27 continues to hold, and the above results can be refined to admit more general forms of unconditional recalibration. We leave these and other ramifications to future work.

4.3 Empirical Reliability Diagrams and Score Decompositions: The CORP Approach

We turn to empirical settings, where calibration checks, scores, and score decompositions address critical practical problems in both model diagnostics and forecast evaluation. The most direct usage is in the evaluation of out-of-sample predictive performance, where forecasts may either take the form of fully specified predictive distributions, or be single-valued point forecasts that arise, implicitly or explicitly, as functionals of predictive distributions. Similarly, in model diagnostics, where in-sample goodness-of-fit is of interest, the model might supply fully specified, parametric or non-parametric conditional distributions, or single-valued regression output that is interpreted as a functional of an underlying, implicit or explicit, probability distribution. Prominent examples for the latter setting include ordinary least squares regression, where the mean or expectation functional is sought, and quantile regression.
In the case of fully specified predictive distributions, we work with tuples of the form

$$(F_1, y_1), \ldots, (F_n, y_n), \tag{4.23}$$

where F_i is a posited conditional CDF for the real-valued observation y_i for $i = 1, \ldots, n$, which we interpret as a sample from an underlying population $\mathbb{P}$ in the prediction space setting of Section 4.2. In the case of stand-alone point forecasts or regression output, we assume throughout that the functional T is of the type stated in Assumption 4.19 and work with tuples of the form

$$(x_1, y_1), \ldots, (x_n, y_n), \tag{4.24}$$

where $x_i = \mathrm{T}(F_i) \in \mathbb{R}$ derives explicitly or implicitly from a predictive distribution F_i for $i = 1, \ldots, n$.
In the remainder of the section, we introduce empirical versions of T-reliability diagrams (Definition 4.20) and score components (Definition 4.22) for samples of the form (4.23) or (4.24), which allow for both diagnostic checks and inference about an underlying population $\mathbb{P}$. While practitioners may think of our empirical

Algorithm 4.1. General T-PAV algorithm based on data of the form (4.24)

Input: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ where $x_1 \leq \cdots \leq x_n$
Output: T-calibrated values $\widehat{x}_1, \ldots, \widehat{x}_n$
partition into groups $G_{1:1}, \ldots, G_{n:n}$ and let $\widehat{x}_i = \mathrm{T}(\delta_i)$ for $i = 1, \ldots, n$
while *there are groups* $G_{k:i}$ *and* $G_{(i+1):l}$ *such that* $\widehat{x}_1 \leq \cdots \leq \widehat{x}_i$ *and* $\widehat{x}_i > \widehat{x}_{i+1}$ **do**
  merge $G_{k:i}$ and $G_{(i+1):l}$ into $G_{k:l}$ and let $\widehat{x}_i = \mathrm{T}(\delta_{k:l})$ for $i = k, \ldots, l$
end

versions exclusively from diagnostic perspectives, we emphasize that they can be interpreted as estimators of the population quantities and be analyzed as such. A key feature of our approach is the use of nonparametric isotonic regression via the pool-adjacent-violators algorithm, as proposed by Dimitriadis et al. (2021) in the particular case of binary outcomes. The generalization that we discuss here is hinted at in the discussion section of their paper.

4.3.1 The T-pool-adjacent-violators (T-PAV) algorithm

Our key tool and workhorse is a very general version of the classical pool-adjacent-violators (PAV) algorithm for nonparametric isotonic regression (Ayer et al., 1955; Van Eeden, 1958). Historically, work on the PAV algorithm has focused on the mean functional, as reviewed by Barlow et al. (1972), Robertson and Wright (1980), and de Leeuw et al. (2009), among others. In contrast, Jordan et al. (2022) study the PAV algorithm in very general terms that accommodate our setting.
We rely on their work and describe the T-*pool-adjacent-violators* algorithm based on tuples $(x_1, y_1), \ldots, (x_n, y_n)$ of the form (4.24), where without loss of generality we may assume that $x_1 \leq \cdots \leq x_n$. Furthermore, we let δ_i denote the point measure in the outcome y_i. More generally, for $1 \leq k \leq l \leq n$ we let

$$\delta_{k:l} = \frac{1}{l-k+1} \sum_{i=k}^{l} \delta_i$$

be the associated empirical measure. Algorithm 4.1 describes the generation of an increasing sequence $\widehat{x}_1 \leq \cdots \leq \widehat{x}_n$ of recalibrated values, which by construction are conditionally T-calibrated with respect to the empirical measure associated with $(\widehat{x}_1, y_1), \ldots, (\widehat{x}_n, y_n)$. The algorithm rests on partitions of the index set $\{1, \ldots, n\}$ into groups $G_{k:l} = \{k, \ldots, l\}$ of consecutive integers. The following result summarizes the remarkable properties of the T-PAV algorithm, as proved in Section 3.2 of Jordan et al. (2022).

Theorem 4.28 (Jordan et al. (2022))**.** *Suppose that the functional* T *is as stated in Assumption 4.19. Then Algorithm 4.1 generates a sequence* $\widehat{x}_1, \ldots, \widehat{x}_n$ *such that the empirical measure associated with* $(\widehat{x}_1, y_1), \ldots, (\widehat{x}_n, y_n)$ *is conditionally* T*-calibrated. This sequence is optimal with respect to any scoring function* S *of the form* (4.17)*, in that*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(\widehat{x}_i, y_i) \leq \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(t_i, y_i) \tag{4.25}$$

for any non-decreasing sequence $t_1 \leq \cdots \leq t_n$*.*

We note that for a functional of interval type, the minimum on the left-hand side of (4.25) is the same under the lower and upper version, respectively. For customary functionals, such as threshold (non) exceedance probabilities, quantiles, expectiles, and moments, the optimality is universal, as functions of the form (4.17) exhaust the class of the T-consistent scoring functions subject to mild conditions (Ehm et al., 2016). While the PAV algorithm has been used extensively for the recalibration of probabilistic classifiers (e.g., Flach, 2012), we are unaware of any extant work that uses Algorithm 4.1 for forecast recalibration, forecast evaluation, or model diagnostics in non-binary settings.

4.3.2 Empirical T-reliability diagrams

Recently, Dimitriadis et al. (2021) introduced the CORP approach for the estimation of reliability diagrams and score decompositions in the case of probability forecasts for binary outcomes. In a nutshell, the acronym CORP refers to an estimator that is Consistent under the assumption of isotonicity for the population recalibration function and Optimal in both finite sample and asymptotic settings, while facilitating Reproducibility, and being based on the PAV algorithm. Here, we extend the CORP approach and employ nonparametric isotonic T-regression via the T-PAV algorithm under Assumption 4.19, where T is the lower or upper version of an identifiable functional, or an identifiable singleton functional.
We begin by defining the empirical T-reliability diagram, which is a sample version of the population diagram in Definition 4.20.

Definition 4.29. Let the functional T be as stated in Assumption 4.19, and suppose that $\widehat{x}_1, \ldots, \widehat{x}_n$ originate from tuples $(x_1, y_1), \ldots, (x_n, y_n)$ with $x_1 \leq \cdots \leq x_n$ via Algorithm 4.1. Then the CORP *empirical* T*-reliability diagram* is the graph of the piecewise linear function that connects the points $(x_1, \widehat{x}_1), \ldots, (x_n, \widehat{x}_n)$ in the Euclidean plane.

A few scattered references in the literature on forecast evaluation have proposed displays of recalibrated against original values for functionals other than binary event probabilities: Figures 3 and 7 of Bentzien and Friederichs (2014) and Figure 8 of Pohle (2020) consider quantiles, and Figures 2–5 of Satopää and Ungar (2015) concern the mean functional. However, none of these papers employ the

PAV algorithm, and the resulting diagrams are subject to issues of stability and efficiency, as illustrated by Dimitriadis et al. (2021) in the case of binary outcomes. For the CORP empirical T-reliability diagram to be consistent in the sense of large sample convergence to the population version of Definition 4.20, the assumption of isotonicity of the population recalibration function needs to be invoked. As argued by Roelofs et al. (2022) and Dimitriadis et al. (2021), such an assumption is natural, and practitioners tend to dismiss nonisotonic recalibration functions as artifacts. Evidently, these arguments transfer to arbitrary functionals, and any violations of the isotonicity assumption entail horizontal segments in CORP reliability diagrams, thereby indicating a lack of reliability. Large sample theory for CORP estimates of the recalibration function and the T-reliability diagram depends on the functional T, the type — discrete or continuous — of the marginal distribution of the point forecast X, and smoothness conditions. Mösching and Dümbgen (2020) establish rates of uniform convergence in the cases of threshold (non) exceedance and quantile functionals that complement classical theory (Barlow et al., 1972; Casady and Cryer, 1976; Wright, 1984; Robertson et al., 1988; El Barmi and Mukerjee, 2005; Guntuboyina and Sen, 2018).

In the case of binary outcomes, Bröcker and Smith (2007, p. 651) argue that reliability diagrams ought to be supplemented by *consistency bars* for "immediate visual evaluation as to just how likely the observed relative frequencies are under the assumption that the predicted probabilities are reliable." Dimitriadis et al. (2021) develop asymptotic and Monte Carlo based methods for the generation of consistency bands to accompany a CORP reliability diagram for dichotomous outcomes, and provide code in the form of the `reliabilitydiag` package (Dimitriadis and Jordan, 2021) for `R` (`R` Core Team, 2022). The consistency bands quantify and visualize the variability of the empirical reliability diagram under the respective null hypothesis, i.e., they show the pointwise range of the CORP T-reliability diagram that we expect to see under a calibrated forecast. Algorithms 4.2 and 4.3 in Appendix 4.A.2 generalize this approach to produce consistency bands from data of the form (4.23) under the assumption of auto-calibration. In the specific case of threshold calibration, where the induced outcome is dichotomous, the assumptions of auto-calibration (in the binary setting) and T-calibration (for the non-exceedance functional) coincide (Gneiting and Ranjan, 2013, Theorem 2.11), and we use the aforementioned algorithms to generate consistency bands (Figure 4.6, top row). Generally, auto-calibration is a strictly stronger assumption than T-calibration, with ensuing issues, which we discuss in Appendix 4.A.2.1. Furthermore, to generate consistency bands from data of the form (4.24), we cannot operate under the assumption of auto-calibration.

As a crude yet viable alternative, we propose in Appendix 4.A.2.2 a Monte Carlo technique for the generation of consistency bands that is based on resampling residuals. As in traditional regression diagnostics, the approach depends on the assumption of independence between point forecasts and residuals. Figure 4.6 shows examples of T-reliability diagrams with associated residual-based 90% consistency bands for the perfect, unfocused, and lopsided forecasts from Section 4.2 for the mean functional (middle row) and the lower quantile functional at level

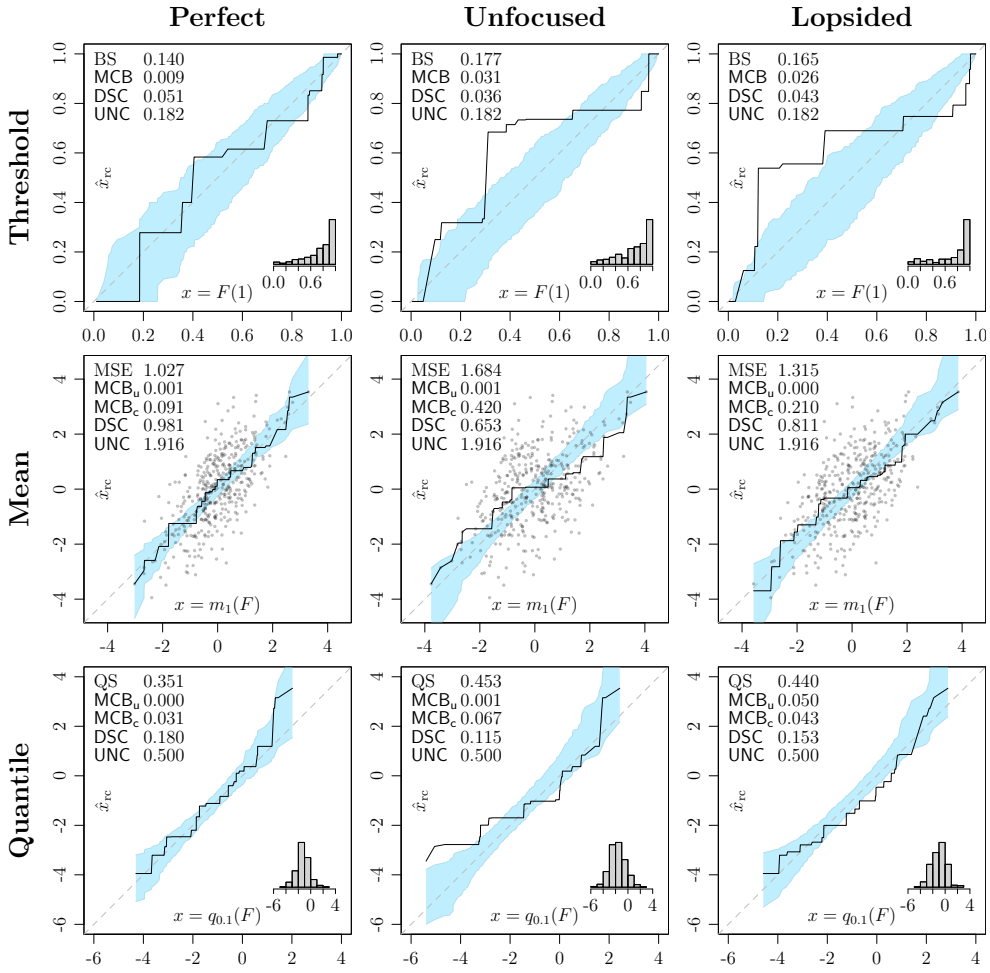


Figure 4.6 CORP empirical threshold (top, $t = 1$), mean (middle) and quantile (bottom, $\alpha = 0.10$) reliability diagrams for the perfect (left), unfocused (middle), and lopsided (right) forecast from Examples 4.1 and 4.2 with 90% consistency bands and CORP score components under the associated canonical loss function based on samples of size 400.

0.10 (bottom row). For further discussion see Appendix 4.A.2.2. In the case of the mean functional, we add the scatter diagram for the original data of the form (4.24), whereas in the other two cases, inset histograms visualize the marginal distribution of the point forecast.

We encourage follow-up work on both Monte Carlo and asymptotic methods for the generation of consistency and confidence bands that are tailored to specific functionals of interest, similar to the analysis by Dimitriadis et al. (2021) in the basic case of probability forecasts for binary outcomes.

4.3.3 Empirical score decompositions

In this section, we consider data $(x_1, y_1), \ldots, (x_n, y_n)$ of the form (4.24), where implicitly or explicitly $x_i = \mathrm{T}(F_i)$ for a single-valued functional T. Let $\widehat{x}_1, \ldots, \widehat{x}_n$ denote the respective T-PAV recalibrated values, and let $\widehat{x}_0 = \mathrm{T}(\widehat{F}_0)$, where $\widehat{F}_0$ is the empirical CDF of the outcomes $y_1, \ldots, y_n$. Let

$$\widehat{\mathrm{S}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(x_i, y_i), \quad \widehat{\mathrm{S}}_{\mathrm{rc}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(\widehat{x}_i, y_i), \quad \text{and} \quad \widehat{\mathrm{S}}_{\mathrm{mg}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(\widehat{x}_0, y_i) \tag{4.26}$$

denote the mean score of the point forecast at hand, the recalibrated point forecast, and the functional T applied to the unconditional, marginal distribution of the outcome, respectively. If all quantities in (4.26) are finite, we refer to

$$\widehat{\mathsf{MCB}}_{\mathrm{S}} = \widehat{\mathrm{S}} - \widehat{\mathrm{S}}_{\mathrm{rc}}, \quad \widehat{\mathsf{DSC}}_{\mathrm{S}} = \widehat{\mathrm{S}}_{\mathrm{mg}} - \widehat{\mathrm{S}}_{\mathrm{rc}}, \quad \text{and} \quad \widehat{\mathsf{UNC}}_{\mathrm{S}} = \widehat{\mathrm{S}}_{\mathrm{mg}} \tag{4.27}$$

as the *miscalibration*, *discrimination* and *uncertainty* components of the mean score $\widehat{\mathrm{S}}$. Our next result generalizes Theorem 1 of Dimitriadis et al. (2021) and decomposes the mean score $\widehat{\mathrm{S}}$ into a signed sum of nonnegative, readily interpretable components.

Theorem 4.30. *Suppose that the functional* T *satisfies the conditions in Assumption 4.19. Let the scoring function* S *be of the form* (4.17)*, suppose that* $\widehat{x}_1, \ldots \widehat{x}_n$ *originate from tuples* $(x_1, y_1), \ldots, (x_n, y_n)$ *via Algorithm 4.1, and let all terms in* (4.26) *be finite. Then*

$$\widehat{\mathrm{S}} = \widehat{\mathsf{MCB}}_{\mathrm{S}} - \widehat{\mathsf{DSC}}_{\mathrm{S}} + \widehat{\mathsf{UNC}}_{\mathrm{S}}, \tag{4.28}$$

where $\widehat{\mathsf{MCB}}_{\mathrm{S}} \geq 0$ *with equality if* $\widehat{x}_i = x_i$ *for* $i = 1, \ldots, n$*, and* $\widehat{\mathsf{DSC}}_{\mathrm{S}} \geq 0$ *with equality if* $\widehat{x}_i = \widehat{x}_0$ *for* $i = 1, \ldots, n$*.*
If S *is strictly consistent, then* $\widehat{\mathsf{MCB}}_{\mathrm{S}} = 0$ *only if* $\widehat{x}_i = x_i$ *for* $i = 1, \ldots, n$ *and* $\widehat{\mathsf{DSC}}_{\mathrm{S}} = 0$ *only if* $\widehat{x}_i = \widehat{x}_0$ *for* $i = 1, \ldots, n$*.*

Proof. Immediate from Theorem 4.28. □

Thus, CORP estimates of score components enjoy the same properties as the respective population quantities (Theorem 4.23, eq. (4.15)). This agreement is not to be taken for granted, as the nonnegativity of the estimated components cannot be guaranteed if approaches other than the T-PAV algorithm are used for recalibration (Dimitriadis et al., 2021, Supplementary Section S5).
Recently, the estimation of calibration error has seen a surge of interest in machine learning (Guo et al., 2017; Kuleshov et al., 2018; Kumar et al., 2019; Nixon et al., 2019; Roelofs et al., 2022). Under the natural assumption of isotonicity of the population recalibration function, $\widehat{\mathsf{MCB}}_{\mathrm{S}}$ is a consistent estimate of the population quantity $\mathsf{MCB}_{\mathrm{S}}$, with canonical loss functions being natural choices for S. As noted, it is often preferable to use the square root of the miscalibration component under squared error as a measure of calibration error that can be

interpreted on natural scales. Asymptotic distributions for our estimators depend on the functional T, the scoring function S, and regularity conditions. Large sample theory can leverage extant theory for nonparametric isotonic regression, as hinted at in the previous section, though score components might show distinct asymptotic behavior. Further development is beyond the scope of the present work and strongly encouraged.

In the remainder of the section, we assume that S is a canonical score and drop the subscript in the score components. If there is a constant $\widehat{c} \in \mathbb{R}$ such that the empirical measure in $(x_1 + \widehat{c}, y_1), \ldots, (x_n + \widehat{c}, y_n)$ is unconditionally T-calibrated, let

$$\widehat{\mathrm{S}}_{\mathrm{urc}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(x_i + \widehat{c}, y_i). \tag{4.29}$$

We then refer to

$$\widehat{\mathsf{MCB}}_{\mathsf{u}} = \widehat{\mathrm{S}} - \widehat{\mathrm{S}}_{\mathrm{urc}} \quad \text{and} \quad \widehat{\mathsf{MCB}}_{\mathsf{c}} = \widehat{\mathrm{S}}_{\mathrm{urc}} - \widehat{\mathrm{S}}_{\mathrm{rc}}$$

as the CORP *unconditional* and *conditional* miscalibration components of the mean canonical score, respectively. Under mild conditions, these estimates are nonnegative and share properties of the respective population quantities in Theorem 4.27.

Theorem 4.31. *Let the conditions of Theorem 4.30 hold, and let* S *be a canonical loss function for* T*. Suppose there is a constant* $\widehat{c} \in \mathbb{R}$ *such that the empirical measure in* $(x_1+\widehat{c}, y_1), \ldots, (x_n+\widehat{c}, y_n)$ *is unconditionally* T*-calibrated, and suppose that all terms in* (4.29) *are finite. Then*

$$\widehat{\mathsf{MCB}} = \widehat{\mathsf{MCB}}_{\mathsf{u}} + \widehat{\mathsf{MCB}}_{\mathsf{c}},$$

where $\widehat{\mathsf{MCB}}_{\mathsf{u}} \geq 0$ *and* $\widehat{\mathsf{MCB}}_{\mathsf{c}} \geq 0$*.*

Proof. Immediate from Theorems 4.26 and 4.28, and the trivial fact that the addition of a constant is a special case of an isotonic mapping. □

In the middle row of Figure 4.6, the extended CORP decomposition,

$$\widehat{\mathrm{S}} = \widehat{\mathsf{MCB}}_{\mathsf{u}} + \widehat{\mathsf{MCB}}_{\mathsf{c}} - \widehat{\mathsf{DSC}} + \widehat{\mathsf{UNC}}, \tag{4.30}$$

which estimates the population decomposition (4.22), is applied to the mean squared error (MSE). Likewise, the extended CORP decomposition of the canonical score for quantiles, i.e., the piecewise linear quantile score (QS) from Table 4.3, is shown in the bottom row. The top row concerns threshold calibration, and we report the standard CORP decomposition (4.28) of the Brier score (BS) from (4.20). While the assumptions of Theorem 4.30 are satisfied in this setting, the addition of the constant $\widehat{c}$ may yield forecast values outside the unit interval, whence we refrain from considering the refined decomposition in (4.30).

In this context, the distinction between out-of-sample forecast evaluation and in-sample model diagnostics is critical. When evaluating out-of-sample forecasts,

both unconditional and conditional miscalibration are relevant. In contrast, in-sample model fits frequently enforce unconditional calibration. For example, if we fit a regression model with intercept by minimizing the canonical loss for a functional T, Theorem 4.26 applied to the associated empirical measure guarantees in-sample unconditional T-calibration. As special cases, this line of reasoning yields classical results in ordinary least squares regression, and the partitioning inequalities of quantile regression in Theorem 3.4 of Koenker and Bassett (1978).

4.3.4 Skill scores and a universal coefficient of determination

Let us revisit the mean scores in (4.26) under the natural assumption that the terms in $\widehat{\mathrm{S}}$ and $\widehat{\mathrm{S}}_{\mathrm{mg}}$ are finite and that $\widehat{\mathrm{S}}_{\mathrm{mg}}$ is strictly positive. In out-of-sample forecast evaluation, the quantity

$$\widehat{\mathrm{S}}_{\mathrm{skill}} = 1 - \frac{\widehat{\mathrm{S}}}{\widehat{\mathrm{S}}_{\mathrm{mg}}} = \frac{\widehat{\mathrm{S}}_{\mathrm{mg}} - \widehat{\mathrm{S}}}{\widehat{\mathrm{S}}_{\mathrm{mg}}} = \frac{\widehat{\mathsf{DSC}}_{\mathrm{S}} - \widehat{\mathsf{MCB}}_{\mathrm{S}}}{\widehat{\mathsf{UNC}}_{\mathrm{S}}} \tag{4.31}$$

is known as *skill score* (Murphy and Epstein, 1989; Murphy, 1996; Gneiting and Raftery, 2007; Jolliffe and Stephenson, 2012) and may attain both positive and negative values. In particular, when $\mathrm{S}(x, y) = (x - y)^2$ is the canonical loss function for the mean functional, $\widehat{\mathrm{S}}_{\mathrm{skill}}$ coincides with the popular Nash-Sutcliffe model efficiency coefficient (NSE; Nash and Sutcliffe, 1970; Moriasi et al., 2007). A positive skill score indicates predictive performance better than the simplistic unconditional reference forecast, whereas a negative skill score suggests that we are better off using the simple reference forecast. Of course, it is possible, and frequently advisable, to base skill scores on reference standards that are more sophisticated than an unconditional, constant point forecast (Hyndman and Koehler, 2006).

In contrast, if the goal is in-sample model diagnostics, the quantity in (4.31) typically is nonnegative. As we demonstrate now, it constitutes a powerful generalization of the coefficient of determination, R^2, or variance explained in least squares regression, and its close cousin, the R^1-measure in quantile regression (Koenker and Machado, 1999). Specifically, we propose the use of

$$\mathrm{R}^* = \frac{\widehat{\mathsf{DSC}}_{\mathrm{S}} - \widehat{\mathsf{MCB}}_{\mathrm{S}}}{\widehat{\mathsf{UNC}}_{\mathrm{S}}}, \tag{4.32}$$

as a universal *coefficient of determination*. In practice, one takes S to be a canonical loss for the functional T at hand, and we drop the subscripts in this case. The classical R^2 measure arises when $\mathrm{S}(x, y) = (x - y)^2$ is the canonical squared error loss function for the mean functional, and the R^1 measure of Koenker and Machado (1999) emerges when $\mathrm{S}(x, y) = 2\,(\mathbb{1}\{x \geq y\} - \alpha)\,(x - y)$ is the canonical piecewise linear loss under the α-quantile functional. Of course, in the case $\alpha = \frac{1}{2}$ of the median, the piecewise linear loss reduces to the absolute error.

In Figure 4.7, we present a numerical example on the toy data from Figure 1 in Kvålseth (1985). The straight lines show the linear (ordinary least squares)

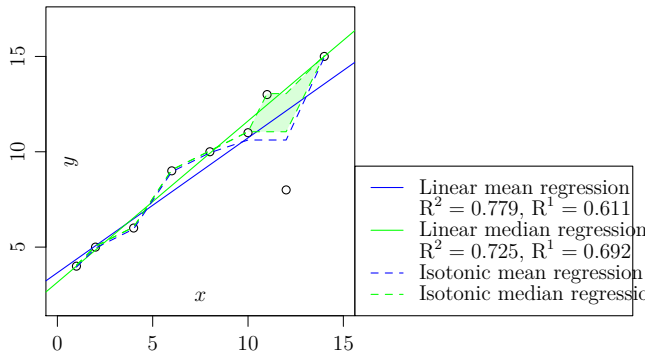


Figure 4.7 Linear mean and linear median regression lines for toy example from Kvålseth (1985, Figure 1), along with nonparametric isotonic mean and median regression fits. The isotonic median regression fit is not unique and framed by the respective lower and upper functional.

mean and linear (Laplace) median regression fits, which Kvålseth (1985) sought to compare. The piecewise linear broken curves illustrate the nonparametric isotonic regression fits, as realized by the T-PAV algorithm, where T is the mean and the lower and the upper median, respectively. As the linear regression fits induce the same ranking of the point forecasts, they yield the same PAV-recalibrated values that enter the terms in the score decomposition (4.27), and thus they have identical discrimination components in (4.28), which equal 10.593 under squared error and 2.333 under absolute error, regardless of which isotonic median is used. The uncertainty components, which equal 12.000 under squared error, and 2.889 under absolute error, are identical as well, since they depend on the observations only. Thus, the differences in R^2 respectively R^1 in Figure 4.7 stem from distinct miscalibration components. Of course, linear mean regression is preferred under squared error, and linear median regression is preferred under absolute error.
Various authors have discussed desiderata for a generally applicable definition of a coefficient of determination (Kvålseth, 1985; Nakagawa and Schielzeth, 2013) for the assessment of in-sample fit. In particular, such a coefficient ought to be dimensionless and take values in the unit interval, with a value of 1 indicating a perfect fit, and a value of 0 representing a complete lack of fit. The universal coefficient of determination R^* enjoys these properties under modest conditions.

Assumption 4.32. Suppose that the functional T is as stated in Assumption 4.19 with associated identification function V. Let the scoring function S be of the form (4.17), and suppose that $\widehat{x}_1, \dots \widehat{x}_n$ in (4.26) originate from tuples $(x_1, y_1), \dots, (x_n, y_n)$ via Algorithm 4.1. Furthermore, let the following hold.

(i) The terms contributing to $\widehat{\mathrm{S}}$ and $\widehat{\mathrm{S}}_{\mathrm{mg}}$ in (4.26) are finite, and $\widehat{\mathrm{S}}_{\mathrm{mg}} > 0$.

(ii) The values $x_1, \dots, x_n$ have been fitted to $y_1, \dots, y_n$ by in-sample empirical loss minimization with respect to S, with any constant fit $x_1 = \cdots = x_n$ being admissible.

For example, suppose that T is the mean functional and S is the canonical squared error scoring function. Then condition (i) is satisfied with the exception of the trivial case where $y_1 = \cdots = y_n$, and condition (ii) is satisfied under linear (ordinary least squares) mean regression with intercept. Similarly, if T is a quantile and S is the canonical piecewise linear loss function, then (i) is satisfied except when $y_1 = \cdots = y_n$, and (ii) is satisfied under linear quantile regression with intercept. In this light, the following theorem covers the classical settings for the R^2 and R^1 measures.

Theorem 4.33. *Under Assumption 4.32 it holds that*

$$\mathrm{R}^* \in [0, 1]$$

with $\mathrm{R}^* = 0$ *if* $x_i = \widehat{x}_0$ *for* $i = 1, \ldots, n$, *and* $\mathrm{R}^* = 1$ *if* $x_i = \mathrm{T}(\delta_i)$ *for* $i = 1, \ldots, n$.

Proof. The claim follows from Theorem 4.28, the trivial fact that a constant fit is a special case of an isotonic mapping, and the assumed form (4.17) of the scoring function. □

We emphasize that Assumption 4.32 and Theorem 4.33 are concerned with, and tailored to, in-sample model diagnostics. At the expense of technicalities, the regularity conditions can be relaxed, but the details are tedious and we leave them to subsequent work. The condition that any constant fit $x_1 = \cdots = x_n$ be admissible is critical and cannot be relaxed.

4.3.5 Empirical examples

We now illustrate the use of reliability diagrams, score decompositions, skill scores, and the coefficient of determination R^* for the purposes of forecast evaluation and model diagnostics.

In the basic setting of tuples $(x_1, y_1), \ldots, (x_n, y_n)$ of the form (4.24), the point forecast x_i represents the functional T of a posited distribution for y_i. The most prominent case of the mean functional and canonical squared error loss (4.19) is illustrated in Figure 4.2, where point forecasts by Tredennick et al. (2021) of (log-transformed) butterfly population size are assessed. The CORP mean reliability diagram along with 90% consistency bands under the hypothesis of mean calibration complements the scatter plot provided by Tredennick et al. (2021, Figure 6). With a mean squared error (MSE) of 0.224, ridge regression performs much better than the null model with an MSE of 0.262. The CORP score decomposition shown in Figure 4.2 refines and supports the analysis.

We move on to discuss the more complex setting of tuples $(F_1, y_1), \ldots, (F_n, y_n)$ of the form (4.24), where F_i is a posited distribution for y_i $(i = 1, \ldots, n)$. As discussed in Section 4.2 and visualized in Figure 4.1, the traditional unconditional notions of calibration, namely, probabilistic and marginal calibration, constitute weak forms of reliability. For this very reason, we recommend that checks for probabilistic and marginal calibration are given priority in this setting, much in

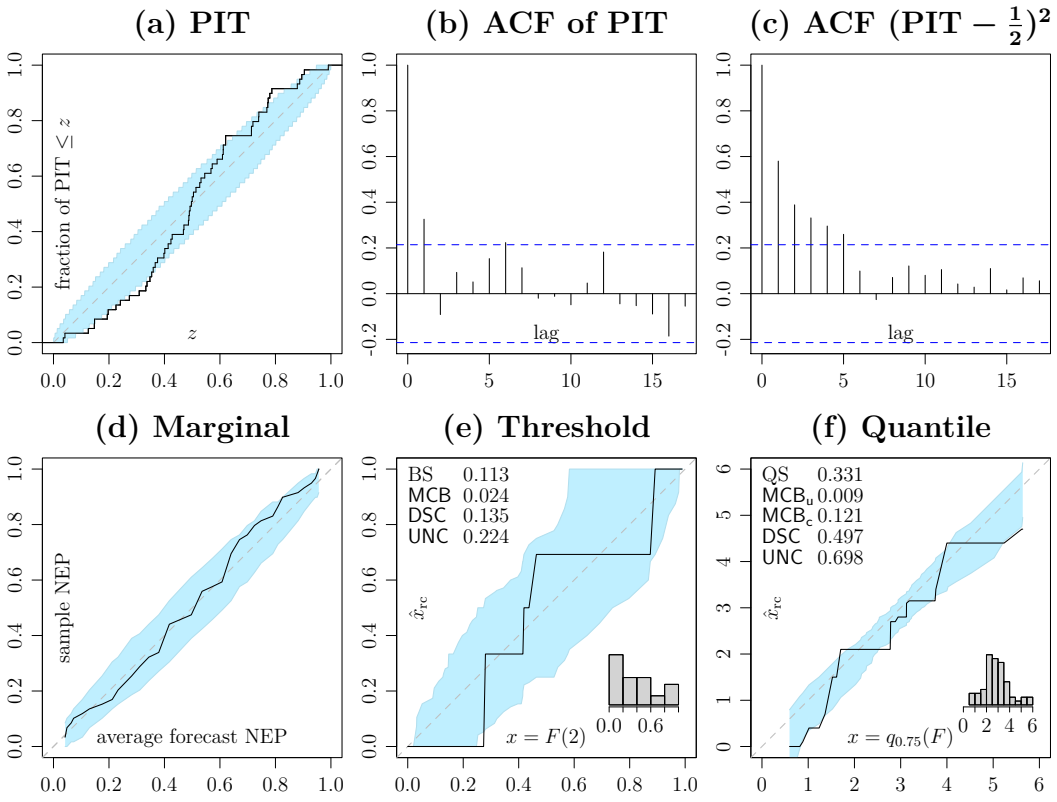


Figure 4.8 Calibration diagnostics for Bank of England forecasts of CPI inflation at a prediction horizon of one quarter: (a) PIT reliability diagram, along with the empirical autocorrelation functions of (b) original and (c) squared, centered PIT values, (d) marginal, (e) threshold, and (f) 75%-quantile reliability diagram. If applicable, we show 90% consistency bands and CORP score components under the associated canonical loss function, namely, the Brier score (BS) and the piecewise linear quantile score (QS), respectively.

line with current practice. Typically, probabilistic calibration is checked by plotting histograms of empirical probability integral transform (PIT) values (Diebold et al., 1998; Gneiting et al., 2007), though this practice is hindered by the need for binning. In Appendix 4.A.2.3, we discuss the *PIT reliability diagram*, a rarely used alternative that avoids binning and retains the spirit of our CORP approach by plotting the CDF of the empirical PIT values. Similarly, as we also discuss in Appendix 4.A.2.3, the *marginal reliability diagram* can be used to assess marginal calibration in the spirit of the CORP approach. If the analysis indicates gross violations of probabilistic or marginal calibration, we note from Section 4.2 and Figure 4.1 that key notions of conditional calibration must be violated as well. Otherwise, we might proceed to check stronger conditional notions of calibration, such as threshold, mean, and quantile calibration.

To illustrate this process, we consider quarterly Bank of England forecasts of consumer price index (CPI) inflation rates, as issued since 2004. The forecast distributions, for which we give details and refer to extant analyses in Appendix

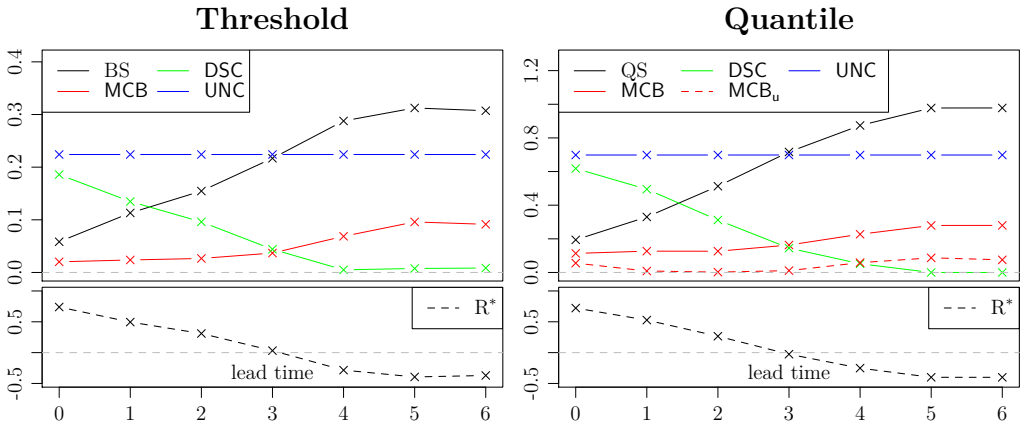


Figure 4.9 Score decomposition (4.28) respectively (4.30) and skill score (4.31) for probability forecasts of not exceeding the 2% inflation target (left) and 75%-quantile forecasts (right) induced by Bank of England fan charts for CPI inflation, under the associated canonical scoring function.

4.A.3.2, are two-piece normal distributions that are communicated to the public via fan charts. The forecasts are at prediction horizons up to six quarters ahead in the time series setting, where k step ahead forecasts that are ideal with respect to the canonical filtration show PIT values that are independent at lags $\geq k+1$ in addition to being uniformly distributed (Diebold et al., 1998). However, as discussed in Appendix 4.A.3.1, independent, uniformly distributed PIT values do not imply auto-calibration, except in a special case. Thus, calibration diagnostics beyond checks of the uniformity and independence of the PIT are warranted.
In Figure 4.8, we consider forecasts one quarter ahead and show PIT and marginal reliability diagrams, along with empirical autocorrelation functions (ACFs) for the first two moments of the PIT. In part, the PIT reliability diagram and the ACFs lie outside the respective 90% consistency bands. For a closer look, we also plot the threshold reliability diagram at the policy target of 2% and the lower α-quantile reliability diagram for $\alpha = 0.75$. The deviations from reliability remain minor, in stark contrast to calibration diagnostics at prediction horizons $k \geq 4$, for which we refer to Appendix 4.A.3.2.
Figure 4.9 shows the standard CORP decomposition (4.28) of the Brier score (BS) for the induced probability forecasts at the 2% target and the extended CORP decomposition (4.30) of the piecewise linear quantile score for α-quantile forecasts at level $\alpha = 0.75$ and lead times up to six quarters ahead. In the latter case, the difference between MCB and $\mathsf{MCB_u}$ equals the $\mathsf{MCB_c}$ component. Generally, the miscalibration components increase while the discrimination components decrease with the lead time. Related results for the quantile functional can be found in Pohle (2020, Table 5, Figures 7 and 8), where there is a notable increase in the discrimination (resolution) component at the largest two lead times, which is caused by counterintuitive decays in the recalibration functions. In contrast, the regularizing constraint of isotonicity prevents overfitting in the CORP approach.

The coefficient of determination or skill score R^* decays with the prediction horizon and becomes negative at lead times $k \geq 4$. This observation suggests that forecasts remain informative at lead times up to at most three quarters ahead, in line with the substantive findings in Pohle (2020) and other extant work, as hinted at in Appendix 4.A.3.2.

4.4 Model Diagnostics and Forecast Evaluation for Quantiles

In Gneiting et al. (2023), we discuss the ubiquitous special case of quantile forecasts. In this section, I provide a brief overview focusing on peculiarities particular to quantile forecasts along with a simple data example (from Gneiting et al., 2023, Section 4.1), which illustrates in-sample and out-of-sample evaluation when fitting multiple quantiles. Additionally, I discuss a test for simultaneous unconditional quantile calibration used in a study by Colonna et al. (2022) and other studies focusing on the validation of expert opinions. The test is simply a multinomial test and thus the discussion nicely complements Chapter 3.

4.4.1 Evaluation of quantile forecasts and models

In Gneiting et al. (2023), we adapt the theory presented so far in this chapter (i.e., the theory from Gneiting and Resin, 2021) to the particular case of quantiles. Quantiles are important functionals frequently addressed in statistical modeling (see Gneiting et al., 2023, and references therein). As probability distributions are characterized by their quantile function, a finite number of posited quantiles at suitable quantile levels may convey sufficient information about the conditional distribution of the observation. Thus, it is common practice to fit or predict the (conditional) quantiles of a real-valued outcome of interest at multiple levels.
Quantiles are an important family of identifiable functionals, to which the CORP approach applies. Beyond CORP reliability diagrams and score decompositions, unconditional quantile calibration may be assessed using coverage plots. In the case of a stand-alone quantile prediction X at level $\alpha \in (0, 1)$, the general unconditional T-calibration condition (4.13) reduces to a classical non-exceedance criterion,

$$\mathbb{P}(Y \leq X) \geq \alpha \quad \text{and} \quad \mathbb{P}(Y \geq X) \leq 1 - \alpha,$$

that takes discreteness into account (Gneiting et al., 2023, Eq. 1.) similar to the traditional unconditional coverage condition (4.10) for probabilistic forecasts. In practice, this non-exceedance criterion can be validated by considering the empirical coverage of the predictions. For data of the form (4.24), *lower* and *upper coverage* are given by

$$c_\alpha^- = \frac{1}{n} \sum_i \mathbb{1}\{y_i < x_i\} \quad \text{and} \quad c_\alpha^+ = \frac{1}{n} \sum_i \mathbb{1}\{y_i \leq x_i\},$$

respectively. Given a forecast specifying multiple quantiles, lower and upper coverage are plotted against the respective quantile levels in a coverage plot. Subject to unconditional quantile calibration, lower and upper coverage typically nest the quantile level α. Hence, lower coverage frequently lies below the diagonal in a coverage plot, whereas upper coverage tends to lie above it. When these tendencies are not observed empirically, consistency intervals illustrate whether the observed fluctuations might be a sign of unconditional miscalibration. In a nutshell, a consistency interval shows the critical values of two one-sided binomial tests with Bonferroni correction as briefly alluded to in Appendix 4.A.2.4 and described in detail in Gneiting et al. (2023). The hypothesis of unconditional quantile calibration at a fixed level is to be scrutinized if either lower coverage exceeds the upper critical value or upper coverage falls below the lower critical value, as observed in the coverage plot for the isotonic out-of-sample fit to Engel's food expenditure data at the lowest and highest quantile levels in Figure 4.11 of Section 4.4.3.

While graphical displays such as coverage plots and reliability diagrams are important diagnostic tools, comparative forecast evaluation calls for the use of consistent scoring functions. Virtually all consistent scoring functions for the α-quantile ($\alpha \in (0, 1)$) are of the generalized piecewise linear (GPL) form, i.e., a consistent scoring function can be written as

$$\mathrm{S}(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)(g(x) - g(y)),$$

where g is a nondecreasing function (Gneiting, 2011a). The scoring function S is strictly consistent if the function g is strictly increasing. The *asymmetric piecewise linear* or *pinball loss*

$$\mathrm{S}(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)(x - y) \tag{4.33}$$

arises when g is the identity function.[3] The elementary losses (4.16) can be used to visualize forecast dominance in Murphy diagrams (Ehm et al., 2016). Note that Ehm et al. (2016) and Gneiting et al. (2023) use elementary loss functions that differ from the ones presented in (4.16). The Murphy diagram compares the Murphy curve of multiple forecasts, i.e., it shows the mean elementary loss of each forecast at each index η. If a forecast's Murphy curve lies entirely below the Murphy curve of a competing forecast, said forecast dominates the competing forecast as it attains a better mean score under virtually all consistent scoring functions. Therefore, a dominant forecast should be preferred by any rational forecast user.

4.4.2 Simultaneous unconditional quantile calibration

"Cooke's classical model" has been used by several authors to assess quantile forecasts typically framed as expert opinions (see Colonna et al., 2022, and references therein). In a nutshell, Cooke's approach uses a p-value to quantify to which

[3]The scoring function (4.33) is a version of the canonical quantile loss function without the unimportant additional factor 2, cf. Table 4.3.

degree the hypothesis of a forecast (or "expert judgment") being calibrated (or "statistically accurate") is supported by the observations (Colonna et al., 2022, Supplementary Information). The p-value is computed from a log-likelihood ratio test as follows. A quantile forecast $X = (X_{\alpha_j})$ at $k \in \mathbb{N}$ levels $\alpha_1 < \alpha_2 < \cdots < \alpha_k$ naturally divides the real line into $k+1$ inter-quantile intervals $I_1(X) = (-\infty, X_1], I_2(X) = (X_1, X_2], \ldots, I_k(X) = (X_{k-1}, X_k], I_{k+1}(X) = (X_k, \infty)$. The test evaluates whether the observed frequencies $(\#\{y_i \in I_j(x_i)\})_{j=1}^k$ of observations falling within each of the respective inter-quantile intervals agree with the hypothesized multinomial probabilities $\pi = (\alpha_1, \alpha_2 - \alpha_1, \ldots, \alpha_k - \alpha_{k-1}, 1 - \alpha_k)$, i.e., the null hypothesis $\mathbb{Q}(Y \in I_j(X)) = \pi_j$ for $j = 1, \ldots, k$. To this end, Cooke's approach uses the log-likelihood ratio statistic and derives the p-value from the test statistic's asymptotic chi-square distribution (Colonna et al., 2022, Supplementary Information).

Notably, Cooke's approach employs a simple multinomial test and other multinomial tests (e.g., tests based on the chi-square or probability mass statistic, see Section 3.2) may be used instead of the asymptotic log-likelihood ratio test. When the number k of quantile levels is small, exact tests may be feasible using the algorithm proposed in Chapter 3, as illustrated in the data example of Section 4.4.3. Especially when the number of observations n is low, exact tests may provide more accurate p-values than their asymptotic approximate counterparts. The above null hypothesis is equivalent to simultaneous unconditional α-quantile calibration at all levels $\alpha = \alpha_1, \ldots, \alpha_k$ if the underlying (conditional) distributions are continuous and hence the probability of the observation matching a predicted quantile, $\mathbb{P}(X_{\alpha_j} = Y)$, is zero for all quantile levels α_j. However, the test does not account for discreteness, and the null hypothesis may fail to coincide with the hypothesis of unconditional quantile calibration if lower and upper coverage cannot be expected to match. In such cases, p-values tend to be very small even for unconditionally calibrated quantile forecasts and hence uninformative or even misleading, as is the case for the in-sample isotonic quantile regression fits in the following data example.

Preliminary results presented in the next section suggest that the test can be adapted to the discrete case by assigning observations matching a predicted quantile dynamically to either adjacent interval in a way that maximizes the p-value. One might call such an assignment a 'most favorable configuration'. Such a configuration balances discrepancies between lower and upper coverage. Since upper and lower coverage nest the quantile level in expectation under unconditional calibration, this adaptation results in a plausible configuration that is not rejected by the multinomial test with high probability. The most favorable configuration might yield a conservative test for the hypothesis of simultaneous unconditional quantile calibration at all quantile levels as p-values might be somewhat inflated, although this effect does not seem to be an issue with the out-of-sample isotonic quantile regression fits in the following data example. A thorough examination of the proposed remedy is beyond the scope of the present discussion, yet I see the investigation of this adaptation as an interesting avenue for future research.

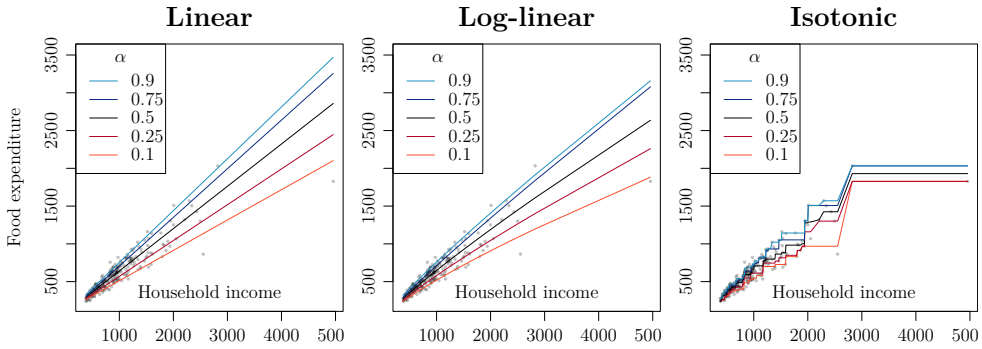


Figure 4.10 Linear, log-linear, and isotonic quantile regression fits for Engel (1857) food expenditure data against household income.

4.4.3 Engel's food expenditure data: In-sample regression diagnostics versus out-of-sample forecast evaluation

This section is adapted from the Annual Review of Statistics and Its Application, Volume 10; copyright 2023 Annual Reviews, `https://www.annualreviews.org`.
We consider quantile regression fits for the classical food expenditure data from Engel (1857) for 19th century European working-class households, as also discussed by Koenker (2005, pp. 78, 297–307). Engel's conclusion that the share of income that is used for food expenditure decreases with income is known as Engel's law and stands in today's work on poverty and especially poverty reduction as one of the most enduring relationships in economics (Blundell et al., 2007). Modeling conditional quantiles for a range of levels ($\alpha = 0.1, 0.25, 0.5, 0.75, 0.9$), instead of the conditional mean, allows a comprehensive assessment of Engel's law. We compare standard (linear) quantile regression, linear quantile regression on log-transformed values (Koenker, 2005, p. 78) and nonparametric isotonic quantile regression (Wright, 1984) both in-sample and out-of-sample, using leave-one-out cross-validation, based on Engel's data of size $n = 235$.
We first fit a standard quantile regression model of food expenditure on income. The parametric form imposes a linear relationship, which in view of Engel's law is too restrictive. Following Koenker (2005, p. 78), we also use a linear model of the log-transformed quantities, where slope values smaller than one support Engel's law, for $\log(y) = \beta_1 + \beta_2 \log(x)$ is equivalent to $y = \exp(\beta_1)x^{\beta_2}$ so that $\beta_2 < 1$ implies a concave relationship. Indeed, we find estimated slope coefficients between 0.80 and 0.92. Finally, we use isotonic quantile regression as a fully flexible nonparametric method. Evidently, the isotonicity assumption is satisfied. Figure 4.10 shows the in-sample model fit for the three methods and five quantile levels. The log-linear model fits show a slightly concave shape, as can be expected by Engel's law, which is confirmed by the nonparametric isotonic estimates.
In Figure 4.11, Figure 4.12 and Table 4.5, we contrast in-sample model diagnostics and out-of-sample (leave-one-out cross-validation) forecast evaluation for

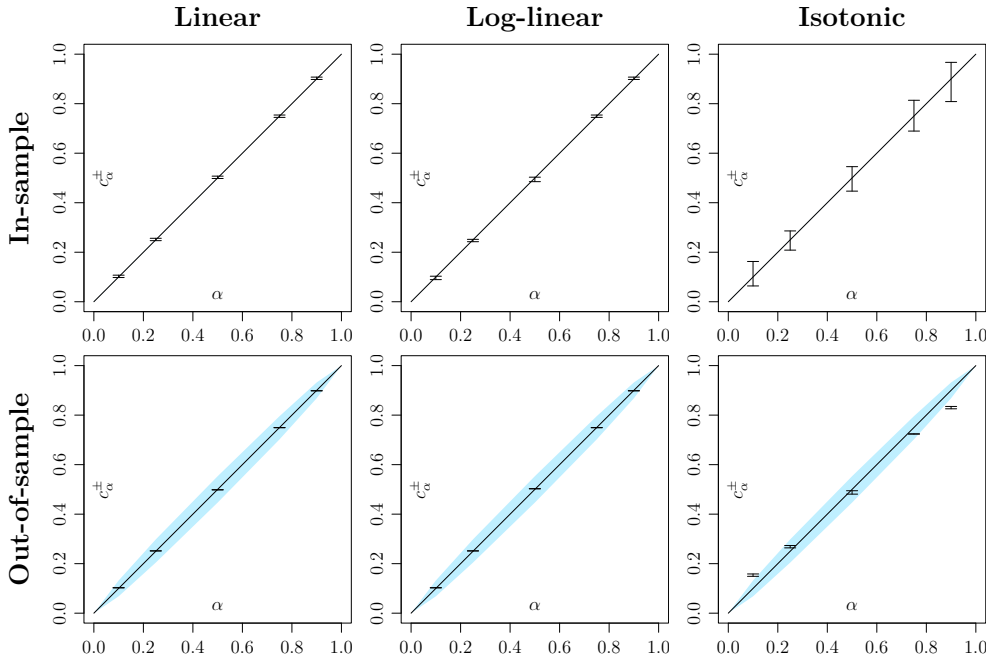


Figure 4.11 In-sample (top row) and out-of-sample (bottom row) coverage plots, depicting the intervals $[c_\alpha^-, c_\alpha^+]$ along with 90% consistency bands, for quantile regression fits to Engel (1857) food expenditure data.

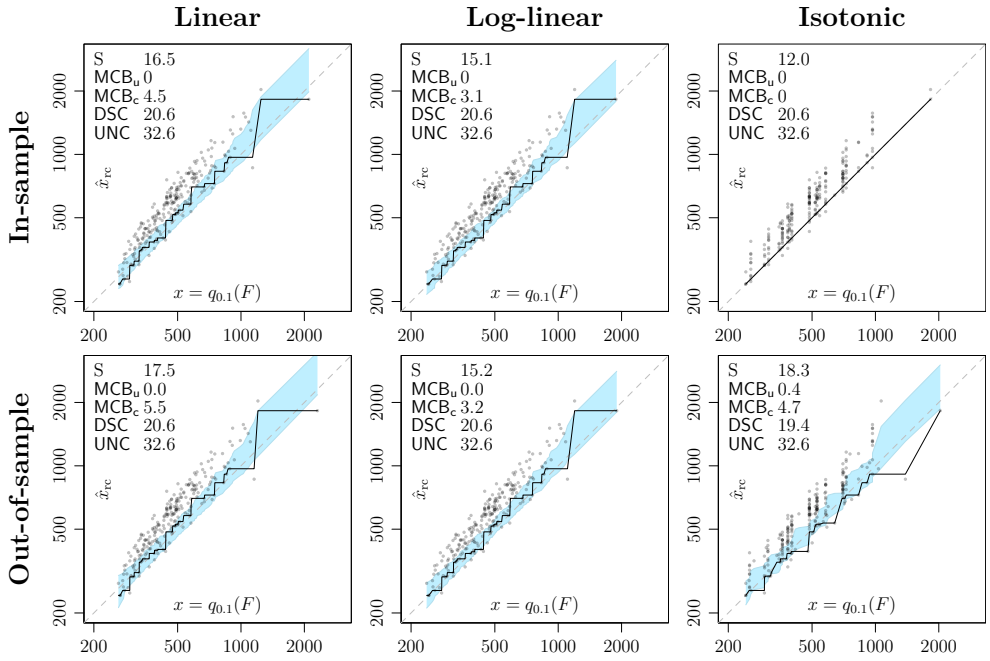


Figure 4.12 In-sample (top row) and out-of-sample (bottom row) α-quantile reliability diagrams ($\alpha = 0.1$), along with 90% consistency bands, for quantile regression fits to Engel (1857) data.

Table 4.5 In-sample and out-of-sample CORP components of the mean pinball loss $\widehat{\mathrm{S}}$ for α-quantile regression fits to Engel (1857) food expenditure data.

Level (UNC)	Components	In-sample			Out-of-sample		
		Linear	Log-linear	Isotonic	Linear	Log-linear	Isotonic
$\alpha = 0.1$ (32.6)	$\widehat{\mathrm{S}}$	16.5	15.1	12.0	17.5	15.2	18.3
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.4
	$\mathsf{MCB_c}$	4.5	3.1	0.0	5.5	3.2	4.7
	DSC	20.6	20.6	20.6	20.6	20.6	19.4
$\alpha = 0.25$ (67.6)	$\widehat{\mathrm{S}}$	30.1	29.2	23.0	30.6	29.8	30.8
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.1
	$\mathsf{MCB_c}$	7.1	6.2	0.0	7.3	6.6	4.4
	DSC	44.6	44.6	44.6	44.3	44.4	41.3
$\alpha = 0.5$ (98.5)	$\widehat{\mathrm{S}}$	37.4	36.6	28.5	38.0	37.3	37.8
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathsf{MCB_c}$	8.9	8.1	0.0	8.9	8.1	5.5
	DSC	70.0	70.0	70.0	69.4	69.3	66.2
$\alpha = 0.75$ (91.6)	$\widehat{\mathrm{S}}$	27.9	27.6	21.0	28.6	28.4	30.0
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathsf{MCB_c}$	6.9	6.6	0.0	6.9	6.6	5.2
	DSC	70.6	70.6	70.6	69.9	69.8	66.8
$\alpha = 0.9$ (61.3)	$\widehat{\mathrm{S}}$	14.4	14.4	10.2	15.0	14.9	17.8
	$\mathsf{MCB_u}$	0.0	0.0	0.0	0.0	0.0	0.6
	$\mathsf{MCB_c}$	4.2	4.2	0.0	4.4	4.3	5.4
	DSC	51.1	51.1	51.1	50.7	50.7	49.5

the three methods.[4] Perfect in-sample coverage is guaranteed by the partitioning inequalities of quantile regression. Similarly, isotonic regression fits show perfect in-sample unconditional and conditional calibration by construction, with reliability diagrams that are constrained to the diagonal. While the linear and log-linear models retain good coverage out-of-sample, unconditional and conditional calibration deteriorate notably for the isotonic model.

Table 4.5 shows the mean pinball loss computed from (4.33) along with the CORP decomposition. In-sample, all three methods show perfect unconditional calibration with a vanishing $\mathsf{MCB_u}$ component, and they also share the DSC component, for they are isotonic transformations of each other. The $\mathsf{MCB_c}$ component vanishes for the isotonic model and is slightly better for the log-linear than for the linear model. The nonparametric isotonic regression technique is prone to overfitting in small samples, which results in the best scores in-sample but worse scores out-of-sample when compared to the parametric linear and log-linear models. Interestingly, isotonic regression also has the worst out-of-sample DSC components. Both in-sample and out-of-sample, the log-linear model has slightly better scores

[4]To generate consistency bands in the reliability diagrams, we resample residuals of log-transformed values, which seems natural here and in other applications with strictly positive data, where variability increases as observed values increase.

Table 4.6 Exact p-values obtained from the unconditional calibration test (based on the log-likelihood ratio statistic) detailed in Section 4.4.2 using both Cooke's classical approach and the proposed adaption for in-sample and out-of-sample α-quantile regression fits to Engel (1857) food expenditure data.

	In-sample			Out-of-sample		
Approach	Linear	Log-linear	Isotonic	Linear	Log-linear	Isotonic
Classical	0.9988	0.9997	0.0004	1.0000	1.0000	0.0006
Proposed	1.0000	1.0000	1.0000	1.0000	1.0000	0.0015

than the linear model, providing additional support for Engel's law.
Table 4.6 contrasts Cooke's classical approach detailed in Section 4.4.2 with the proposed adaption using a most favorable configuration. As hinted at in the preceding section, Cooke's approach does not cope well with posited quantiles matching the respective observations. This issue is strikingly apparent with the in-sample isotonic regression fits, which are perfectly calibrated despite the minuscule p-value assigned by the unconditional calibration test. If the p-value of a favorable configuration is considered, the issue disappears, while the p-value of the miscalibrated out-of-sample fits remains small. Notably, the out-of-sample linear and log-linear fits show nearly perfect coverage, as evidenced by p-values of almost 1. This unexpected behavior appears to be due to the leave-one-out cross-validation scheme and the linear nature of the respective models as observations that lie above the in-sample fit end up lying above the out-of-sample fit, while observations below the in-sample fit end up below the out-of-sample fit.

4.5 Discussion

We have developed a comprehensive theoretical and methodological framework for the analysis of calibration and reliability, serving the purposes of both (out-of-sample) forecast evaluation and (in-sample) model diagnostics. A common principle is that fitted or predicted distributions ought to be calibrated or reliable, ideally in the sense of auto-calibration, which stipulates that the outcomes are random draws from the posited distributions. For general real-valued outcomes, we have seen that auto-calibration is stronger than both classical unconditional and recently proposed conditional notions of calibration. We have developed hierarchies of calibration in the spirit of Van Calster et al. (2016), as highlighted in Figure 4.1, and proposed a generic notion of conditional calibration in terms of statistical functionals. Specifically, a posited distribution is conditionally T-calibrated if the induced point forecast for the functional T can be taken at face value. This concept continues to apply when stand-alone point forecasts or regression output in terms of the functional T are to be evaluated and can be assessed via T-reliability diagrams and associated score decompositions. Importantly, our tools apply regardless of how forecasts are generated, be it through the use of

traditional statistical regression models, modern machine learning techniques, or even subjective human judgment.

We have adopted and generalized the nonparametric approach of Dimitriadis et al. (2021), who obtained consistent, optimally binned, reproducible, and PAV based (CORP) estimators of T-reliability diagrams and score components in the case of probability forecasts for binary outcomes. While our tools apply in the much broader setting of identifiable functionals and real-valued outcomes, the arguments put forth by Dimitriadis et al. (2021) continue to apply, in that CORP estimators are bound to, simultaneously, improve statistical efficiency, reproducibility (Stodden et al., 2016), and stability (Yu and Kumbier, 2020). In a nutshell, the CORP approach is flexible, due to its use of nonparametric regression for recalibration, and yet it avoids overfitting, owing to the regularizing constraint of isotonicity. Notably, the CORP score decomposition yields a new, universal coefficient of determination, R^*, that nests and generalizes the classical R^2 in ordinary least squares (mean) regression, and its cousin R^1 in quantile regression. In independent work, Allen (2021) also observes the link between skill scores, score decompositions, and the coefficient of determination. We have illustrated the CORP approach on Bank of England forecasts of inflation, along with a brief ecological example. Furthermore, a brief review of the particular case of forecasts in the form of (one or multiple) quantiles, accompanied by a case study, was provided. Code in R (R Core Team, 2022) for reproducing our results is available (Resin, 2021b; Wolffram et al., 2022).

Follow-up work on the CORP approach for specific functionals T is essential, including but not limited to the ubiquitous cases of quantiles and the mean functional, where the newly developed tools can supplement classical approaches to regression diagnostics, as hinted at in the ecological example. In particular, we have applied a crude, all-purpose, residual-based permutation approach to generate consistency bands for T-reliability diagrams under the hypothesis of T-calibration. Clearly, this approach can be refined, and we anticipate vigorous work on consistency and confidence bands, based on either resampling or large sample theory, akin to the developments in Dimitriadis et al. (2021) for probability forecasts of binary outcomes. Similarly, CORP estimates of miscalibration components under canonical loss functions are natural candidates for the quantification of calibration error in empirical work. Reliability and discrimination ability are complementary attributes of point forecasts and regression output, and discrimination can be assessed quantitatively via the respective score component. When many forecasts are to be compared with each other, scatter plots of CORP miscalibration (MCB) and discrimination (DSC) components admit succinct visual displays of predictive performance. In this type of display, forecasts with the same score or, equivalently, identical coefficient of determination, R^*, gather on lines with unit slope, and multiple facets of forecast quality can be assessed simultaneously, for a general alternative to the widely used Taylor (2001) diagram.

Formal tests of hypotheses of calibration are critical in both specific applications, such as banking regulation (e.g., Nolde and Ziegel, 2017), and in generic tasks,

such as the assessment of goodness-of-fit in regression (Dimitriadis et al., 2021, Section S2). In Appendix 4.A.2.4, we comment on this problem from the perspective of the theoretical and methodological advances presented here. While specific developments need to be deferred to future work, it is our belief that the progress in our understanding of notions and hierarchies of calibration, paired with the CORP approach to estimating reliability diagrams and score components, can spur a wealth of new and fruitful developments in these directions.

4.A Appendix

4.A.1 Supporting calculations for Section 4.2

Here, we provide supporting computations and discussion for Examples 4.2 and 4.4, Definitions 4.9 and 4.24, Figures 4.4 and 4.5, and Table 4.4, along with a discussion of the relation between probabilistic calibration and unconditional quantile calibration, and a counterexample hinted at in the main text. For subsequent use, the first three (non-centered) moments of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ are μ, $\mu^2 + \sigma^2$, and $\mu^3 + 3\mu\sigma^2$. As in the main text, we let φ and Φ denote the density and the cumulative distribution function (CDF), respectively, of a standard normal variable.

4.A.1.1 Unfocused forecast

For fixed $a, b \in \mathbb{R}$, the function

$$y \mapsto \Phi_a(y - b) = \frac{1}{2}\left(\Phi(y - b) + \Phi(y - a - b)\right)$$

is a CDF. The random CDF (4.2) for the unfocused forecast in Example 4.2 can be written as $F(y) = \Phi_\eta(y - \mu)$, where η and μ are independent random variables and $\eta = \pm\eta_0$ for some constant $\eta_0 > 0$. Then the conditional CDF for the outcome Y given the posited (non) exceedance probability $F(t)$ at any fixed threshold $t \in \mathbb{R}$ or, equivalently, given the quantile forecast $F^{-1}(\alpha)$ at any fixed level $\alpha \in (0, 1)$ is

$$\begin{aligned}\mathbb{P}(Y \le y \mid F(t) = \alpha) &= \mathbb{P}(Y \le y \mid F^{-1}(\alpha) = t) = \mathbb{P}(Y \le y \mid \mu = t - \Phi_\eta^{-1}(\alpha)) \\ &= \frac{1}{\sum_{s=\pm1} \varphi(t - \Phi_{s\eta_0}^{-1}(\alpha))} \sum_{s=\pm1} \varphi(t - \Phi_{s\eta_0}^{-1}(\alpha))\, \Phi(y - (t - \Phi_{s\eta_0}^{-1}(\alpha))).\end{aligned}$$

As F is symmetric, conditioning on the mean is the same as conditioning on the median. The second moment is $m_2(F) = 1 + \mu^2 + \mu\eta + \frac{1}{2}\eta^2 \ge 1 + \frac{1}{4}\eta^2$, so that

$$\mathbb{P}(Y \le y \mid m_2(F) = m) = \mathbb{P}\left(Y \le y \mid \mu = -\frac{1}{2}\eta \pm \sqrt{m - 1 - \frac{1}{4}\eta^2}\right)$$

is a mixture of normal distributions. Similarly, the third moment is $m_3(F) = \mu^3 + \frac{3}{2}\eta\mu^2 + 3\left(\frac{1}{2}\eta^2 + 1\right)\mu + \frac{1}{2}\eta\left(\eta^2 + 3\right) = f(\mu; \eta)$, so that $\mathbb{P}(Y \le y \mid m_3(F) = m) = \mathbb{P}(Y \le y \mid f(\mu; \eta) = m)$ also is a mixture of normal distributions. To compute the roots of the mapping $x \mapsto f(x; \eta)$, we fix η at $\pm\eta_0$ and use a numeric solver (`polyroot` in R).

As regards the score decomposition (4.15) with $\mathrm{S}(x, y) = (x - y)^2$ for the implied mean-forecast, $m_1(F) = \mu + \frac{1}{2}\eta$, the expected score of the recalibrated mean-forecast is

$$\bar{\mathrm{S}}_{\mathrm{rc}} = \mathbb{E}\left[\frac{\sum_{s=\pm1} \varphi\left(m_1(F) + \frac{s}{2}\eta_0\right)\left(m_1(F) + \frac{s}{2}\eta_0\right)}{\sum_{s=\pm1} \varphi\left(m_1(F) + \frac{s}{2}\eta_0\right)} - Y\right]^2$$

$$= \mathbb{E}\left[\frac{\sum_{s=\pm1}\varphi\big(\mu+\frac{1}{2}\eta+\frac{s}{2}\eta_0\big)\left(\frac{1}{2}\eta+\frac{s}{2}\eta_0\right)}{\sum_{s=\pm1}\varphi\left(\mu+\frac{1}{2}\eta+\frac{s}{2}\eta_0\right)} - (Y-\mu)\right]^2$$
$$= \eta_0^2\,\mathbb{E}\left[\frac{\varphi(\mu+\eta)}{\varphi(\mu)+\varphi(\mu+\eta)}\right]^2 + \mathbb{E}[Y-\mu]^2 = \eta_0^2\,\mathbb{E}\big[\Psi_{\eta_0}^2(\mu)\big] + 1,$$

where we define $\Psi_a(x) = \varphi(x+a)/(\varphi(x)+\varphi(x+a))$ for $a \in \mathbb{R}$ and note that $\mathbb{E}[\Psi_\eta^2(\mu) \mid \eta] = \mathbb{E}[\Psi_{\eta_0}^2(\mu)]$. The associated integral

$$I(\eta_0) = \mathbb{E}\big[\Psi_{\eta_0}^2(\mu)\big] = \int_{-\infty}^{\infty}\left(\frac{\varphi(x+\eta_0)}{\varphi(x)+\varphi(x+\eta_0)}\right)^2\varphi(x)\,\mathrm{d}x$$

needs to be evaluated numerically.

4.A.1.2 Lopsided forecast

We proceed in analogy to the development for the unfocused forecast. For fixed $a \in [0,1]$ and $b \in \mathbb{R}$, the function

$$y \mapsto \Phi_a(y-b) = (1-a)\Phi(y-b)\mathbb{1}\{y \le b\} + ((1+a)\Phi(y-b)-a)\,\mathbb{1}\{y > b\}$$

is a CDF. The CDF for the lopsided forecast with random density (4.3) from Example 4.2 can be written as $F(y) = \Phi_\delta(y-\mu)$, where δ and μ are independent random variables and $\delta = \pm\delta_0$ for some $\delta_0 \in (0,1)$. As $\mathbb{E}[\Phi_\delta(y-\mu) \mid \mu] = \Phi(y-\mu)$, the lopsided forecast is marginally calibrated. It fails to be probabilistically calibrated since $Z_F = \Phi_\delta(Y-\mu)$ has CDF

$$\mathbb{P}(Z_F \le u) = \frac{1}{2}\sum_{s=\pm1}\left(\frac{u}{1-s\delta_0}\mathbb{1}\left\{\frac{u}{1-s\delta_0} \le \frac{1}{2}\right\} + \frac{u+s\delta_0}{1+s\delta_0}\mathbb{1}\left\{\frac{u}{1-s\delta_0} > \frac{1}{2}\right\}\right)$$

for $u \in (0,1)$ by the law of total probability.
The conditional CDF for the outcome Y given the posited (non) exceedance probability $F(t)$ at any fixed threshold $t \in \mathbb{R}$ or, equivalently, given the quantile forecast $F^{-1}(\alpha)$ at any fixed level $\alpha \in (0,1)$ is

$$\mathbb{P}(Y \le y \mid F(t) = \alpha) = \mathbb{P}(Y \le y \mid F^{-1}(\alpha) = t) = \mathbb{P}(Y \le y \mid \mu = t - \Phi_\delta^{-1}(\alpha))$$
$$= \frac{1}{\sum_{s=\pm1}\varphi(t-\Phi_{s\delta_0}^{-1}(\alpha))}\sum_{s=\pm1}\varphi(t-\Phi_{s\delta_0}^{-1}(\alpha))\,\Phi(y-(t-\Phi_{s\delta_0}^{-1}(\alpha))),$$

where $\Phi_a^{-1}(\alpha) = \Phi^{-1}(\alpha/(1-a))$ if $\alpha \le \frac{1}{2}(1-a)$ and $\Phi_a^{-1}(\alpha) = \Phi^{-1}((a+\alpha)/(a+1))$ otherwise.
As F is a mixture of truncated normal distributions, its moments are mixtures of the component moments, for which we refer to Orjebin (2014). The first moment is $m_1(F) = \mu + 2\delta\varphi(0)$, so that

$$\mathbb{P}(Y \le y \mid m_1(F) = m) = \mathbb{P}(Y \le y \mid \mu = m - 2\delta\varphi(0))$$
$$= \frac{1}{\sum_{s=\pm1}\varphi(m-2s\delta_0\varphi(0))}\sum_{s=\pm1}\varphi(m-2s\delta_0\varphi(0))\,\Phi(y-(m-2s\delta_0\varphi(0)))$$

is a mixture of normal distributions. Similarly, the second and third moments are $m_2(F) = \mu^2+1+4\delta\varphi(0)\mu \geq 1-4\delta^2\varphi(0)^2$ and $m_3(F) = \mu^3+3\mu+2\delta\varphi(0)(3\mu^2+2) = f(\mu;\delta)$, respectively, so that

$$\mathbb{P}(Y \leq y \mid m_2(F) = m) = \mathbb{P}(Y \leq y \mid \mu = -2\delta\varphi(0) \pm \sqrt{4\delta^2\varphi(0)^2 - 1 + m}),$$
$$\mathbb{P}(Y \leq y \mid m_3(F) = m) = \mathbb{P}(Y \leq y \mid f(\mu;\delta) = m)$$

also admit expressions as mixtures of normal distributions. Again, we use a numeric solver to find the roots of $x \mapsto f(x;\pm\delta_0)$.
As the implied mean-forecast, $m_1(F) = \mu + 2\delta\varphi(0)$, agrees with the implied mean-forecast of the unfocused forecast with $\eta = (8/\pi)^{1/2}\delta$, the terms in the score decomposition (4.15) with $\mathrm{S}(x,y) = (x-y)^2$ derive from the respective terms in the score decomposition for the unfocused forecast, as illustrated in Figure 4.5.

4.A.1.3 Piecewise uniform forecast

Given any fixed index $i \in \{1,2,3\}$, let the tuple $(p_1^{(i)}, p_2^{(i)}, p_3^{(i)}; q_1^{(i)}, q_2^{(i)}, q_3^{(i)})$ attain the value $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}; \frac{5}{10}, \frac{1}{10}, \frac{4}{10})$ if $i = 1$, the value $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}; \frac{1}{10}, \frac{8}{10}, \frac{1}{10})$ if $i = 2$, and the value $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}; \frac{4}{10}, \frac{1}{10}, \frac{5}{10})$ if $i = 3$. Furthermore, let P_i be the CDF of a mixture of uniform measures on $[0,1], [1,2]$, and $[2,3]$ with weights $p_1^{(i)}, p_2^{(i)}$, and $p_3^{(i)}$, respectively. Similarly, let Q_i be the CDF of the respective mixture with weights $q_1^{(i)}, q_2^{(i)}$, and $q_3^{(i)}$, respectively.
The random CDF for the piecewise uniform forecast in Example 4.4 can then be written as $F(x) = P_\iota(x-\mu)$, where the random variables ι and μ are independent, and the integer-valued variable ι is such that

$$(p_1, p_2, p_3; q_1, q_2, q_3) = \left(p_1^{(\iota)}, p_2^{(\iota)}, p_3^{(\iota)}; q_1^{(\iota)}, q_2^{(\iota)}, q_3^{(\iota)}\right).$$

The conditional CDF for the outcome Y given the posited (non) exceedance probability $F(t)$ at any fixed threshold $t \in \mathbb{R}$ or, equivalently, given the quantile forecast $F^{-1}(\alpha)$ at any fixed level $\alpha \in (0,1)$ then is

$$\begin{aligned}\mathbb{P}(Y \leq y \mid F(t) = \alpha) &= \mathbb{P}(Y \leq y \mid F^{-1}(\alpha) = t) = \mathbb{P}(Y \leq y \mid \mu = t - P_\iota^{-1}(\alpha)) \\ &= \frac{1}{\sum_{i=1,2,3} \varphi\left(\frac{t-P_i^{-1}(\alpha)}{c}\right)} \sum_{i=1,2,3} \varphi\left(\frac{t - P_i^{-1}(\alpha)}{c}\right) Q_i(y - (t - P_i^{-1}(\alpha))),\end{aligned}$$

where c is the standard deviation of μ, as defined in Example 4.4. The first moment of F is $m_1(F) = \mu + 1 + \frac{1}{4}\iota$, so that

$$\begin{aligned}\mathbb{P}(Y \leq y \mid m_1(F) = m) &= \mathbb{P}(Y \leq y \mid \mu = m - 1 - \tfrac{1}{4}\iota) \\ = \frac{1}{\sum_{i=1,2,3} \varphi\left(\frac{m-1-\frac{1}{4}i}{c}\right)} &\sum_{i=1,2,3} \varphi\left(\frac{m-1-\frac{1}{4}i}{c}\right) Q_i(y - (m - 1 - \tfrac{1}{4}i))\end{aligned}$$

is a mixture of shifted versions of Q_1, Q_2, and Q_3. The associated first moment is the respective mixture of $m + \frac{3}{20}$, m, and $m - \frac{3}{20}$.
Given any integer $k \geq 0$, let $\beta_k = \sum_{j=1,2,3} \left(j^{k+1} - (j-1)^{k+1}\right) p_j^{(\iota)}$. The second moment of F is $m_2(F) = \mu^2 + \beta_1 \mu + \frac{1}{3}\beta_2$, whence

$$\mathbb{P}(Y \leq y \mid m_2(F) = m) = \mathbb{P}\left(Y \leq y \mid \mu = -\frac{1}{2}\beta_1 \pm \sqrt{\frac{1}{4}\beta_1^2 - \frac{1}{3}\beta_2 + m}\right)$$

also admits an expression in terms of mixtures of shifted versions of Q_1, Q_2, and Q_3. Finally, the third moment of F is $m_3(F) = \mu^3 + \frac{3}{2}\beta_1\mu^2 + \beta_2\mu + \frac{1}{4}\beta_3 = f(\mu; \iota)$, so that the conditional distribution $\mathbb{P}(Y \leq y \mid m_3(F) = m) = \mathbb{P}(Y \leq y \mid f(\mu; \iota) = m)$ and the associated third moment can be computed analogously.

4.A.1.4 Identification functions, unconditional calibration, and canonical loss

In this section, we demonstrate that Definitions 4.9 and 4.24 are unambiguous and do not depend on the choice of the identification function, which is essentially unique. To this end, we first contrast the notions of identification functions in Fissler and Ziegel (2016) and Jordan et al. (2022). Fissler and Ziegel (2016) call $V\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ a (strict $\mathcal{F}$-)identification function if $V(x, \cdot)$ is integrable with respect to all $F \in \mathcal{F}$ for all $x \in \mathbb{R}$. Jordan et al. (2022) additionally require V to be increasing and left-continuous in its first argument. Furthermore, there is a subtle difference in the way that the functional is induced. While Fissler and Ziegel (2016) define the induced functional as the set

$$\mathrm{T}_0(F) = \left\{x \in \mathbb{R} : \int V(x, y)\,\mathrm{d}F(y) = 0\right\},$$

Jordan et al. (2022) define it to be the closed interval $T(F) = [T^-(F), T^+(F)]$, where $\mathrm{T}^-(F)$ and $\mathrm{T}^+(F)$ are defined in (4.7) and (4.8), respectively. The approach by Jordan et al. (2022) allows for quantiles to be treated in full generality and ensures that the interval $\mathrm{T}(F)$ coincides with the closure of $\mathrm{T}_0(F)$ if the latter is nonempty.
In the setting of Fissler and Ziegel (2016), if V is an identification function, then so is $(x, y) \mapsto h(x)V(x, y)$ whenever $h(x) \neq 0$ for all $x \in \mathbb{R}$. If the class $\mathcal{F}$ is sufficiently rich, then any two locally bounded identification functions V and $\widetilde{V}$ that induce a functional T_0 of singleton type relate to each other in the stated form almost everywhere on the interior of $\mathrm{T}_0(\mathcal{F}) \times \mathbb{R}$ (Dimitriadis et al., 2022b, Theorem 4), which implies that increasing identification functions of prediction error form are unique up to a positive constant. The following proposition provides an elementary proof under slightly different conditions that are tailored to our setting. Notably, identification functions of prediction error form induce functionals that are equivariant under translation by Proposition 4.7 of Fissler and Ziegel (2019), a result which can easily be transferred to the setting of Jordan et al. (2022).

Proposition 4.34. *Let $\mathcal{F}$ be a convex class of probability measures such that $\delta_y \in \mathcal{F}$ for all $y \in \mathbb{R}$. If the functional T is induced on $\mathcal{F}$ by an identification function $V(x,y) = v(x-y)$ of prediction error form, where v is increasing and left-continuous with $v(-r) < 0$ and $v(r) > 0$ for some $r > 0$, then any other identification function of the stated form that induces T on $\mathcal{F}$ is a positive multiple of V.*

Proof. Let $V\colon (x,y) \mapsto v(x-y)$ and $\widetilde{V}\colon (x,y) \mapsto \widetilde{v}(x-y)$ induce the functionals T and $\widetilde{\mathrm{T}}$, respectively. We proceed to show that $\mathrm{T} = \widetilde{\mathrm{T}}$ implies $\widetilde{v} = h_0 \cdot v$ for some constant $h_0 > 0$.
To this end, suppose that $\mathrm{T} = \widetilde{\mathrm{T}}$, and let

$$r^- = \sup\{r : \widetilde{v}(r) < 0\} = \widetilde{\mathrm{T}}^-(\delta_0) = \mathrm{T}^-(\delta_0) = \sup\{r : v(r) < 0\} > -\infty,$$
$$r^+ = \inf\{r : \widetilde{v}(r) > 0\} = \widetilde{\mathrm{T}}^+(\delta_0) = \mathrm{T}^+(\delta_0) = \inf\{r : v(r) > 0\} < \infty.$$

By left-continuity and monotonicity of v and $\widetilde{v}$, it follows that $v(r) = \widetilde{v}(r) = 0$ for $r \in (r^-, r^+]$, $v(r) < 0$ and $\widetilde{v}(r) < 0$ for $r < r^-$, and $v(r) > 0$ and $\widetilde{v}(r) > 0$ for $r > r^+$.
Let $h(r) = \widetilde{v}(r)/v(r) > 0$ for $r \in \mathbb{R} \setminus [r^-, r^+]$. If $r < r^- \leq r^+ < s$, then $\widetilde{v}(r) = h(r)v(r) < 0$ and $\widetilde{v}(s) = h(s)v(s) > 0$. Assume $h(r) < h(s)$, and let $p \in (0,1)$ be such that

$$\left(1 - \frac{h(s)v(s)}{h(r)v(r)}\right)^{-1} < p < \left(1 - \frac{v(s)}{v(r)}\right)^{-1}.$$

Then $(1-p)h(r)v(r) + ph(s)v(s) > 0 > (1-p)v(r) + pv(s)$ and $\widetilde{\mathrm{T}}^+(p\delta_{-s} + (1-p)\delta_{-r}) < 0 \leq \mathrm{T}^-(p\delta_{-s} + (1-p)\delta_{-r})$, a contradiction. An analogous argument applies if we assume that $h(r) > h(s)$, and we conclude that $h(r) = h(s)$.
If $r, s < r^-$, then $h(r) = h(s) = h(t)$ for any $t > r^+$ by the above line of reasoning. An analogous argument yields $h(r) = h(s)$ for $r, s > r^+$. Therefore, the function h is constant and $v(r) = h_0 \cdot \widetilde{v}(r)$ for a constant $h_0 > 0$ and all $r \in \mathbb{R} \setminus \{r^-\}$. Finally, we obtain $v(r^-) = \lim_{r \uparrow r^-} v(r) = \lim_{r \uparrow r^-} h_0 \cdot \widetilde{v}(r) = h_0 \cdot \widetilde{v}(r^-)$ by left-continuity. □

Hence, if we assume an identification function of type (i) in Assumption 4.8, Definitions 4.9 and 4.24 do not depend on the choice of the identification function, as it is unique up to a positive constant. Trivially, the same holds true for type (ii). To complete the argument that the definitions are unambiguous, the following technical argument is needed.

Remark 4.35. If a functional T of singleton type is identified by both an identification function $V(x,y) = v(x-y)$ of type (i) and an identification function $\widetilde{V}(x,y) = x - \mathrm{T}(\delta_y)$ of type (ii), then $\widetilde{V}$ is also of type (i). To confirm this claim, let z denote the unique value at which the sign of v changes, and note that $z = \mathrm{T}(\delta_y) - y$ for all y since V induces the functional T for each Dirac measure δ_y. Hence, $\mathrm{T}(\delta_y) = y + z$ and $\widetilde{V}(x,y) = x - y - z$ is of type (i).

We close this section with comments on the role of the class $\mathcal{F}$. As expressed by Assumption 4.8, we prefer to work with identification functions that elicit the target functional T on a large, convex class $\mathcal{F}$ of probability measures, to avoid unnecessary constraints on forecast(er)s. Furthermore, when evaluating stand-alone point forecasts, the underlying predictive distributions typically are implicit, and assumptions other than the existence of the functional at hand are unwarranted and contradict the prequential principle. Evidently, if the class $\mathcal{F}$ is sufficiently restricted, additional identification functions arise. For example, the piecewise constant identification function associated with the median can be used to identify the mean within any class of symmetric distributions.

4.A.1.5 Strong threshold calibration does not imply auto-calibration

As pointed out by Sahoo et al. (2021, p. 5), strong threshold calibration does not imply auto-calibration. Here, we provide a simple example illustrating this fact as Sahoo et al. (2021) do not present such. The example is similar in spirit to the continuous forecast of Example 4.14(a) (as $c \to 0$) but with strictly increasing distribution functions satisfying Assumption 4.15.

Example 4.36. Let F be a mixture of uniform distributions on the intervals $[0,1], [1,2], [2,3]$, and $[3,4]$ with weights p_1, p_2, p_3, and p_4, respectively, and let Y be from a mixture with weights q_1, q_2, q_3, and q_4. Furthermore, let the tuple $(p_1, p_2, p_3, p_4; q_1, q_2, q_3, q_4)$ attain each of the values

$$\begin{aligned}
&\left(\tfrac{4}{10}, \tfrac{1}{10}, \tfrac{4}{10}, \tfrac{1}{10}; \tfrac{16}{25}, \tfrac{4}{25}, \tfrac{4}{25}, \tfrac{1}{25}\right), \quad \left(\tfrac{1}{10}, \tfrac{4}{10}, \tfrac{1}{10}, \tfrac{4}{10}; \tfrac{4}{25}, \tfrac{16}{25}, \tfrac{1}{25}, \tfrac{4}{25}\right),\\
&\left(\tfrac{4}{10}, \tfrac{1}{10}, \tfrac{1}{10}, \tfrac{4}{10}; \tfrac{4}{25}, \tfrac{1}{25}, \tfrac{4}{25}, \tfrac{16}{25}\right), \quad \left(\tfrac{1}{10}, \tfrac{4}{10}, \tfrac{4}{10}, \tfrac{1}{10}; \tfrac{1}{25}, \tfrac{4}{25}, \tfrac{16}{25}, \tfrac{4}{25}\right)
\end{aligned}$$

with equal probability. The equal average of the distribution of the PIT conditional on either forecast from the top row, and either forecast from the bottom row, is uniform. As any nontrivial conditioning in terms of a threshold yields a combination of two forecast cases, one from the top row and one from the bottom row, the forecast F is strongly threshold calibrated.

4.A.1.6 Remarks on Figure 4.4

The root transforms in the moment reliability diagrams in the bottom row of Figure 4.4 bring the first, second, and third moment to the same scale. The peculiar dent in the reliability curve for the (third root of the) third moment of the piecewise uniform forecast results from the transform, which magnifies small deviations between $x = m_3(F)$ and x_{rc} when x is close to zero. For comparison, Figure 4.13 shows moment reliability diagrams for all three forecasts without applying the root transform.

4.A.2 Consistency resamples and calibration tests

Monte Carlo based consistency bands for T-reliability diagrams can be generated from resamples, at any desired nominal level. The consistency bands then show

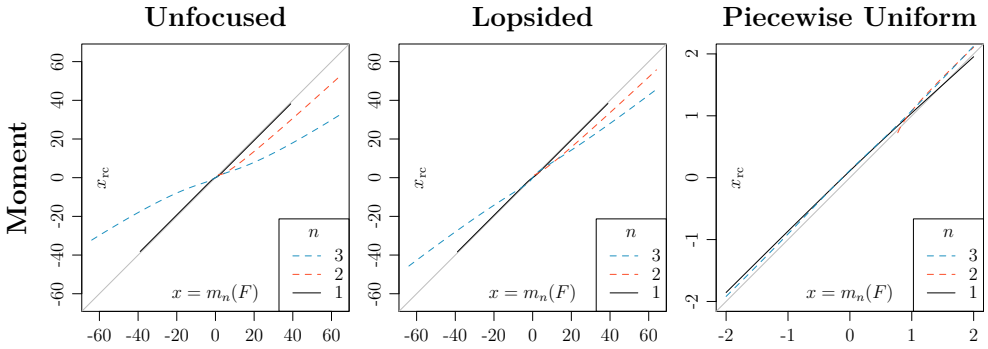


Figure 4.13 Same as the lower row of Figure 4.4 but with displays on original (rather than root-transformed) scales: Moment reliability diagrams for point forecasts induced by (left) the unfocused forecast with $\eta_0 = 1.5$ and (middle) the lopsided forecast with $\delta_0 = 0.7$ from Example 4.2, and (right) the piecewise uniform forecast with $c = 0.5$ from Example 4.4.

the pointwise range of the resampled calibration curves. For now, let us assume that we have data $(x_1, y_1), \ldots, (x_n, y_n)$ of the form (4.24) along with m resamples at hand, and defer the critical question of how to generate the resamples.

Algorithm 4.2. Consistency bands for T-reliability curves based on resamples

Input: resamples $(x_1, \widetilde{y}_1^{(j)}), \ldots, (x_n, \widetilde{y}_n^{(j)})$ for $j = 1, \ldots, m$
Output: $\alpha \times 100\%$ consistency band
for $j \in \{1, \ldots, m\}$ **do**
  apply Algorithm 4.1 to obtain $\widehat{x}_1^{(j)}, \ldots, \widehat{x}_n^{(j)}$ from $(x_1, \widetilde{y}_1^{(j)}), \ldots, (x_n, \widetilde{y}_n^{(j)})$
end
for $i \in \{1, \ldots, n\}$ **do**
  let l_i and u_i be the empirical quantiles of $\widehat{x}_i^{(1)}, \ldots, \widehat{x}_i^{(m)}$ at level $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$
end
interpolate the point sets $(x_1, l_1), \ldots, (x_n, l_n)$ and $(x_1, u_1), \ldots, (x_n, u_n)$ linearly, to obtain the lower and upper bound of the consistency band, respectively

Complementary to consistency bands, tests for the assumed type of calibration, as quantified by the functional T and a generic miscalibration measure MCB, can be performed as usual. Specifically, we compute MCB_j for each resample $j = 1, \ldots, m$, and, if r of the resampled measures $\mathsf{MCB}_1, \ldots, \mathsf{MCB}_m$ are less than or equal to the miscalibration measure computed from the original data, we declare a Monte Carlo p-value of $1 - \frac{r}{m+1}$.

4.A.2.1 Consistency resamples under the hypothesis of auto-calibration

When working with original data of the form (4.23), we can generate resamples under the hypothesis of auto-calibration in the obvious way, as follows.

Algorithm 4.3. Consistency resamples under the hypothesis of auto-calibration

Input: $(F_1, y_1), \ldots, (F_n, y_n)$
Output: resamples $(x_1, \widetilde{y}_1^{(j)}), \ldots, (x_n, \widetilde{y}_n^{(j)})$ for $j = 1, \ldots, m$
for $i \in \{1, \ldots, n\}$ **do**
  let $x_i = \mathrm{T}(F_i)$
end
for $j \in \{1, \ldots, m\}$ **do**
  for $i = 1, \ldots, n$ **do**
    sample $\widetilde{y}_i^{(j)}$ from F_i
  end
end

As noted, in the case of threshold calibration, the induced outcome is binary, whence the assumptions of auto-calibration and T-calibration coincide. For other types of functionals, auto-calibration is a strictly stronger assumption than T-calibration, and it is important to note that the resulting inferential procedures may be confounded by forecast attributes other than T-calibration. For illustration, let us return to the setting of Example 4.1 and suppose that, conditionally on a standard normal variate μ, the outcome Y is normal with mean μ and variance 1. Given any fixed $\sigma > 0$, the forecast $F_\sigma = \mathcal{N}(\mu, \sigma^2)$ is auto-calibrated if, and only if, $\sigma = 1$. However, if T is the mean or median functional, then F_σ is T-calibrated under any $\sigma > 0$. Clearly, if we use Algorithm 4.3 to generate resamples, then the consistency bands generated by Algorithm 4.2 might be misleading with regard to the assessment of T-calibration. For example, if $\sigma < 1$ the confidence bands tend to be narrow and might erroneously suggest a lack of T-calibration, despite the forecast being T-calibrated.

4.A.2.2 Consistency resamples under the hypothesis of T-calibration

The issues just described call for an alternative to Algorithm 4.3. Residual-based approaches can be used to generate resamples under the weaker hypothesis of T-calibration. In developing such a method, we restrict the discussion to single-valued functionals T under which $y_i = \mathrm{T}(\delta_i)$, which covers all cases of key interest, such as the mean functional, lower or upper quantiles, and expectiles. As is standard in regression diagnostics, residual-based approaches operate on the basis of tuples $(x_1, y_1), \ldots, (x_n, y_n)$ of the form (4.24) under the assumptions of independence between the point forecast, x_i, and the residual, $y_i - x_i$, and exchangeability of the residuals. For a discussion in the context of backtests in banking regulation, see Example 3 of Nolde and Ziegel (2017).
Interestingly, Theorem 4.21 demonstrates that under these assumptions a forecast

is conditionally T-calibrated if, and only if, it is unconditionally T-calibrated. Thus, we draw resamples in a two-stage procedure. First, we find the constant c from Theorem 4.27 such that the empirical distribution of $(x_1 + c, y_1), \ldots, (x_n + c, y_n)$ or, equivalently, $(x_1, y_1 - c), \ldots, (x_n, y_n - c)$, is unconditionally T-calibrated, and then we resample from the respective residuals, as follows.

Algorithm 4.4. Consistency resamples under the joint hypothesis of T-calibration and independence between point forecasts and residuals

Input: $(x_1, y_1), \ldots, (x_n, y_n)$
Output: resamples $(x_1, \widetilde{y}_1^{(j)}), \ldots, (x_n, \widetilde{y}_n^{(j)})$ for $j = 1, \ldots, m$
for $i = 1, \ldots, n$ **do**
  let $r_i = y_i - x_i$
end
find c such that $(x_1 + c, y_1), \ldots, (x_n + c, y_n)$ is unconditionally T-calibrated
for $j \in \{1, \ldots, m\}$ **do**
  sample $\widetilde{r}_1, \ldots, \widetilde{r}_n$ from $\{r_1, \ldots, r_n\}$ with replacement
  for $i = 1, \ldots, n$ **do**
    let $\widetilde{y}_i = x_i + \widetilde{r}_i - c$
  end
end

As noted in the main text, the consistency bands for the threshold reliability diagrams in Figures 4.6 and 4.8 have been generated by Algorithms 4.2 and 4.4. This approach is similar to the Monte Carlo technique proposed by Dimitriadis et al. (2021) that applies in the case of (induced) binary outcomes (only). However, unlike Dimitriadis et al. (2021), we do not resample the forecasts themselves. To generate consistency bands for the mean and quantile reliability diagrams in these figures, we apply Algorithm 4.2 to $m = 1000$ resamples generated by Algorithm 4.4. Evidently, this procedure is crude and relies on classical assumptions. Nonetheless, we believe that in many practical settings, where visual tools for diagnostic checks of calibration are sought, the consistency bands thus generated provide useful guidance.

Further methodological development on consistency and confidence bands needs to be tailored to the specific functional T of interest, and follow-up work on Monte Carlo techniques and large sample theory is strongly encouraged. Extant asymptotic theory for nonparametric isotonic regression, as implemented by Algorithm 4.1, is available for quantiles and the mean or expectation functional, as developed and reviewed by Barlow et al. (1972), Casady and Cryer (1976), Wright (1984), Robertson et al. (1988), El Barmi and Mukerjee (2005), and Mösching and Dümbgen (2020), and can be leveraged, though with hurdles, as rates of convergence depend on distributional assumptions and limit distributions involve nuisance parameters that need to be estimated, whereas the use of bootstrap methods might be impacted by the issues described by Sen et al. (2010).

4.A.2.3 Reliability diagrams and consistency bands for probabilistic and marginal calibration

For the classical notions of unconditional calibration in Section 4.2.2, the CORP approach does not apply directly, but its spirit can be retained and adapted.
As for probabilistic calibration, the prevalent practice is to plot histograms of empirical probability integral transform (PIT) values, as proposed by Diebold et al. (1998), Gneiting et al. (2007), and Czado et al. (2009), though this practice is hindered by the necessity for binning, as analyzed by Heinrich (2021) in the nearly equivalent setting of rank histograms. The population version of our suggested alternative is the *PIT reliability diagram*, which is simply the graph of the CDF of the PIT Z_F in (4.1). The PIT reliability diagram coincides with the diagonal in the unit square if, and only if, F is probabilistically calibrated. For tuples of the form (4.23) the *empirical PIT reliability diagram* shows the empirical CDF of the (potentially randomized) PIT values. This approach does not require binning and can be interpreted in much the same way as a PIT diagram: An inverse S-shape corresponds to a U-shape in histograms and indicates underdispersion of the forecast, as typically encountered in practice. Evidently, this idea is not new and extant implementations can be found in work by Pinson and Hagedorn (2012) and Henzi et al. (2021).
As regards marginal calibration, we define the population version of the *marginal reliability diagram* as the point set

$$\{(\mathbb{E}[F(y)], \mathbb{P}(Y \leq y)) \in [0,1]^2 : y \in \mathbb{R}\}.$$

The marginal reliability diagram is concentrated on the diagonal in the unit square if, and only if, F is marginally calibrated. For tuples of the form (4.23) the *empirical marginal reliability diagram* is a plot of the empirical non-exceedance probability (NEP) $\widehat{F}_0(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y \geq y_i\}$ against the average forecast NEP $\bar{F}(y) = \frac{1}{n}\sum_{i=1}^{n} F_i(y)$ at the unique values y of the outcomes $y_1, \ldots, y_n$, and interpolated linearly in between. Of course, this idea is not new either and the resulting diagram can be interpreted as a P-P plot.
For marginal calibration diagrams, we obtain consistency bands under the assumption of marginal calibration by drawing resamples $y_1^{(j)}, \ldots, y_n^{(j)}$ from $\bar{F} = \frac{1}{n}\sum_{i=1}^{n} F_i$, computing the respective marginal reliability curve, and repeating over Monte Carlo replicates $j = 1, \ldots, m$. Then we find consistency bands in the spirit of Algorithm 4.2. For PIT reliability diagrams, a trivial technique applies as we may obtain consistency bands under the assumption of probabilistic calibration by (re)sampling n independent standard uniform variates, computing the respective empirical CDF, and repeating over Monte Carlo replicates. Evidently, there are alternatives based on empirical process theory (Shorack and Wellner, 2009).
Figure 4.14 illustrates PIT and marginal reliability diagrams on our customary examples, along with 90% consistency bands based on $m = 1000$ Monte Carlo replicates.

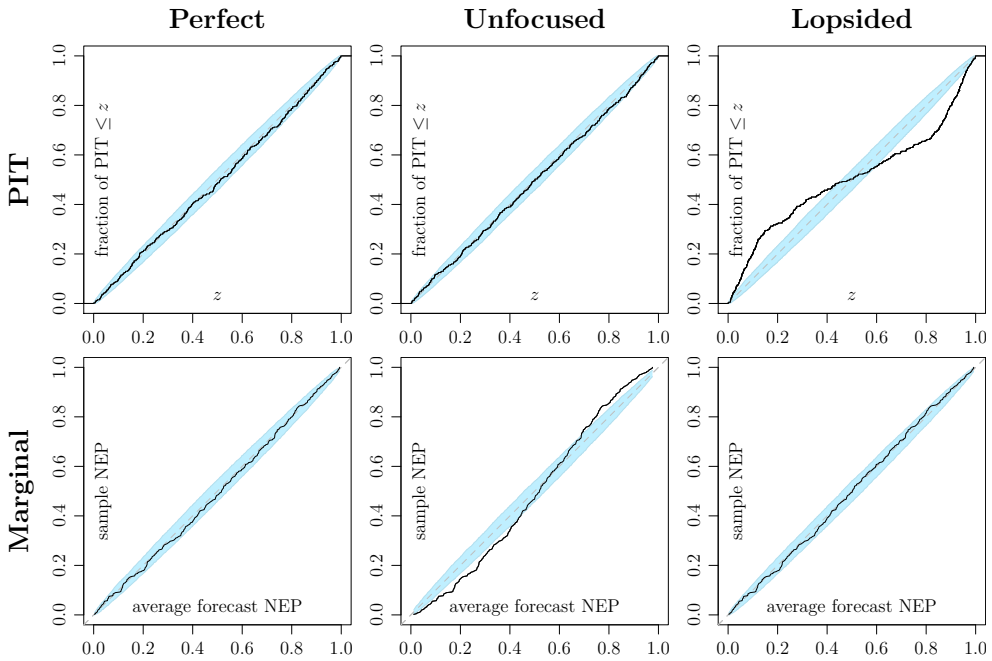


Figure 4.14 PIT (top) and marginal (bottom) reliability diagrams for the perfect (left), unfocused (middle), and lopsided (right) forecast from Examples 4.1 and 4.2, along with 90% consistency bands based on samples of size 400.

4.A.2.4 Testing hypotheses of calibration

While the explicit development of calibration tests exceeds the scope of this chapter, we believe that the results and discussion in Section 4.2 convey an important general message: It is critical that the assessed notion of calibration be carefully and explicitly specified. Throughout, we consider tests under the assumption of independent, identically distributed data from a population. For extensions to dependent samples, we refer to Strähl and Ziegel (2017), who generalized the prediction space concept to allow for serial dependence, and point at methods introduced by, e.g., Corradi and Swanson (2007), Knüppel (2015), and Bröcker and Ben Bouallègue (2020).

The most basic case is that of tuples $(x_1, y_1), \ldots, (x_n, y_n)$ of the form (4.24), where implicitly or explicitly $x_i = \mathrm{T}(F_i)$ for a single-valued functional T. We first discuss tests of unconditional calibration. If the simplified condition (4.11) is sufficient, a two-sided t-test based on $\widehat{v} = \frac{1}{n}\sum_{i=1}^{n} V(x_i, y_i)$ can be used to test for unconditional calibration. In the general case, two one-sided t-tests can be used along with a Bonferroni correction. In the special case of quantiles, there is no need to resort to the approximate t-tests, and exact binomial tests can be used instead. Essentially, this special case is the setting of backtests for value-at-risk reports in banking regulation, for which we refer to Nolde and Ziegel (2017, Sections 2.1–2.2).

As noted earlier in the section, resamples generated under the hypothesis of conditional T-calibration can readily be used to perform Monte Carlo tests for the re-

spective hypothesis, based on CORP score components that are computed on the surrogate data. Alternatively, one might leverage extant large sample theory for nonparametric isotonic regression (Barlow et al., 1972; Casady and Cryer, 1976; Wright, 1984; Robertson et al., 1988; El Barmi and Mukerjee, 2005; Mösching and Dümbgen, 2020). Independently of the use of resampling or asymptotic theory, CORP based tests avoid the issues and instabilities incurred by binning (Dimitriadis et al., 2021, Section S2) and may simultaneously improve efficiency and stability. In passing, we hint at relations to the null hypothesis of Mincer-Zarnowitz regression (Krüger and Ziegel, 2021) and tests of predictive content (Galbraith, 2003; Breitung and Knüppel, 2021).

We move on to the case of fully specified distributions, where we work with tuples $(F_1, y_1), \ldots, (F_n, y_n)$ of the form (4.23), where F_i is a posited conditional CDF for y_i $(i = 1, \ldots, n)$. Tests for probabilistic calibration then amount to tests for the uniformity of the (potentially, randomized) PIT values. Wallis (2003) and Wilks (2019, p. 769) suggest chi-square tests for this purpose, which depend on binning, and thus are subject to the aforementioned instabilities. To avoid binning, we recommend the use of test statistics that operate on the empirical CDF of the PIT values, such as the classical Kolmogorov-Smirnov (KS) statistic, as suggested and used to test for PIT calibration by Noceti et al. (2003) and Knüppel (2015), or, more generally, tests based on distance measures between the empirical CDF of the PIT values, and the CDF of the standard uniform distribution that arises under the hypothesis of probabilistic calibration. Recently proposed alternatives arise via e-values (Henzi and Ziegel, 2022). Similarly, tests for marginal calibration can be based on resamples and distance measures between $\bar{F}$ and $\widehat{F}_0$, or leverage asymptotic theory.

In the distributional setting, arbitrarily many types of reliability can be tested for, and all of the aforementioned tests for unconditional or conditional T-calibration apply. Multiple testing needs to be accounted for properly, and the development of simultaneous tests for various types of calibration would be useful. In this context, let us recall from Theorem 4.16 that, subject to technical conditions, CEP, threshold, and quantile calibration are equivalent and tests for CEP calibration (Held et al., 2010; Strähl and Ziegel, 2017), quantile and threshold calibration assess identical hypotheses.

4.A.3 Time series settings and the Bank of England example

In typical time series settings, as exemplified by our analysis of Bank of England forecasts in Section 4.3, the assumption of independent replicates of forecasts and observations is too restrictive. While the diagnostic methods proposed in this chapter continue to apply, statistical inference requires care, as discussed by Corradi and Swanson (2007) and Knüppel (2015), among other authors. Here, we elucidate the role of uniform and independent probability integral transform (PIT) values for calibration in time series settings, and give further details and results for the Bank of England example.

4.A.3.1 The role of uniform and independent PITs

In a landmark paper, Diebold et al. (1998, p. 867) showed that a sequence of continuous predictive distributions F_t for a sequence Y_t of observations at time $t = 0, 1, \ldots$ results in a sequence of independent, uniformly distributed PITs if F_t is ideal relative to the σ-algebra generated by past observations, $\mathcal{A}_t = \sigma(Y_0, Y_1, \ldots, Y_{t-1})$. This property does not depend on the continuity of F_t and continues to hold under general predictive CDFs and the randomized definition (4.1) of the PIT (Rüschendorf and de Valk, 1993, Theorem 3).
In the case of continuous predictive distributions, Tsyplakov (2011, Section 2) noted without proof that if the forecasts F_t are based only on past observations, i.e., if F_t is $\mathcal{A}_t$-measurable, then the converse holds, namely, uniform and independent PITs arise only if F_t is ideal relative to $\mathcal{A}_t$. The following result formalizes Tsyplakov's claim and proves it in the general setting, without any assumption of continuity.

Theorem 4.37. *Let $(Y_t)_{t=0,1,\ldots}$ be a sequence of random variables, and let $\mathcal{A}_t = \sigma(Y_0, \ldots, Y_{t-1})$ for $t = 0, 1, \ldots$ Furthermore, let $(F_t)_{t=0,1,\ldots}$ be a sequence of CDFs, such that F_t is $\mathcal{A}_t$-measurable for $t = 0, 1, \ldots$, and let $(U_t)_{t=0,1,\ldots}$ be a sequence of independent, uniformly distributed random variables, independent of the sequence (Y_t). Then the sequence of randomized PITs, $(Z_t) = (F_t(Y_t-) + U_t(F_t(Y_t) - F_t(Y_t-)))$ is an independent sequence of uniform random variables on the unit interval if, and only if, F_t is ideal relative to $\mathcal{A}_t$, i.e., $F_t = \mathcal{L}(Y_t \mid \mathcal{A}_t)$ almost surely for $t = 0, 1, \ldots$*

The proof utilizes the following simple lemma.

Lemma 4.38. *Let X, Y, Z be random variables. If $X = Z$ almost surely, then $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid Z]$ almost surely.*

Proof. Problem 14 of Breiman (1992, Chapter 4), which is proved by Schmidt (2011, Satz 18.2.10), states that for random variables X_1 and X_2 such that $\sigma(Y, X_1)$ is independent of $\sigma(X_2)$, $\mathbb{E}[Y \mid X_1, X_2] = \mathbb{E}[Y \mid X_1]$ almost surely. The statement of the lemma follows as $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X, X - Z] = \mathbb{E}[Y \mid Z, X - Z] = \mathbb{E}[Y \mid Z]$ almost surely. □

Proof of Theorem 4.37. Since F_t is measurable with respect to $\mathcal{A}_t$, there exists a measurable function $f_t \colon \mathbb{R}^t \to \mathcal{F}$ such that $F_t = f_t(Y_0, \ldots, Y_{t-1})$ for each t by the Doob–Dynkin Lemma (Schmidt, 2011, Satz 7.1.16).[5] We define

$$G_t := f_t(G_0^{-1}(Z_0), \ldots, G_{t-1}^{-1}(Z_{t-1}))$$

[5]Note that f_0 is constant, and f_t is not a random quantity but a fixed function that encodes how the predictive distributions are generated from past observations. The σ-algebra on $\mathcal{F}$, which is implicitly used throughout, is given by

$$\mathcal{A}_{\mathcal{F}} = \sigma(\{\{F \in \mathcal{F} : F(x) \in B\} : x \in \mathbb{Q},\, B \in \mathcal{B}(\mathbb{R})\}),$$

where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$. For each $x \in \mathbb{Q}$ there exists a measurable function $f_{x,t}$ such that $F_t(x) = f_{x,t}(Y_0, \ldots, Y_{t-1})$ by the Doob-Dynkin Lemma, and f_t is essentially the countable (and hence measurable) collection $(f_{x,t})_{x \in \mathbb{Q}}$.

recursively for all t, and show the "only if" assertion by induction.
To this end, let $t \geq 0$ and assume the induction hypothesis that F_i is ideal relative to $\mathcal{A}_i$ for $i = 0, \ldots, t-1$. By Rüschendorf and de Valk (1993, Theorem 3(a)) and the construction of G_t, the induction hypothesis implies

$$(Y_0, \ldots, Y_{t-1}) = (F_0^{-1}(Z_0), \ldots, F_{t-1}^{-1}(Z_{t-1})) = (G_0^{-1}(Z_0), \ldots, G_{t-1}^{-1}(Z_{t-1}))$$

almost surely, where the last vector is $\sigma(Z_0, \ldots, Z_{t-1})$-measurable. By Lemma 4.38, it follows that

$$\mathcal{L}(Z_t \mid \mathcal{A}_t) = \mathcal{L}(Z_t \mid \sigma(G_0^{-1}(Z_0), \ldots, G_{t-1}^{-1}(Z_{t-1}))) = \mathcal{U}([0,1])$$

almost surely, where the second equality stems from the fact that Z_t is independent of $\sigma(G_0^{-1}(Z_0), \ldots, G_{t-1}^{-1}(Z_{t-1})) \subset \sigma(Z_0, \ldots, Z_{t-1})$. This independence implies that F_t is ideal relative to $\mathcal{A}_t$ because

$$F_t(y) = \mathbb{P}(Z_t < F_t(y) \mid \mathcal{A}_t) \leq \mathbb{P}(Y_t \leq y \mid \mathcal{A}_t) \leq \mathbb{P}(Z_t \leq F_t(y) \mid \mathcal{A}_t) = F_t(y)$$

almost surely, and hence $F_t(y) = \mathbb{P}(Y_t \leq y \mid \mathcal{A}_t)$ almost surely for all $y \in \mathbb{Q}$, thereby completing both the induction step and the claim for the base case $t = 0$. □

Evidently, the assumption that no information other than the history of the time series itself has been utilized to construct the forecasts is very limiting. In this light, it is not surprising that, while the "if" part of Theorem 4.37 is robust, the "only if" claim fails if F_t is allowed to use information beyond the canonical filtration, even if that information is uninformative. A simple counterexample is given by the unfocused forecast from Example 4.2, which is probabilistically calibrated but fails to be auto-calibrated. Its PIT nevertheless is uniform and independent even for autoregressive variants (Tsyplakov, 2011, Section 6).

4.A.3.2 Details and further results for the Bank of England example

Bank of England forecasts of inflation rates are available within the data accompanying the quarterly Monetary Policy Report (formerly Inflation Report), which is available online at `https://www.bankofengland.co.uk/sitemap/monetary-policy-report`. The forecasts are visualized and communicated in the form of *fan charts* that span prediction intervals at increasing forecast horizons, and derive from two-piece normal forecast distributions. A detailed account of the parametrizations for the two-piece normal distribution used by the Bank of England can be found in Julio (2006), and we have implemented the formulas in this reference. Historical quarterly CPI inflation rates are published by the UK Office for National Statistics (ONS) and available online at `https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/d7g7`.
We consider forecasts of consumer price index (CPI) inflation based on market expectations for future interest rates at prediction horizons of zero to six quarters ahead, valid for the third quarter of 2005 up to the first quarter of 2020, for a total

of $n = 59$ quarters. These and earlier Bank of England forecasts of inflation rates have been checked for reliability by Wallis (2003), who considered probabilistic calibration, by Clements (2004) in terms of probabilistic, mean, and threshold calibration, by Galbraith and Van Norden (2012), who considered probabilistic and mean calibration, by Strähl and Ziegel (2017) with focus on conditional exceedance probability (CEP) calibration, and by Pohle (2020), who considered quantile calibration. The 2% inflation target is discussed on the Bank of England website at `https://www.bankofengland.co.uk/monetary-policy/inflation`. Figures 4.15–4.20 show calibration diagnostics for inflation forecasts at prediction horizons of $k \in \{0, 2, 3, 4, 5, 6\}$ quarters ahead, in the same format as Figure 4.8 in the main text, which concerns forecasts at a lead time of one quarter.

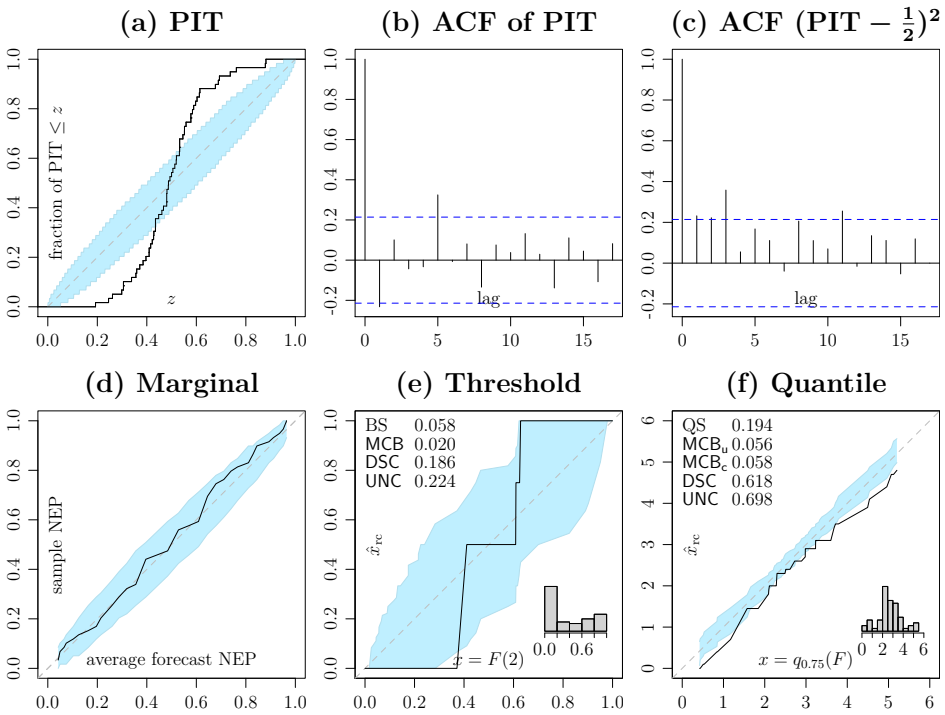


Figure 4.15 Same as Figure 4.8 but at a prediction horizon of zero quarters.

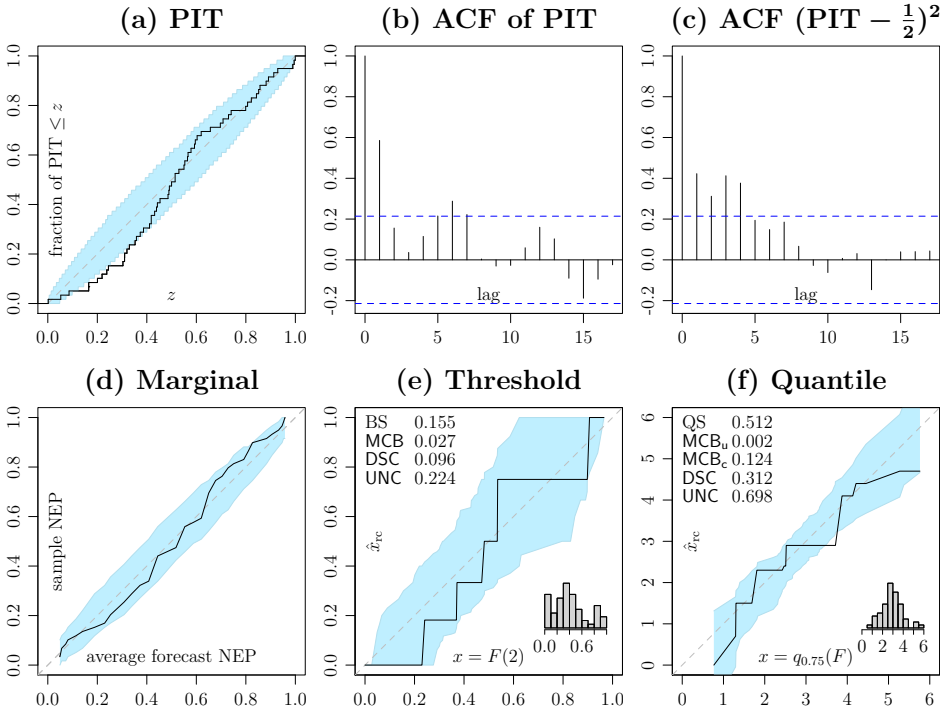


Figure 4.16 Same as Figure 4.15 but at a prediction horizon of two quarters.

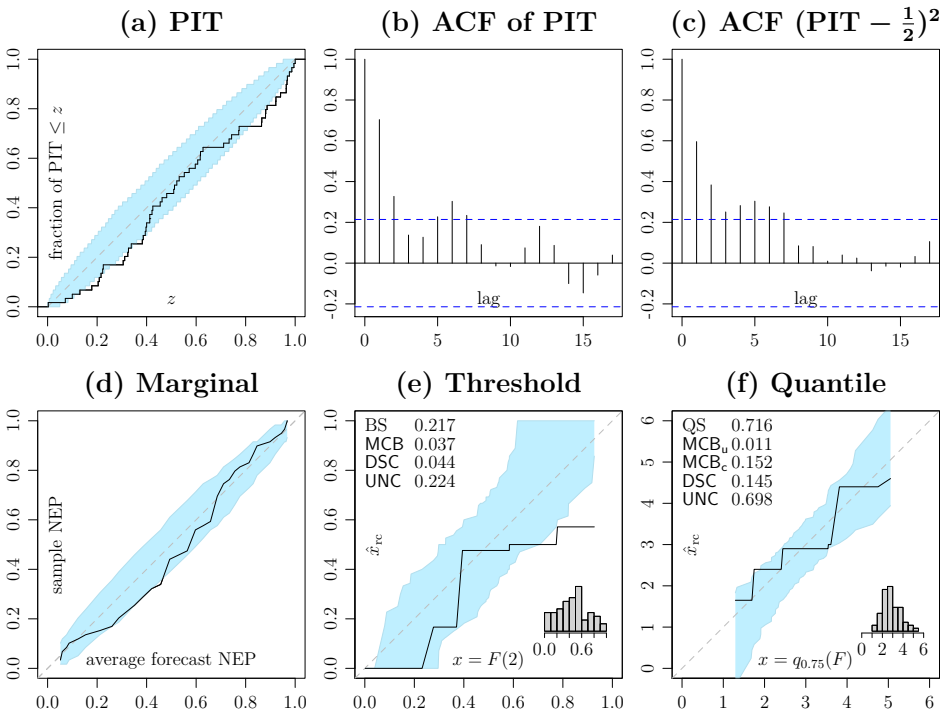


Figure 4.17 Same as Figure 4.15 but at a prediction horizon of three quarters.

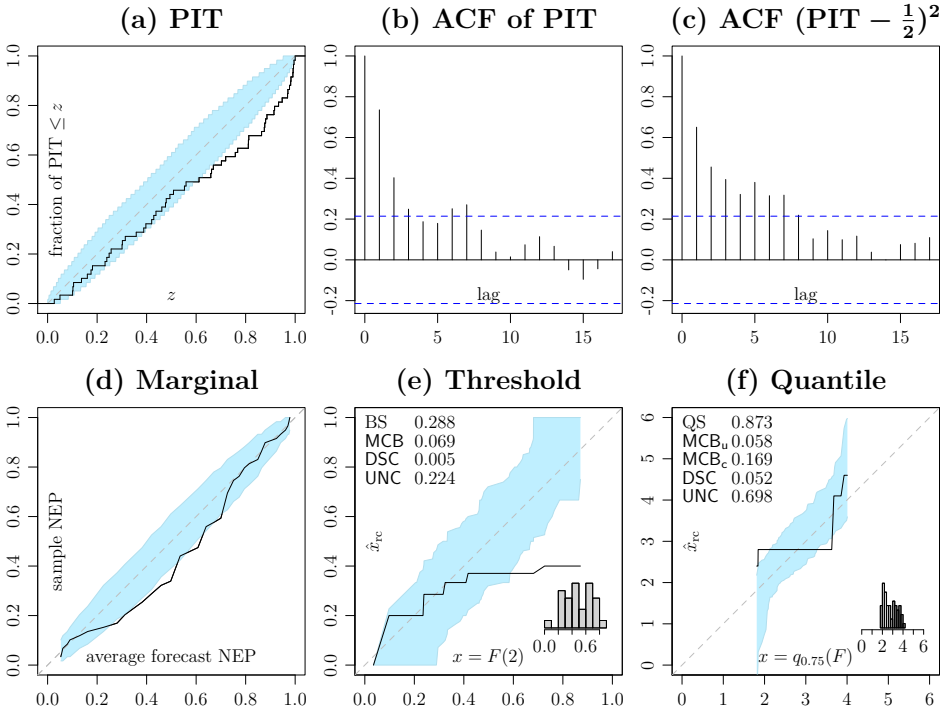


Figure 4.18 Same as Figure 4.15 but at a prediction horizon of four quarters.

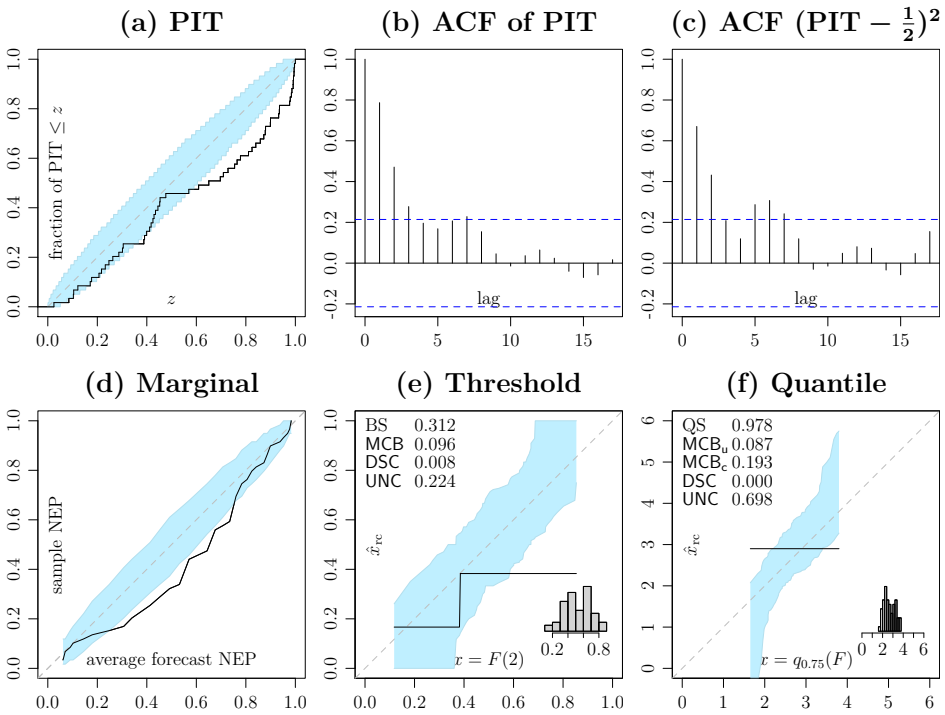


Figure 4.19 Same as Figure 4.15 but at a prediction horizon of five quarters.

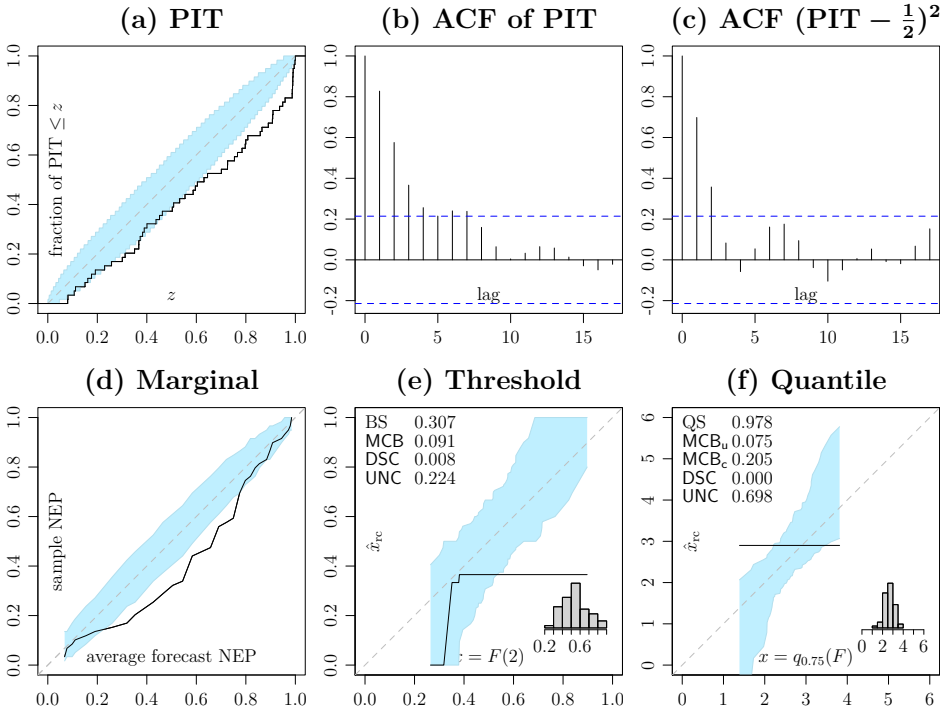


Figure 4.20 Same as Figure 4.15 but at a prediction horizon of six quarters.

5 | Elicitability of Probabilistic Top List Functionals

In this chapter, I propose a family of consistent scoring functions for evaluating probabilistic top-k list predictions, that is, probabilistic forecasts specifying the k most likely outcomes along with their predictive probabilities in classification settings. The proposed scoring functions are based on symmetric proper scoring rules, and the scores are computed by assigning a simple proxy probability to the remaining classes. I show that this construction results in consistent top-k list scoring functions for fixed k, which can be used to compare top list predictions of varying length in a balanced manner. If the underlying scoring rule is strictly proper, the scoring function is strictly consistent for the top-k list functional.

5.1 Introduction

In the face of uncertainty, predictions ought to quantify their level of confidence (Gneiting and Katzfuss, 2014). This idea has been recognized for decades in the literature on weather forecasting (Brier, 1950; Murphy, 1977) and probabilistic forecasting (Dawid, 1984; Gneiting and Raftery, 2007). Ideally, a prediction specifies a probability distribution over potential outcomes. Such predictions are evaluated and compared by means of proper scoring rules, which quantify their value in a way that rewards truthful prediction (Gneiting and Raftery, 2007). In statistical classification and machine learning, the need for reliable uncertainty quantification has not gone unnoticed, as exemplified by the growing interest in the calibration of probabilistic classifiers (Guo et al., 2017; Vaicenavicius et al., 2019). However, classifier evaluation often focuses on the most likely class (i.e., the mode of the predictive distribution) through the use of classification accuracy and related metrics derived from the confusion matrix (Tharwat, 2020; Hui and Belkin, 2021).

In this chapter, I propose *probabilistic top lists* as a way of producing probabilistic classifications in settings where specifying entire predictive distributions may be undesirable, impractical, or even impossible. While multi-label classification serves as a key example of such a setting, the theory presented here applies to classification in general. I envision the probabilistic top list approach to be particularly useful in settings eluding traditional probabilistic forecasting, where the specification of probability distributions on the full set of classes is hindered by a large number of classes and missing (total) order. Consistent evaluation is achieved through the use of proper scoring rules.

Whereas in traditional classification an instance is associated with a single class (e.g., *cat* or *dog*), multi-label classification problems (reviewed by Tsoumakas and Katakis, 2007; Zhang and Zhou, 2014; Tarekegn et al., 2021) admit multiple labels for an instance (e.g., *cat* or *dog* or *cat and dog*).[1] Applications of multi-label classification include text categorization (Zhang and Zhou, 2006), image recognition (Chen et al., 2019), and functional genomics (Barutcuoglu et al., 2006; Zhang and Zhou, 2006). Multi-label classification methods often output confidence scores for each label independently, and the final label set prediction is determined by a simple cut-off (Zhang and Zhou, 2014). As this approach does not account for label correlations, computing label set probabilities in a postprocessing step can improve predictions and probability estimates (Li et al., 2020) over simply multiplying probabilities to obtain label set probabilities. Probabilistic top lists offer a flexible approach to multi-label classification, which embraces the value of probabilistic information. In fact, the *BR-rerank* method introduced by Li et al. (2020) produces top list predictions. Yet, comparative performance evaluation focuses on (set) accuracy and the improper instance F1 score. This discrepancy has been a key motivation for this research.

In probabilistic forecasting, a scoring rule assigns a numerical score to a predictive distribution based on the true outcome (Gneiting and Raftery, 2007). It is proper if the expected score is optimized by the true distribution of the outcome of interest. Popular examples in classification are the Brier (or quadratic) score and the logarithmic (or cross-entropy) loss (Gneiting and Raftery, 2007; Hui and Belkin, 2021). When one is not interested in full predictive distributions, simple point predictions are frequently preferred. A meaningful point prediction admits interpretation in terms of a statistical functional (Gneiting, 2011a). Point predictions are evaluated by means of consistent scoring or loss functions. Similar to proper scoring rules, a scoring function is consistent for a functional if the expected score is optimized by the true functional value of the underlying distribution. For example, accuracy (or zero-one loss) is consistent for the mode in classification (Gneiting, 2017).

Probabilistic top lists bridge the gap between mode forecasts and full predictive distributions in classification. In this chapter, I define a probabilistic top-k list as a collection of k classes deemed most likely together with confidence scores quantifying the predictive probability associated with each of the k classes. The key question tackled in this chapter is how to evaluate such top list predictions consistently. To this end, I propose what I call *padded symmetric scores*, which are based on proper symmetric scoring rules. I show that the proposed padded symmetric scores are consistent for the probabilistic top-k list functional. The padded symmetric score of a probabilistic top list prediction is obtained from a symmetric proper scoring rule by padding the top list to obtain a full distribution. The padded distribution divides the probability mass not accounted for by the top list's confidence scores equally among the classes that are not included in the

[1] Multi-label classification is a special case of classification if classes are (re-)defined as subsets of labels.

list. Padded symmetric scores exhibit an interesting property, which allows for fair comparison of top lists of different length. Notably, the expected score of a correctly specified top list only depends on the top list itself and is invariant to other aspects of the true distribution. Comparability of top lists of differing length is ensured as the expected score does not deteriorate upon increasing the length of the predicted top list. Nonetheless, if the scoring function is based on the Brier score, there is little incentive to provide unreasonably large top lists.
The remaining chapter proceeds as follows. Section 5.2 recalls the traditional multi-class classification problem with a focus on probabilistic classification and suitable evaluation metrics. A quick introduction to the multi-label classification problem is also provided. Section 5.3 introduces probabilistic top lists, and related notation and terminology used throughout the chapter. Section 5.4 introduces some preliminary results on symmetric proper scoring rules and some results relating to the theory of majorization. These results are used in Section 5.5 to show that the padded symmetric scores yield consistent scoring functions for the top list functionals. Section 5.6 discusses the comparison of various types of predictions using the padded Brier and logarithmic scores. A theoretical argument as well as numerical examples illustrate that the padded Brier score is well suited for this task. Section 5.7 concludes the chapter.

5.2 Statistical Classification

The top list functionals and the proposed scoring functions are motivated by multi-label classification, but they apply to other classification problems as well. Here, I give a short formal introduction to the general classification problem and related evaluation metrics from the perspective of probabilistic forecasting.

5.2.1 Traditional multi-class classification

Recall from Chapter 2 that in the classical (multi-class) classification problem, one tries to predict the distinct class Y of an instance characterized by a vector of features $\mathbf{X}$. Formally, the outcome Y is a random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in the set of classes $\mathcal{Y}$ of cardinality $m \in \mathbb{N}$, and the feature vector $\mathbf{X}$ is a random vector taking values in some feature space $\mathcal{X} \subseteq \mathbb{R}^d$. Ideally, one learns the entire conditional distribution $p(\mathbf{X}) = \mathcal{L}(Y \mid \mathbf{X})$ of Y given $\mathbf{X}$ through a probabilistic classifier $c\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ mapping the features of a given instance to a probability distribution from the set of probability distributions $\mathcal{P}(\mathcal{Y})$ on $\mathcal{Y}$. The set $\mathcal{P}(\mathcal{Y})$ of probability distributions is typically identified with the probability simplex

$$\Delta_{m-1} = \{p \in [0,1]^m \mid p_1 + \cdots + p_m = 1\}$$

by (arbitrarily) labeling the classes as $1, \ldots, m$, and probability distributions are represented by vectors $p \in \Delta_{m-1}$, where the i-th entry p_i is the probability assigned to class i for $i = 1, \ldots, m$. To ease notation in what follows, vectors in

Δ_{m-1} are indexed directly by the classes in $\mathcal{Y}$ without explicit mention of any (re-)labeling.

Proper scoring rules quantify the value of a probabilistic classification and facilitate the comparison of multiple probabilistic classifiers (Gneiting and Raftery, 2007). A scoring rule is a mapping $\mathrm{S}\colon \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$, which assigns a, possibly infinite, score $\mathrm{S}(p, y)$ from the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ to a predictive distribution p if the true class is y. Typically, scores are negatively oriented in that lower scores are preferred. A scoring rule S is called *proper* if the true distribution p of Y minimizes the expected score,

$$\mathbb{E}[\mathrm{S}(p, Y)] \leq \mathbb{E}[\mathrm{S}(q, Y)] \quad \text{for } Y \sim p \text{ and } p, q \in \mathcal{P}(\mathcal{Y}). \tag{5.1}$$

It is *strictly proper* if the inequality (5.1) is strict unless $p = q$. Prominent examples are the logarithmic score

$$\mathrm{S}_{\log}(p, y) = -\log p_y \tag{5.2}$$

and the Brier score

$$\mathrm{S}_{\mathrm{B}}(p, y) = (1 - p_y)^2 + \sum_{z \neq y} p_z^2 = 1 - 2p_y + \sum_{z \in \mathcal{Y}} p_z^2. \tag{5.3}$$

Frequently, current practice does not focus on learning the full conditional distribution but, rather, on simply predicting the most likely class, i.e., the mode of the conditional distribution $p(\mathbf{X})$. This practice is formalized by a *hard* classifier $c\colon \mathcal{X} \to \mathcal{Y}$ aspiring to satisfy the functional relationship $c(\mathbf{X}) \in \mathrm{Mode}(p(\mathbf{X}))$, where the *mode functional* is given by

$$\mathrm{Mode}(p) = \arg\max_{y \in \mathcal{Y}} p_y = \{z \in \mathcal{Y} \mid p_z = \max_{y \in \mathcal{Y}} p_y\} \tag{5.4}$$

for $p \in \Delta_{m-1}$. Other functionals may be learned as well. When it comes to point forecasts of real-valued outcomes, popular choices are the mean or a quantile, see Chapter 4. Formally, a statistical functional $\mathrm{T}\colon \mathcal{P}(\mathcal{Y}) \to 2^{\mathcal{T}}$ reduces probability measures to certain facets in some space $\mathcal{T}$. Note that the functional T maps a distribution to a subset in the power set $2^{\mathcal{T}}$ of $\mathcal{T}$ owing to the fact that the functional value may not be uniquely determined. For example, the mode (5.4) of a distribution is not unique if multiple classes are assigned the maximum probability. The probabilistic top lists introduced in Section 5.3 are a nonstandard example of a statistical functional, which lies at the heart of this chapter.

Analogously to the evaluation of probabilistic classifiers through the use of proper scoring rules, predictions aimed at a statistical functional are evaluated by means of consistent scoring functions. Given a functional T, a scoring function is a mapping $\mathrm{S}\colon \mathcal{T} \times \mathcal{Y} \to \overline{\mathbb{R}}$, which assigns a score $\mathrm{S}(t, y)$ to a predicted facet t if the true class is y. A scoring function S is *consistent* for the functional T if the expected score is minimized by any prediction that is related to the true distribution of Y by the functional, i.e.,

$$\mathbb{E}[\mathrm{S}(t, Y)] \leq \mathbb{E}[\mathrm{S}(s, Y)] \quad \text{for } Y \sim p, t \in \mathrm{T}(p), p \in \mathcal{P}(\mathcal{Y}), \text{ and } s \in \mathcal{T}. \tag{5.5}$$

It is *strictly consistent* for T if the inequality (5.5) is strict unless $s \in \mathrm{T}(p)$. A functional T is called *elicitable* if a strictly consistent scoring function for T exists. For example, the mode (5.4) is elicited by the zero-one scoring function or misclassification loss (Gneiting, 2017)

$$\mathrm{S}(x, y) = \mathbb{1}\{x \neq y\},$$

which is simply a negatively oriented version of the ubiquitous classification accuracy measure. As discussed by Gneiting (2017) and references therein, decisions based on the mode are suboptimal if the losses invoked by different misclassifications are not uniform, which is frequently the case.
(Strictly) Proper scoring rules arise as a special case of (strictly) consistent scoring functions if T is the identity on $\mathcal{P}(\mathcal{Y})$. Furthermore, any consistent scoring function yields a proper scoring rule if predictive distributions are reduced by means of the respective functional first (Gneiting, 2011a, Theorem 3). On the other hand, a point prediction $x \in \mathcal{Y}$ can be assessed by means of a scoring rule as the classes can be embedded in the probability simplex by identifying a class $y \in \mathcal{Y}$ with the point mass $\delta_y \in \mathcal{P}(\mathcal{Y})$ in y. For example, applying the Brier score to a class prediction in this way yields twice the misclassification loss, $\mathrm{S_B}(x, y) = \mathrm{S_B}(\delta_x, y) = 2 \cdot \mathbb{1}\{x \neq y\}$.
Naturally, the true conditional distributions are unknown in practice, and expected scores are estimated by the mean score attained across all instances available for evaluation purposes.

5.2.2 Multi-label classification

In multi-label classification problems, an instance may be assigned multiple (class) labels. Here, I frame this problem as a special case of multi-class classification instead of an entirely different problem.
Let L be the set of labels and $\mathcal{Y} \subseteq 2^L$ be the set of label sets, i.e., classes are subsets of labels. In this setting, it may be difficult to specify a sensible predictive distribution on $\mathcal{Y}$, even for moderately sized sets of labels L, since the number of classes may grow exponentially with the number of labels. Extant comparative evaluation practices in multi-label classification focus mainly on hard classifiers ignoring the need for uncertainty quantification through probabilistic assessments (e.g., Tsoumakas and Katakis, 2007; Zhang and Zhou, 2014; Li et al., 2020; Tarekegn et al., 2021) with the exception of Read et al. (2011), who also consider a sum of binary logarithmic losses to evaluate the confidence scores associated with individual labels.
Classification accuracy is typically referred to as (sub-)set accuracy in multi-label classification. Other popular evaluation metrics typically quantify the overlap between the predicted label set and the true label set. For example, the comparative evaluation by Li et al. (2020) reports instance F1 scores in addition to set accuracy, where instance F1 of a single instance is defined as

$$\mathrm{S_{F1}}(x, y) = \frac{2 \sum_{\ell \in L} \mathbb{1}\{\ell \in x\} \mathbb{1}\{\ell \in y\}}{\sum_{\ell \in L} \mathbb{1}\{\ell \in x\} + \sum_{\ell \in L} \mathbb{1}\{\ell \in y\}}.$$

(and the overall score is simply the average across all instances as usual). Note that this metric is positively oriented, i.e., higher instance F1 scores are preferred. Caution is advised as the instance F1 score is not consistent for the mode, as illustrated by the following example. Hence, evaluating the same predictions using set accuracy and instance F1 seems to be a questionable practice.

Example 5.1. Let the label set $L = \{1, 2, 3, 4, 5\}$ consist of five labels and the set of classes $\mathcal{Y} = 2^L$ be the power set of the label set L. Consider the distribution $p \in \mathcal{P}(\mathcal{Y})$ that assigns all probability mass to four label sets as follows:

$$p(\{1, 2\}) = 0.28, \quad p(\{1, 3\}) = 0.24, \quad p(\{1, 4\}) = 0.24, \quad p(\{1, 5\}) = 0.24.$$

Then the expected instance F1 score of the most likely label set $\{1, 2\}$,

$$\mathbb{E}[\mathrm{S}_{\mathrm{F1}}(\{1, 2\}, Y)] = 0.64,$$

given $Y \sim p$ is surpassed by predicting only the single label $\{1\}$,

$$\mathbb{E}[\mathrm{S}_{\mathrm{F1}}(\{1\}, Y)] = \tfrac{2}{3}.$$

5.3 Probabilistic Top Lists

In what follows, I develop a theory informing principled evaluation of top list predictions based on proper scoring rules. To this end, a concise mathematical definition of probabilistic top lists is fundamental.
Let $k \in \{0, \ldots, m\}$ be fixed. A *(probabilistic) top-k list* is a collection $t = (\widehat{Y}, \widehat{t})$ of a set $\widehat{Y} \subset \mathcal{Y}$ of $k = \#\widehat{Y}$ classes together with a vector $\widehat{t} = (\widehat{t}_y)_{y \in \widehat{Y}} \in [0, 1]^k$ of *confidence scores* (or predicted probabilities) indexed by the set $\widehat{Y}$ whose sum does not exceed one, i.e., $\sum_{y \in \widehat{Y}} \widehat{t}_y \leq 1$, and equals one if $k = m$. Let $\mathcal{T}_k$ denote the set of probabilistic top-k lists. On the one hand, the above definition includes the empty top-0 list $t_\emptyset = (\emptyset, ())$ for technical reasons. At the other extreme, top-m lists specify entire probability distributions on $\mathcal{Y}$, i.e., $\mathcal{T}_m \equiv \mathcal{P}(\mathcal{Y})$. The *proxy probability*

$$\pi(t) := \frac{1 - \sum_{y \in \widehat{Y}} \widehat{t}_y}{m - k}$$

associated with a top-k list $t = (\widehat{Y}, \widehat{t}) \in \mathcal{T}_k$ of size $k < m$ is the probability mass not accounted for by the top list t divided by the number of classes not listed. For a top-m list $t \in \mathcal{T}_m$, the proxy probability $\pi(t) \equiv 0$ is defined to be zero. The *padded probability distribution* $\widetilde{t} = (\widetilde{t}_y)_{y \in \mathcal{Y}} \in \Delta_{m-1}$ associated with a probabilistic top-k list $t = (\widehat{Y}, \widehat{t}) \in \mathcal{T}_k$ assigns the proxy probability $\pi(t)$ to all classes not in $\widehat{Y}$, i.e.,

$$\widetilde{t}_y = \begin{cases} \widehat{t}_y, & \text{if } y \in \widehat{Y}, \\ \pi(t), & \text{if } y \notin \widehat{Y} \end{cases} \tag{5.6}$$

for $y \in \mathcal{Y}$.

A top-k list $t = (\widehat{Y}, \widehat{t})$ is *calibrated* relative to a distribution $p = (p_y)_{y\in\mathcal{Y}} \in \Delta_{m-1}$ if the confidence score $\widehat{t}_y$ of class y matches the true class probability p_y for all $y \in \widehat{Y}$. A top-k list $t = (\widehat{Y}, \widehat{t})$ is *true* relative to a distribution $p \in \mathcal{P}(\mathcal{Y})$ if it is calibrated relative to p and $\widehat{Y}$ consists of k most likely classes. There may be multiple true top-k lists for a given $k \in \mathbb{N}$ if the class probabilities are not pairwise distinct (i.e., if some classes have the same probability). References to the true distribution of the outcome Y are usually omitted in what follows. For example, a calibrated top list is understood to be calibrated relative to the distribution of Y. The *(probabilistic) top-k list functional* $\mathrm{T}_k \colon \mathcal{P}(\mathcal{Y}) \to \mathcal{T}_k$ maps any probability distribution $p \in \mathcal{P}(\mathcal{Y})$ to the set

$$\mathrm{T}_k(p) = \left\{ (\widehat{Y}, (p_y)_{y\in\widehat{Y}}) \in \mathcal{T}_k \;\middle|\; \widehat{Y} \in \operatorname*{arg\,max}_{S\subset\mathcal{Y}:|S|=k} \sum_{y\in S} p_y \right\}$$

of top-k lists that are true relative to p. The top-m list functional T_m identifies $\mathcal{P}(\mathcal{Y})$ with $\mathcal{T}_m$. A top-k list $t \in \mathcal{T}_k$ is *valid* if it is true relative to some probability distribution, i.e., there exists a distribution $p \in \mathcal{P}(\mathcal{Y})$ such that $t \in \mathrm{T}_k(p)$. Equivalently, a top-k list $t = (\widehat{Y}, \widehat{t})$ is valid if the associated proxy probability does not exceed the least confidence score, i.e., $\min_{y\in\widehat{Y}} \widehat{t}_y \geq \pi(t)$. Let $\widetilde{\mathcal{T}}_k \subset \mathcal{T}_k$ denote the set of valid top-k lists. The following is a simple example illustrating the previous definitions.

Example 5.2. Let $k = 2$, $m = 4$, $\mathcal{Y} = \{1, 2, 3, 4\}$, and $Y \sim p = (0.5, 0.2, 0.2, 0.1)$, i.e., $\mathbb{P}(Y = y) = p_y$. There are two true top-2 lists, namely, $T_2(p) = \{(\{1, 2\}, (0.5, 0.2)), (\{1, 3\}, (0.5, 0.2))\}$. The list $s = (\{1, 4\}, (0.5, 0.1))$ is calibrated (relative to p) but fails to be valid because it cannot be true relative to a probability distribution on $\mathcal{Y}$. On the other hand, the list $r = (\{1, 4\}, (0.5, 0.2))$ is valid as it is true relative to $q = (0.5, 0.2, 0.1, 0.2)$ but fails to be calibrated.

An invalid top-k list $t = (\widehat{Y}, \widehat{t})$ contains a *largest valid sublist* $t' = (\widehat{Y}', (\widehat{t}_y)_{y\in\widehat{Y}'})$. The largest valid sublist is uniquely determined by recursively removing the class $z \in \arg\min_{y\in\widehat{Y}} \widehat{t}_y$ with the lowest confidence score from the invalid list until a valid list remains. Removing a class $x \in \widehat{Y}$ with $\pi(t) > \widehat{t}_x$ cannot result in a valid top list $t' = (\widehat{Y} \setminus \{x\}, (\widehat{t}_y)_{y\in\widehat{Y}\setminus\{x\}})$ as long as there is another class z such that $\widehat{t}_x \geq \widehat{t}_z$ because $\pi(t) > \pi(t') > \widehat{t}_x \geq \widehat{t}_z$. Similarly, removing a class $x \in \widehat{Y}$ with $\pi(t) \leq \widehat{t}_x$ cannot prevent the removal of a class z if $\pi(t) > \widehat{t}_z$, because it does not decrease the proxy probability, $\pi(t') \geq p(t)$. Hence, *no* sublist containing a class with minimal confidence score in the original list is valid, and removal results in a superlist of the largest valid sublist.

In what follows, I show how to construct consistent scoring functions for the top-k list functional using proper scoring rules. Recall from Section 5.2.1 that a scoring function $\mathrm{S} \colon \mathcal{T}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$ is *consistent* for the top list functional T_k if the expected score under any probability distribution $p \in \mathcal{P}(\mathcal{Y})$ is minimized by any true top-k list $t \in \mathrm{T}_k(p)$, i.e.,

$$\mathbb{E}[\mathrm{S}(t, Y)] \leq \mathbb{E}[\mathrm{S}(s, Y)]$$

holds for $Y \sim p$ and any $s \in \mathcal{T}_k$. It is *strictly consistent* if the expected score is minimized only by the true top-k lists $t \in \mathrm{T}_k(p)$, i.e., the inequality is strict for $s \notin \mathrm{T}_k(p)$. The functional T_k is *elicitable* if a strictly consistent scoring function for T_k exists. In what follows, such a scoring function is constructed, giving rise to the following theorem.

Theorem 5.3 (Elicitability of the top list functional)**.** *The top-k list functional* T_k *is elicitable.*

Proof. The theorem is an immediate consequence of either Theorem 5.12 or 5.14. □

As the image of T_k is $\widetilde{\mathcal{T}}_k$ by definition, invalid top-k lists may be ruled out a priori, and the domain of S may be restricted to $\widetilde{\mathcal{T}}_k \times \mathcal{Y}$ in the above definitions. On the other hand, the use of a consistent scoring function on the larger domain $\mathcal{T}_k \times \mathcal{Y}$ merely encourages valid predictions, but it does not prohibit invalid predictions. Any scoring function that is consistent for valid top list predictions can be extended by assigning an infinite score to any invalid top list regardless of the observation. In a sense, this approach reconciles both points of view as an invalid prediction could not outperform any arbitrary valid prediction, thereby disqualifying it in comparison. In what follows, I focus on the construction of consistent scoring functions for valid top lists at first and propose a way of extending such scoring functions to invalid top lists that is less daunting than simply assigning an infinite score.

5.4 Mathematical Preliminaries

This section introduces some preliminary results, which are used heavily in the following section.

5.4.1 Symmetric scoring rules

The proposed scoring functions are based on symmetric proper scoring rules. Recall from Chapter 2 that (subject to mild regularity conditions) any proper scoring rule $\mathrm{S}\colon \mathcal{P}(\mathcal{Y}) \to \overline{\mathbb{R}}$ admits a *Savage representation*,

$$\mathrm{S}(p, y) = G(p) - \langle G'(p), p \rangle + G'_y(p), \tag{5.7}$$

in terms of a concave function $G\colon \Delta_{m-1} \to \mathbb{R}$ and a supergradient $G'\colon \Delta_{m-1} \to \mathbb{R}^m$ of G, i.e., a function satisfying the *supergradient inequality*

$$G(q) \leq G(p) + \langle G'(p), q - p \rangle \tag{5.8}$$

for all $p, q \in \Delta_{m-1}$. The function G is strictly concave if S is strictly proper. It is called the *entropy (function)* of S, which is simply the expected score $G(p) = \mathbb{E}[\mathrm{S}(p, Y)]$ under the posited distribution, $Y \sim p$. The supergradient inequality (5.8) is strict if G is strictly concave and $p \neq q$ (Jungnickel, 2015, Satz 5.1.12).

Let $\mathrm{Sym}(\mathcal{Y})$ denote the symmetric group on $\mathcal{Y}$, i.e., the set of all permutations of $\mathcal{Y}$. A scoring rule is called *symmetric* if scores are invariant under permutation of classes, i.e.,

$$\mathrm{S}((p_y), y) = \mathrm{S}((p_{\tau^{-1}(y)}), \tau(y))$$

holds for any permutation $\tau \in \mathrm{Sym}(\mathcal{Y})$ and all $y \in \mathcal{Y}, p \in \mathcal{P}(\mathcal{Y})$. Clearly, the entropy function G of a symmetric scoring rule is also symmetric, i.e., invariant to permutation in the sense that $G(p) = G((p_{\tau(y)}))$ holds for any permutation $\tau \in \mathrm{Sym}(\mathcal{Y})$ and any distribution $p \in \mathcal{P}(\mathcal{Y})$. Vice versa, any symmetric entropy function admits a symmetric proper scoring rule.

Proposition 5.4. *Let $G\colon \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Y})$ be a concave symmetric function. Then there exists a supergradient $G'\colon \mathcal{P}(\mathcal{Y}) \to \mathbb{R}^m$ such that the Savage representation* (5.7) *yields a symmetric scoring rule.*

Proof. Let $\bar{G}'$ be a supergradient of G. Using the shorthand $v_\tau = (v_{\tau^{-1}(y)})_{y\in\mathcal{Y}}$ for vectors $v = (v_y)_{y\in\mathcal{Y}} \in \mathbb{R}^m$ indexed by $\mathcal{Y}$ and permutations $\tau \in \mathrm{Sym}(\mathcal{Y})$, define G' by

$$G'(p) = \frac{1}{|\mathrm{Sym}(\mathcal{Y})|} \sum_{\tau\in\mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\tau^{-1}}(p_\tau)$$

for $p \in \mathcal{P}(\mathcal{Y})$. By symmetry of G and the supergradient inequality,

$$G(q) = G(q_\tau) \le G(p_\tau) + \langle \bar{G}'(p_\tau), q_\tau - p_\tau\rangle = G(p) + \langle \bar{G}'_{\tau^{-1}}(p_\tau), q - p\rangle$$

holds for all $p, q \in \mathcal{P}(\mathcal{Y})$ and $\tau \in \mathrm{Sym}(\mathcal{Y})$. Summation over all $\tau \in \mathrm{Sym}(\mathcal{Y})$ and division by the cardinality of the symmetric group $\mathrm{Sym}(\mathcal{Y})$ yields

$$G(q) \le \frac{1}{|\mathrm{Sym}(\mathcal{Y})|} \sum_{\tau\in\mathrm{Sym}(\mathcal{Y})} (G(p) + \langle \bar{G}'_{\tau^{-1}}(p_\tau), q - p\rangle) = G(p) + \langle G'(p), q - p\rangle$$

for any $p, q \in \mathcal{P}(\mathcal{Y})$. Therefore, G' is a supergradient, and the Savage representation (5.7) yields a symmetric scoring rule since

$$\begin{aligned}
G'(p) &= \frac{1}{|\mathrm{Sym}(\mathcal{Y})|} \sum_{\tau\in\mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\tau^{-1}}(p_\tau) = \frac{1}{|\mathrm{Sym}(\mathcal{Y})|} \sum_{\tau\in\mathrm{Sym}(\mathcal{Y})} \bar{G}'_{(\tau\circ\rho)^{-1}}(p_{\tau\circ\rho}) \\
&= \frac{1}{|\mathrm{Sym}(\mathcal{Y})|} \sum_{\tau\in\mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\rho^{-1}\circ\tau^{-1}}(p_{\tau\circ\rho}) = \frac{1}{|\mathrm{Sym}(\mathcal{Y})|} \sum_{\tau\in\mathrm{Sym}(\mathcal{Y})} (\bar{G}'_{\tau^{-1}}(p_{\tau\circ\rho}))_{\rho^{-1}} \\
&= \left(\frac{1}{|\mathrm{Sym}(\mathcal{Y})|} \sum_{\tau\in\mathrm{Sym}(\mathcal{Y})} \bar{G}'_{\tau^{-1}}((p_\rho)_\tau) \right)_{\rho^{-1}} = G'_{\rho^{-1}}(p_\rho)
\end{aligned}$$

and

$$\langle G'(p), p\rangle = \langle G'_{\rho^{-1}}(p_\rho), p\rangle = \langle G'(p_\rho), p_\rho\rangle$$

holds for any permutation $\rho \in \mathrm{Sym}(\mathcal{Y})$ and all $p \in \mathcal{P}(\mathcal{Y})$. □

On the other hand, not all proper scoring rules with symmetric entropy function are symmetric. The following result provides a necessary condition satisfied by supergradients of symmetric proper scoring rules.

Lemma 5.5. *Let* S *be a symmetric proper scoring rule. If* $p \in \Delta_{m-1}$ *satisfies* $p_x = p_z$ *for* $x, z \in \mathcal{Y}$*, then the supergradient* $G'(p) = (G'_y(p))_{y\in\mathcal{Y}}$ *at* p *in the Savage representation* (5.7) *satisfies* $G'_x(p) = G'_z(p)$*.*

Proof. Let $\tau = (x\ z)$ be the permutation swapping x and z while keeping all other classes fixed. Using notation as in the proof of Proposition 5.4, the equality $\mathrm{S}(p, x) = \mathrm{S}(p_\tau, \tau(x))$ holds by symmetry of S. Since $p = p_\tau$, the Savage representation (5.7) yields $G'_x(p) = G'_{\tau(x)}(p) = G'_z(p)$. □

The Brier score (5.3) and the logarithmic score (5.2) are both symmetric scoring rules. The entropy function of the Brier score is given by

$$G(p) = 1 - \sum_{y\in\mathcal{Y}} p_y^2, \tag{5.9}$$

whereas the entropy of the logarithmic score is given by

$$G(p) = -\sum_{y\in\mathcal{Y}} p_y \log(p_y)$$

(see Gneiting and Raftery, 2007).

5.4.2 Majorization and Schur-concavity

In this section, I adopt some definitions and results on majorization and Schur-concavity from Marshall et al. (2011). The theory of majorization is essentially a theory of inequalities, which covers many classical results and a plethora of mathematical applications not only in stochastics.
For a vector $v \in \mathbb{R}^m$, the vector $v_{[\,]} := (v_{[i]})_{i=1}^m$, where

$$v_{[1]} \geq \cdots \geq v_{[m]}$$

denote the components of v in decreasing order, is called the *decreasing rearrangement* of v. A vector $w \in \mathbb{R}^m$ is a permutation of $v \in \mathbb{R}^m$ (i.e., w is obtained by permuting the entries of v) precisely if $v_{[\,]} = w_{[\,]}$. For vectors $v, w \in \mathbb{R}^m$ with equal sum of components, $\sum_i v_i = \sum_i w_i$, the vector v is said to *majorize* w, or $v \succ w$ for short, if the inequality

$$\sum_{i=1}^k v_{[i]} \geq \sum_{i=1}^k w_{[i]}$$

holds for all $k = 1, \ldots, m-1$.
Let $D \subseteq \mathbb{R}^m$. A function $f\colon D \to \mathbb{R}$ is *Schur-concave on* D if $v \succ w$ implies $f(v) \leq f(w)$ for all $v, w \in D$. A Schur-concave function f is *strictly Schur-concave* if $f(v) < f(w)$ holds whenever $v \succ w$ and $v_{[\,]} \neq w_{[\,]}$. In particular, any

symmetric concave function is Schur-concave and strictly Schur-concave if it is strictly concave (Marshall et al., 2011, Chapter 3, Proposition C.2 and C.2.c). Hence, the following lemma holds.

Lemma 5.6. *The entropy function of any symmetric proper scoring rule is Schur-concave. It is strictly Schur-concave if the scoring rule is strictly proper.*

A set $D \subset \mathbb{R}^m$ is called *symmetric* if $v \in D$ implies $w \in D$ for all vectors $w \in \mathbb{R}^m$ such that $v_{[]} = w_{[]}$. By the Schur-Ostrowski criterion (Marshall et al., 2011, Chapter 3, Theorem A.4 and A.4.a) a continuously differentiable function $f\colon D \to \mathbb{R}$ on a symmetric convex set D with non-empty interior is Schur-concave if, and only if, f is symmetric and the partial derivatives $f_{(i)}(v) = \frac{\partial}{\partial v_i} f(v)$ increase as the components v_i of v decrease, i.e., $f_{(i)}(v) \leq f_{(j)}(v)$ if (and only if) $v_i \geq v_j$. Unfortunately, supergradients of concave functions do not share this property. The following is a slightly weaker condition, which applies to supergradients of symmetric concave functions.

Lemma 5.7 (Schur-Ostrowski condition for concave functions)**.** *Let $f\colon D \to \mathbb{R}$ be a symmetric concave function on a symmetric convex set D, $v \in D$ and $f'(v) = (f'_1(v), \dots, f'_m(v))$ be a supergradient of f at v, i.e., a vector satisfying the supergradient inequality*

$$f(w) \leq f(v) + \langle f'(v), w - v \rangle \tag{5.10}$$

for all $w \in D$. Then $v_i > v_j$ implies $f'_i(v) \leq f'_j(v)$.

Proof. For $i = 1, \dots, m$, let $e_i = (\mathbb{1}\{i = j\})_{j=1}^m$ denote the i-th vector of the standard basis of $\mathbb{R}^m$. Let $v \in D$ be such that $v_i > v_j$ for some indices i, j and let $0 < \varepsilon \leq v_i - v_j$. Define $w = v - \varepsilon e_i + \varepsilon e_j$. Then $v \succ w$ (by Marshall et al., 2011, Chapter 2, Theorem B.6) because w is obtained from v through a so-called 'T-transformation' (see Marshall et al., 2011, p. 32), i.e., $w_i = \lambda v_i + (1-\lambda) v_j$ and $w_j = \lambda v_j + (1-\lambda) v_i$ with $\lambda = \frac{v_i - v_j - \varepsilon}{v_i - v_j}$. Therefore, Schur-concavity of f implies $f(v) \leq f(w)$, and the supergradient inequality (5.10) yields

$$\varepsilon(f'_j(v) - f'_i(v)) = \langle f'(v), w - v \rangle \geq f(w) - f(v) \geq 0.$$

Hence, the inequality $f'_j(v) \geq f'_i(v)$ holds. □

With this result, there is no need to restrict attention to differentiable entropy functions when applying the Schur-Ostrowski condition in what follows. Furthermore, true top-k lists can be characterized using majorization.

Lemma 5.8. *Let $Y \sim p$ be distributed according to $p \in \mathcal{P}(\mathcal{Y})$. The padded distribution $\widetilde{t}$ associated with a true top-k list $t \in \mathrm{T}_k(p)$ majorizes the padded distribution $\widetilde{s}$ associated with any calibrated top-k list $s \in \mathcal{T}_k$.*

Proof. The sum of confidence scores $\sum_{i=1}^k \widetilde{t}_{[i]} = \sum_{i=1}^k p_{[i]} \geq \sum_{i=1}^k \widetilde{s}_{[i]}$ of a true top-k list is maximal among calibrated top-k lists by definition. Hence, the

confidence score $\widehat{t}_{[i]} = \widetilde{t}_{[i]}$ of the true top-k list $t = (\widehat{Y}, \widehat{t}\,)$ matches the i-th largest class probability $p_{[i]}$ for $i = 1, \ldots, k$. Therefore, the partial sums $\sum_{i=1}^{\ell} \widetilde{t}_{[i]} = \sum_{i=1}^{\ell} p_{[i]} \geq \sum_{i=1}^{\ell} \widetilde{s}_{[i]}$ across the largest confidence scores are also maximal for $\ell = 1, \ldots, k-1$. Furthermore, the proxy probability $\pi(t) = \frac{1-\sum_{i=1}^{k} \widetilde{t}_{[i]}}{m-k}$ associated with a true top-k list is minimal among calibrated top-k lists. Hence, the partial sums

$$\sum_{i=1}^{\ell} \widetilde{t}_{[i]} = 1 - (m-\ell)\pi(t) \geq 1 - (m-\ell)\pi(s) = \sum_{i=1}^{\ell} \widetilde{s}_{[i]}$$

are maximal for $\ell > k$. □

5.5 Consistent Top List Scores

Having reviewed the necessary preliminaries, this section shows that the proposed padded symmetric scores constitute a family of consistent scoring functions for the probabilistic top list functionals. The padded symmetric scores are defined for valid top lists and can be extended to invalid top lists by scoring the largest valid sublist, which yields a consistent scoring function. Strict consistency is preserved by adding an additional penalty term to the score of an invalid prediction.

5.5.1 Padded symmetric scores

From now on, let $\mathrm{S}\colon \mathcal{P}(\mathcal{Y}) \to \overline{\mathbb{R}}$ be a proper symmetric scoring rule with entropy function G. The scoring rule S is extended to valid top-k lists for $k = 0, 1, \ldots, m-1$ by setting

$$\mathrm{S}(t, y) := \mathrm{S}(\widetilde{t}, y)$$

for $y \in \mathcal{Y}, t \in \widetilde{\mathcal{T}}_k$, where $\widetilde{t} \in \Delta_{m-1}$ is the padded distribution (5.6) associated with the top-k list t. I call the resulting score $\mathrm{S}\colon \bigcup_{k=0}^{m} \widetilde{\mathcal{T}}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$ a *padded symmetric score*. For example, the logarithmic score (5.2) yields the *padded logarithmic score*

$$\mathrm{S}_{\log}((\widehat{Y}, \widehat{t}\,), y) = \begin{cases} -\log(\widehat{t}_y), & \text{if } y \in \widehat{Y}, \\ \log(m-k) - \log(1 - \sum_{z \in \widehat{Y}} \widehat{t}_z), & \text{otherwise,} \end{cases} \tag{5.11}$$

whereas the Brier score (5.3) yields the *padded Brier score*

$$\mathrm{S}_{\mathrm{B}}((\widehat{Y}, \widehat{t}\,), y) = 1 + \sum_{z \in \widehat{Y}} \widehat{t}_z^{\,2} + \frac{(1 - \sum_{z \in \widehat{Y}} \widehat{t}_z)^2}{m-k} - 2 \cdot \begin{cases} \widehat{t}_y, & \text{if } y \in \widehat{Y}, \\ \frac{1-\sum_{z \in \widehat{Y}} \widehat{t}_z}{m-k}, & \text{otherwise.} \end{cases} \tag{5.12}$$

The following example shows that padded symmetric scores should not be applied to invalid top lists without further considerations.

Example 5.9. If a padded symmetric score based on a strictly proper scoring rule is used to evaluate the invalid top-2 list s in Example 5.2, it attains a lower expected score than a true top list $t \in \mathrm{T}_2(p)$ because $\widetilde{s} = p$, whereas $\widetilde{t} \neq p$. Hence, the score would fail to be consistent.

The following lemma shows that the expected score of a calibrated top list is fully determined by the top list itself and does not depend on (further aspects of) the underlying distribution.

Lemma 5.10. *Let* S *be a padded symmetric score. If* $p \in \mathcal{P}(\mathcal{Y})$ *is the true distribution of* $Y \sim p$ *and* t *is a calibrated valid top list, then the expected score of the top list* t *matches the entropy of the padded distribution* $\widetilde{t}$*,*

$$\mathbb{E}[\mathrm{S}(t, Y)] = G(\widetilde{t}).$$

Proof. Let $t = (\widehat{Y}, \widehat{t}) \in \widetilde{\mathcal{T}}_k(p)$. Assume w.l.o.g. $k < m$ (the claim is trivial if $k = m$), and let $z \in \mathcal{Y} \setminus \widehat{Y}$. By Lemma 5.5 the supergradient at $\widetilde{t}$ satisfies $G'_y(\widetilde{t}) = G'_z(\widetilde{t})$ for all $y \notin \widehat{Y}$. Hence, the Savage representation (5.7) of the underlying scoring rule yields

$$\begin{aligned}
\mathbb{E}[\mathrm{S}(t, Y)] &= G(\widetilde{t}) - \langle G'(\widetilde{t}), \widetilde{t} \rangle + \sum_{y \in \mathcal{Y}} p_y G'_y(\widetilde{t}) \\
&= G(\widetilde{t}) - \sum_{y \in \widehat{Y}} (p_y - \widehat{t}_y) G'_y(\widetilde{t}_y) - \left(\sum_{y \notin \widehat{Y}} p_y - (m-k)\pi(t) \right) G'_z(\widetilde{t}) \\
&= G(\widetilde{t})
\end{aligned}$$

because t is calibrated. □

Padded symmetric scores exhibit an interesting property that admits fair comparison of top list predictions of varying length. A top list score $\mathrm{S}\colon \bigcup_{k=0}^{m} \widetilde{\mathcal{T}}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$ exhibits the *comparability property* if the expected score does not deteriorate upon extending a true top list, i.e., for $k = 0, 1, \ldots, m-1$ and any distribution $p \in \mathcal{P}(\mathcal{Y})$ of $Y \sim p$,

$$\mathbb{E}[\mathrm{S}(t_{k+1}, Y)] \leq \mathbb{E}[\mathrm{S}(t_k, Y)] \tag{5.13}$$

holds for $t_k \in \mathrm{T}_k(p)$ and $t_{k+1} \in \mathrm{T}_{k+1}(p)$. The following theorem shows that padded symmetric scores in fact exhibit the comparability property. I use the comparability property to show consistency of the individual padded symmetric top-k list scores $\mathrm{S}|_{\widetilde{\mathcal{T}}_k \times \mathcal{Y}}$ and to extend these scores to invalid top lists. Section 5.6 provides further discussion and some numerical insights.

Theorem 5.11. *Padded symmetric scores exhibit the comparability property.*

Proof. Let S be a padded symmetric score and G be the concave entropy function of the underlying proper scoring rule. Let $Y \sim p$ be distributed according to some distribution $p \in \mathcal{P}(\mathcal{Y})$, and let $t_k = (\widehat{Y}_k, (p_y)_{y \in \widehat{Y}_k})$ be a calibrated valid top-k list for some $k = 0, 1, \ldots, m-1$, which is extended by a calibrated valid top-$(k+1)$ list $t_{k+1} = (\widehat{Y}_{k+1}, (p_y)_{y \in \widehat{Y}_{k+1}})$ in the sense that $\widehat{Y}_{k+1} = \widehat{Y}_k \cup \{z\}$ for some $z \in \mathcal{Y}$. It is easy to verify that $\widetilde{t}_{k+1} \succ \widetilde{t}_k$ since $p_z \geq \pi(t_k) \geq \pi(t_{k+1})$. Hence, the inequality

$G(\widetilde{t}_{k+1}) \leq G(\widetilde{t}_k)$ holds by Schur-concavity of G (Lemma 5.6), which yields the desired inequality of expected scores by Lemma 5.10.
Clearly, there exists a true top-$(k+1)$ list $t_{k+1} \in \mathrm{T}_{k+1}(p)$ extending a true top-k list $t_k \in \mathrm{T}_k(p)$ in the above sense. By a symmetry argument, all true top lists of a given length have the same expected score, and hence S exhibits the comparability property. □

Note that the proof of Theorem 5.11 shows that (5.13) holds for any calibrated valid extension t_{k+1} of a calibrated valid top list t_k and not only for true top lists. I proceed to show that padded symmetric scores restricted to valid top-k lists are consistent for the top-k list functional.

Theorem 5.12. *Let $k \in \{0, 1, \ldots, m\}$ be fixed and $\mathrm{S}\colon \bigcup_{\ell=0}^{m} \widetilde{\mathcal{T}}_\ell \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a padded symmetric score. Then the restriction $\mathrm{S}\,|_{\widetilde{\mathcal{T}}_k \times \mathcal{Y}}$ of the score S to the set of valid top-k lists $\widetilde{\mathcal{T}}_k$ is consistent for the top-k list functional T_k. It is strictly consistent if the underlying scoring rule $\mathrm{S}\,|_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ is strictly proper.*

Proof. Let $p = (p_y)_{y \in \mathcal{Y}} \in \mathcal{P}(\mathcal{Y})$ be the true probability distribution of $Y \sim p$. Clearly, all true top-k lists in $\mathrm{T}_k(p)$ attain the same expected score by symmetry of the underlying scoring rule. Let $t = (\widehat{Y}, (p_y)_{y \in \widehat{Y}}) \in \mathrm{T}_k(p)$ be a true top-k list and $s = (\widehat{Z}, (\widehat{s}_y)_{y \in \widehat{Z}}) \in \widetilde{\mathcal{T}}_k$ be an arbitrary valid top-k list. To show consistency of $\mathrm{S}\,|_{\widetilde{\mathcal{T}}_k \times \mathcal{Y}}$, it suffices to show that the valid top-k list s does not attain a lower (i.e., better) expected score than the true top-k list t. Strict consistency follows if the expected score of any $s \notin \mathrm{T}_k(p)$ is higher than that of the true top-k list t. First, consider $s \notin \mathrm{T}_k(p)$ to be a calibrated top-k list, i.e., $\widehat{s}_y = p_y$ for all $y \in \widehat{Z}$. Since $\widetilde{t}$ majorizes $\widetilde{s}$ by Lemma 5.8, the inequality

$$\mathbb{E}[\mathrm{S}(t, Y)] = G(\widetilde{t}) \leq G(\widetilde{s}) = \mathbb{E}[\mathrm{S}(s, Y)]$$

holds by Schur-concavity of the entropy function G (Lemma 5.6) and Lemma 5.10. If the underlying scoring rule is strictly proper, the entropy function is strictly (Schur-)concave, and hence the inequality is strict.
Now, consider s to be an uncalibrated top-k list, and let $r = (\widehat{Z}, (p_y)_{y \in \widehat{Z}})$ be the respective calibrated top-k list on the same classes. The calibrated top-k list r may not be valid and cannot be scored if it is invalid. However, its largest valid sublist $r' = (\widehat{Z}', (p_y)_{y \in \widehat{Z}'})$ with $\widehat{Z}' \subseteq \widehat{Z}$ can be scored. Let $z \in \mathcal{Y} \setminus \widehat{Z}$. The difference in expected scores

$$\begin{aligned}
&\mathbb{E}[\mathrm{S}(s, Y)] - \mathbb{E}[\mathrm{S}(r', Y)] \\
&= G(\widetilde{s}) - G(\widetilde{r}') - \langle G'(\widetilde{s}), \widetilde{s} \rangle + \langle G'(\widetilde{r}'), \widetilde{r}' \rangle + \sum_{y \in \mathcal{Y}} p_y (G'_y(\widetilde{s}) - G'_y(\widetilde{r}')) \\
&\qquad\qquad \text{(by the Savage representation (5.7))} \\
&\geq \langle G'(\widetilde{r}') - G'(\widetilde{s}), \widetilde{r}' \rangle + \sum_{y \in \mathcal{Y}} p_y (G'_y(\widetilde{s}) - G'_y(\widetilde{r}')) \\
&\qquad\qquad \text{(by the supergradient inequality (5.8))}
\end{aligned}$$

$$= \sum_{y \in \widehat{Z} \setminus \widehat{Z}'} (p_y - \pi(r'))(G'_y(\widetilde{s}) - G'_z(\widetilde{r}')) + \sum_{y \in \mathcal{Y} \setminus \widehat{Z}} (p_y - \pi(r'))(G'_z(\widetilde{s}) - G'_z(\widetilde{r}'))$$

(by Lemma 5.5)

$$= \sum_{y \in \widehat{Z} \setminus \widehat{Z}'} (p_y - \pi(r'))(G'_y(\widetilde{s}) - G'_z(\widetilde{s}))$$

(as $\sum_{y \in \mathcal{Y} \setminus \widehat{Z}} (p_y - \pi(r')) = -\sum_{y \in \widehat{Z} \setminus \widehat{Z}'} (p_y - \pi(r'))$)

is nonnegative by the fact that $(p_y - \pi(r')) \leq 0$ for $y \in \widehat{Z} \setminus \widehat{Z}'$ (since r' is the largest valid sublist) and Lemma 5.7 (and Lemma 5.5 if $\widehat{s}_y = \pi(s)$ for some $y \in \widehat{Z} \setminus \widehat{Z}'$). Let $k' = |\widehat{Z}'|$. Then, r' scores no better than a true top-k' list $t_{k'} \in \mathrm{T}_{k'}(p)$, which in turn scores no better than t by the comparability property. Therefore,

$$\mathbb{E}[\mathrm{S}(s, Y)] \geq \mathbb{E}[\mathrm{S}(r', Y)] \geq \mathbb{E}[\mathrm{S}(t_{k'}, Y)] \geq \mathbb{E}[\mathrm{S}(t, Y)]$$

holds. If the underlying scoring function is strictly proper, the difference in expected scores $\mathbb{E}[\mathrm{S}(s, Y)] - \mathbb{E}[\mathrm{S}(r', Y)]$ above is strictly positive by strictness of the supergradient inequality (Jungnickel, 2015, Satz 5.1.12), and hence $\mathbb{E}[\mathrm{S}(t, Y)] < \mathbb{E}[\mathrm{S}(s, Y)]$ holds in this case, which concludes the proof. □

5.5.2 Penalized extensions of padded symmetric scores

The comparability property can be used to extend a padded symmetric score S to invalid top lists in a consistent manner. To this end, recall that t' denotes the largest valid sublist of a top list $t = (\widehat{Y}, \widehat{t}) \in \mathcal{T}_k$. Assigning the score of the largest valid sublist to an invalid top-k list yields a consistent score by the comparability property. Strict consistency of the padded symmetric score S is preserved by adding a positive penalty term $c_{\text{invalid}} > 0$ to the score of the largest valid sublist in the case of an invalid top list prediction. I call the resulting score extension $\mathrm{S}\colon \bigcup_{k=0}^{m} \mathcal{T}_k \times \mathcal{Y} \to \overline{\mathbb{R}}$, which assigns the score

$$\mathrm{S}(t, y) = \mathrm{S}(t', y) + c_{\text{invalid}} \tag{5.14}$$

to an invalid top list $t \in \mathcal{T}_k \setminus \widetilde{\mathcal{T}}_k$ for $k = 1, 2, \ldots, m-1$, a *penalized extension* of a padded symmetric score. The following example illustrates that the positive penalty is necessary to obtain a strictly consistent scoring function.

Example 5.13. Consider a setting similar to that of Example 5.2, where $Y \sim p = (0.4, 0.2, 0.2, 0.2)$. The padded distribution associated with the largest valid sublist $t' = (\{1\}, (0.4))$ of the invalid list $t = (\{1, 2\}, (0.4, 0.1))$ matches the true distribution, $\widetilde{t}' = p$, and hence the expected score of t in (5.14) is minimal if $c_{\text{invalid}} = 0$.

The following theorem summarizes the properties of the proposed score extension.

Theorem 5.14. *Let $k \in \{0, 1, \ldots, m\}$ be fixed and $\mathrm{S}\colon \bigcup_{\ell=0}^{m} \mathcal{T}_\ell \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a penalized extension* (5.14) *of a padded symmetric score with penalty term $c_{\text{invalid}} \geq$*

0. *Then the restriction* $\mathrm{S}\,|_{\mathcal{T}_k \times \mathcal{Y}}$ *of the score* S *to the set of top-k lists* $\mathcal{T}_k$ *is consistent for the top-k list functional* T_k*. It is strictly consistent if the underlying scoring rule* $\mathrm{S}\,|_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ *is strictly proper and the penalty term* c_{invalid} *is nonzero.*

Proof. In light of Theorem 5.12, it remains to show that an invalid top-k list attains a worse expected score than a true top-k list $t \in \mathrm{T}_k(p)$ under the true distribution $p \in \mathcal{P}(\mathcal{Y})$ of $Y \sim p$. To this end, let $s \in \mathcal{T}_k$ be invalid. By construction of the penalized extension, the top list s is assigned the score of its largest valid sublist s' plus the additional penalty c_{invalid}. By consistency of the padded symmetric score and the comparability property, the expected score of s' cannot fall short of the expected score of t. Hence, $\mathrm{S}\,|_{\mathcal{T}_k \times \mathcal{Y}}$ is consistent for the top-k list functional. If a positive penalty $c_{\text{invalid}} > 0$ is added, the score extension is strictly consistent given a strictly consistent padded symmetric score. □

5.6 Comparability

The comparability property (5.13) ensures that additional information provided by an extended true top list does not adversely influence the expected score. The information gain is quantified by a reduction in entropy, which depends on the underlying scoring rule. Ideally, a top list score encourages the prediction of classes that account for a substantial portion of probability mass, while offering little incentive to provide unreasonably large top lists. In what follows, I argue that the padded Brier score satisfies this requirement.

Let S be a padded symmetric score with entropy function G (of the underlying proper scoring rule). Furthermore, let $1 \leq k < m$ and $t = (\widehat{Y}, (\widehat{t}_y)_{y \in \widehat{Y}})$ be a top-k list that accounts for most of the probability mass. In particular, assume that the unaccounted probability $\alpha = \alpha(t) = 1 - \sum_{y \in \widehat{Y}} \widehat{t}_y$ is less than the least confidence score but nonzero, i.e.,

$$0 < \alpha < \min_{y \in \widehat{Y}} \widehat{t}_y. \tag{5.15}$$

Let $Q = Q(t) = \{p \in \mathcal{P}(\mathcal{Y}) \mid t \in \mathrm{T}_k(p)\}$ be the set of all probability measures relative to which t is a true top-k list. Let $p \in Q$ assign the remaining probability mass α to a single class. Then p majorizes any $q \in Q$, and the distribution p attains the lowest entropy, i.e., $G(p) = \min_{q \in Q} G(q)$, by Schur-concavity of the entropy function (Lemma 5.6). As the expected score of the top list t is invariant under distributions in Q by Lemma 5.10, the relative difference in expected scores between the true top list t and the true distribution $q \in Q$ is bounded by the relative difference in expected scores between t and p,

$$\frac{G(\widetilde{t}) - G(q)}{G(q)} \leq \frac{G(\widetilde{t}) - G(p)}{G(p)}.$$

The upper bound can be simplified by bounding the entropy of p from below as $G(p) \geq G((1 - \alpha, \alpha, 0, \ldots, 0))$ by Schur-concavity of G.

Table 5.1 Expected padded Brier scores and expected padded logarithmic scores of various types of true predictions and multiple distributions discussed in Example 5.15. Relative score differences (in percent) with respect to the optimal scores are in brackets.

			$\mathbb{E}[\mathrm{S}(\cdot, Y)]$		
p	S	Mode(p)	$\mathrm{T}_1(p)$	$\mathrm{T}_2(p)$	p
$p^{(\mathrm{h})}$	$\mathrm{S_B}$	0.02 (1.01%)	0.0199 (0.38%)	0.0198 (0%)	0.0198
$p^{(\mathrm{m})}$	$\mathrm{S_B}$	1 (70.59%)	0.6875 (17.28%)	0.5552 (0.04%)	0.5550
$p^{(\mathrm{l})}$	$\mathrm{S_B}$	1.5 (88.87%)	0.7969 (0.34%)	0.7955 (0.16%)	0.7942
$p^{(\mathrm{h})}$	$\mathrm{S_{log}}$	∞	0.0699 (24.75%)	0.0560 (0%)	0.0560
$p^{(\mathrm{m})}$	$\mathrm{S_{log}}$	∞	1.3863 (32.49%)	0.9425 (0.56%)	0.9373
$p^{(\mathrm{l})}$	$\mathrm{S_{log}}$	∞	1.6021 (0.45%)	1.5984 (0.23%)	1.5948

If $\mathrm{S} = \mathrm{S_B}$ is the padded Brier score (5.12) with entropy (5.9), the lower bound reduces to $G(p) \geq G((1-\alpha, \alpha, 0, \ldots, 0)) = 2(\alpha - \alpha^2) > \alpha$ since $\alpha < 0.5$ by assumption (5.15) and hence $2\alpha^2 < \alpha$. Therefore, the relative difference in expected scores has a simple upper bound,

$$\frac{G(\widetilde{t}) - G(p)}{G(p)} \leq \frac{\alpha^2 - \alpha\pi(t)}{2(\alpha - \alpha^2)} < \frac{\alpha^2}{\alpha} = \alpha.$$

For the padded logarithmic score, no such bound exists, and the deviation of the expected top list score from the optimal score can be severe, as illustrated in the following numerical example. The example sheds some light on the behavior of the (expected) padded symmetric scores and demonstrates that top lists of length $k > 1$ may provide valuable additional information over a simple mode prediction.

Example 5.15. Suppose there are $m = 5$ classes labeled $1, 2, \ldots, 5$ and the true (conditional) distribution $p = \mathcal{L}(Y \mid \mathbf{X} = \mathbf{x})$ of Y (given a feature vector $\mathbf{x} \in \mathcal{X}$) is known. Table 5.1 features expected padded Brier and logarithmic scores of various types of truthful predictions under several distributions, as well as relative differences with respect to the optimal score. The considered distributions

$$\begin{gathered} p^{(\mathrm{h})} = (0.99, 0.01, 0, 0, 0), \quad p^{(\mathrm{m})} = (0.5, 0.44, 0.03, 0.02, 0.01), \\ p^{(\mathrm{l})} = (0.25, 0.22, 0.2, 0.18, 0.15). \end{gathered} \tag{5.16}$$

exhibit varying degrees of predictability. Distribution $p^{(\mathrm{h})}$ exhibits high predictability in the sense that a single class can be predicted with high confidence. Distribution $p^{(\mathrm{m})}$ exhibits moderate predictability in that it is possible to narrow predictions down to a small subset of classes with high confidence, but getting the class exactly right is a matter of luck. Distribution $p^{(\mathrm{l})}$ exhibits low predictability in the sense that all classes may well realize. Predictions are of increasing

information content. The first prediction is the true mode, i.e., a hard classifier without uncertainty quantification that predicts class 1 under all considered distributions. The hard mode is interpreted as assigning all probability mass to the predicted class. Scores are obtained by embedding the predicted class in the probability simplex or, equivalent, by scoring the top-1 list $(\{1\}, 1)$. The second prediction is the true top-1 list $(\{1\}, p_1)$, i.e., the mode with uncertainty quantification. The third prediction is the true top-2 list $(\{1, 2\}, (p_1, p_2))$, and the final prediction is the true distribution p itself.
By consistency of the padded symmetric scores, the true top-1 lists score better in expectation than the mode predictions, and, by the comparability property, the true top-2 lists score better than the top-1 lists, while the true distributions attain the optimal scores. The mode predictions perform significantly worse than the probabilistic predictions, which highlights the importance of truthful uncertainty quantification. Note that the log score assigns an infinite score in cases where the true outcome is predicted as having zero probability, hence the mode prediction is assigned an infinite score with positive probability.
The expected padded Brier score of the probabilistic top-1 list under the highly predictable distribution $p^{(\mathrm{h})}$ is not far from optimal, whereas the respective logarithmic score is inflated by the discrepancies between the padded and true distributions, even though the top list accounts for most of the probability mass ($\alpha = 0.01$). Deviations from the optimal scores are more pronounced under the logarithmic score in all considered cases.
Under the distribution exhibiting moderate predictability, the top-2 list prediction is much more informative than the top-1 list prediction, which results in a significantly improved score that is not far from optimal. Under the distribution exhibiting low predictability, all probabilistic predictions perform well as there is little information to be gained.

Estimation of small probabilities is frequently hindered by finite sample size. The specification of top list predictions in conjunction with the padded Brier score circumvents this issue as the Brier score is driven by absolute differences in probabilities, whereas the logarithmic score emphasizes relative differences in probabilities. In other words, the padded distribution is deemed a good approximation of the true distribution if the true top list accounts for most of the probability mass by the Brier score.
In light of these considerations, I conclude that the padded Brier score is suitable for the comparison of top list predictions of varying length.

5.7 Concluding Remarks

In this chapter, I argued for the use of evaluation metrics rewarding truthful probabilistic assessments in classification. To this end, I introduced the probabilistic top list functionals, which offer a flexible probabilistic framework in the general classification problem. Padded symmetric scores yield consistent scoring functions, which admit comparison of various types of predictions. The padded

Brier score appears particularly suitable as top lists accounting for most of the probability mass obtain an expected padded Brier score that is close to optimal.
The entropy of a distribution is a measure of uncertainty or information content. Majorization provides a relation characterizing common decreases in entropy shared by all symmetric proper scoring rules. In particular, for two distributions $p \in \mathcal{P}(\mathcal{Y})$ and $q \in \mathcal{P}(\mathcal{Y})$, the entropy of the distribution p does not exceed the entropy of q, i.e., $G(p) \leq G(q)$, if p majorizes q. The inequality is strict if the scoring rule is strictly proper and q is not a permutation of p.
Typically, classes cannot simply be averaged, and combining multiple class predictions may be difficult as majority voting may result in a tie, while learning individual voting weights or a meta-learner requires training data (see Kotsiantis et al., 2006, Section 8.3 for a review of classifier combination techniques). Probabilistic top lists facilitate the combination of multiple predictions as confidence scores can simply be averaged, which may be an easy way to improve the prediction.
The prediction of probabilistic top lists appears particularly useful in problems, where classification accuracy is not particularly high, as is frequently the case in multi-label classification. I suspect that shifting focus towards probabilistic predictions may well increase prediction quality and usefulness in various decision problems, where misclassification losses are not uniform. Applying the proposed scores in a study with real predictions (e.g., the study conducted by Li et al. (2020)) is left as a topic for future work.

6 | Conclusion

While the research presented in this thesis was initially motivated by topics in forecast evaluation, this initial focus led to research on another topic in computational statistics, namely, the development of an improved algorithm for the computation of exact multinomial tests presented in Chapter 3. Besides the exposition and analysis of the algorithm, the chapter was aimed at providing a succinct overview on the topic of simple multinomial tests, which also resulted in an asymptotic chi-square approximation to the exact multinomial test based on what I called the probability mass statistic. The idea of ordering the sample space by the sample null probabilities is appealing. In the case of a uniform multinomial null hypothesis, this ordering yields an unbiased test, and in the case of a general multinomial null hypothesis, the test power appears to be comparable to the power of other popular multinomial tests.

Uncertainty quantification in the calibration simplex via multinomial p-values improves the visual representation of probabilistic forecasts for ternary outcomes. As the null distributions in each bin are merely approximated by multinomial distributions, resampling-based methods might be preferable. However, in my experience, the multinomial p-values provided good estimates, while offering a computationally fast approach. To provide some further examples, Figure 6.1 shows the exemplary calibration simplexes from Chapter 2 with added color-coded multinomial p-values. Of course, the p-values are not intended to be interpreted as strict statistical tests but, rather, as a measure of uncertainty relating the observed discrepancies between predicted probabilities and observed frequencies to the distribution of the empirical frequencies subject to calibration. A few small p-values are to be expected, yet if miscalibration results in certain trends observed across multiple bins, the p-values help to highlight the affected areas.

The hierarchies presented in Chapter 4 help to further differentiate between various notions of calibration for probabilistic forecasts in the case of real-valued outcomes and confirm a conjecture by Gneiting and Ranjan (2013) that has in part motivated this research. In the case of continuous, strictly increasing distributions the hierarchy shows that multiple notions of conditional calibration encountered in the literature coincide, which is not the case in general. Under minimal assumptions, there are some open questions as to the connection between the conditional notions of calibration. In particular, we conjectured that either CEP or quantile calibration implies threshold calibration, yet a proof has been elusive. I am confident that these implications can be confirmed in future work.

The chapter continues by studying a general notion of conditional T-calibration. T-Reliability diagrams and score decompositions are treated both from a theoretical point of view and as a practical tool for forecast verification. In empirical

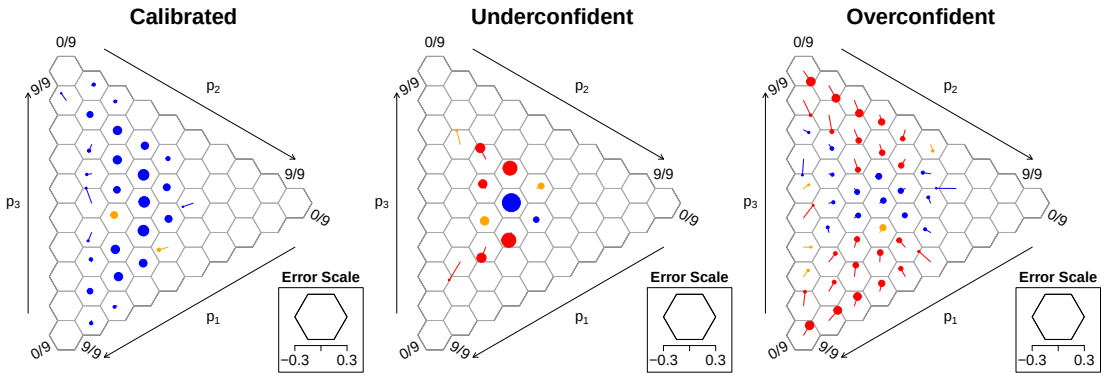


Figure 6.1 Calibration simplexes as in Figure 2.4 with color-coded p-values as introduced in Section 3.4.2.

settings, the use of empirical T-reliability diagrams that generalize the binary CORP reliability diagrams (Dimitriadis et al., 2021) to identifiable functionals was investigated. In contrast to the binary case, uncertainty quantification in general T-reliability diagrams is difficult to achieve without restrictive assumptions, and the consistency bands featured throughout Chapter 4 should be taken with a pinch of salt as the assumption of independent residuals may not hold true in practice. The approach outlined in Appendix 4.A.2 nonetheless provides a simple form of uncertainty quantification for general functionals T. Improvements, presumably, need to be tailored to specific functionals. In this regard, the recently proposed confidence bands for isotonic quantile curves by Dümbgen and Lüthi (2022) appear as a promising alternative to our consistency bands for quantile reliability diagrams, which should be explored in future work. The confidence bands for isotonic quantile curves are similar in spirit to confidence bands proposed by Dimitriadis et al. (2022a) for use in the binary CORP reliability diagram. Whereas consistency bands visualize typical deviations of the reliability curve from the diagonal subject to calibration, confidence bands quantify the estimator's variability and surround the estimated reliability curve. The confidence bands by Dimitriadis et al. (2022a) and Dümbgen and Lüthi (2022) are appealing as they ensure that the true reliability curve is entirely enclosed by the band with high confidence. Such 'simultaneous' confidence bands give rise to goodness-of-fit tests as the hypothesis of perfect calibration can be rejected if the band does not contain the diagonal in its entirety.

As an alternative to traditional statistical tests and p-values, interest in so-called e-values has surged in the statistical community (e.g., Ramdas et al., 2022). Recently, calibration tests based on e-values have been proposed for probability forecasts of binary events (Henzi and Ziegel, 2022; Dimitriadis et al., 2022c) and to assess probabilistic calibration of probabilistic forecasts (Arnold et al., 2021). Investigating the use of e-values in the context of multinomial tests (Lindon and Malek, 2020) and multi-class classifier calibration appears as a promising avenue for future research.

As an improvement on current evaluation practices, I propose the use of padded symmetric scores in Chapter 5, which conform to well-founded principles from the theory on forecast evaluation. The padded symmetric scores elicit probabilistic top lists giving rise to an overarching framework including various types of probabilistic predictions. I contend that the proposed probabilistic framework may help to improve current practices in multi-class classification and related forecasting problems. While the proposed approach has undergone in-depth theoretical treatment in this thesis, studies involving real data are a necessary next step to further substantiate its practical relevance.

Bibliography

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. Dover Publications, Inc., New York. Tenth Printing.

Allen, S. (2021). *Advanced Statistical Post-Processing of Ensemble Weather Forecasts*. PhD thesis, University of Exeter, UK.

Arnold, S. (2020). Isotonic distributional approximation. Master's thesis, Universität Bern, Switzerland.

Arnold, S., Henzi, A., and Ziegel, J. F. (2021). Sequentially valid tests for forecast calibration. Preprint, arXiv:2109.11761.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silvermann, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26:641–647.

Baglivo, J., Olivier, D., and Pagano, M. (1992). Methods for exact goodness-of-fit tests. *Journal of the American Statistical Association*, 87:464–469.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, New York.

Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22:830–836.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525:47–55.

Beals, R. and Wong, R. (2010). *Special Functions: A Graduate Text*, volume 126 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

Bejerano, G. (2006). Branch and bound computation of exact p-values. *Bioinformatics*, 22:2158–2159.

Bejerano, G., Friedman, N., and Tishby, N. (2004). Efficient exact p-value computation for small sample, sparse, and surprising categorical data. *Journal of Computational Biology*, 11:867–886.

Benjamin, S. G., Brown, J. M., Brunet, G., Lynch, P., Saito, K., and Schlatter, T. W. (2018). 100 years of progress in forecasting and NWP applications. *Meteorological Monographs*, 59:13.1 – 13.67.

Bentzien, S. and Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140:1924–1934.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75:1613–1669.

Breiman, L. (1992). *Probability*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, SIAM Classics edition.

Breitung, J. and Knüppel, M. (2021). How far can we forecast? Statistical tests of the predictive content. *Journal of Applied Econometrics*, 36:369–392.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Bröcker, J. and Ben Bouallègue, Z. (2020). Stratified rank histograms for ensemble forecast verification under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 146:1976–1990.

Bröcker, J. and Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22:651–661.

Casady, R. J. and Cryer, J. D. (1976). Monotone percentile regression. *Annals of Statistics*, 4:532–541.

Chafai, D. and Concordet, D. (2009). Confidence regions for the multinomial parameter with small sample size. *Journal of the American Statistical Association*, 104:1071–1079.

Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. (2021). Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *The Economic Journal*, 114:844–866.

Cohen, A. and Sackrowitz, H. B. (1975). Unbiasedness of the chi-square, likelihood ratio, and other goodness of fit tests for the equal cell case. *The Annals of Statistics*, 3:959–964.

Colonna, K. J., Nane, G. F., Choma, E. F., Cooke, R. M., and Evans, J. S. (2022). A retrospective assessment of COVID-19 model performance in the USA. *Royal Society Open Science*, 9:220021.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

Corradi, V. and Swanson, N. R. (2007). Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics*, 135:187–228.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46:440–464.

Cressie, N. and Read, T. R. C. (1989). Pearson's X^2 and the loglikelihood ratio statistic G^2: A comparative review. *International Statistical Review*, 57:19–43.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65:1254–1261.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A*, 147:278–292.

Dawid, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley-Interscience.

Dawid, A. P. (2016). Contribution to the discussion of "Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings" by W. Ehm, T. Gneiting, A. Jordan and F. Krüger. *Journal of the Royal Statistical Society Series B*, 78:505–562.

de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883.

Dimitriadis, T., Dümbgen, L., Henzi, A., Puke, M., and Ziegel, J. (2022a). Honest calibration assessment for binary outcome predictions. Preprint, arXiv:2203.04065.

Dimitriadis, T., Fissler, T., and Ziegel, J. F. (2022b). Osband's principle for identification functions. Preprint, arXiv:2208.07685.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences of the United States of America*, 118:e2016191118.

Dimitriadis, T., Henzi, A., Puke, M., and Ziegel, J. (2022c). A safe hosmer-lemeshow test. Preprint, arXiv:2203.00426.

Dimitriadis, T. and Jordan, A. I. (2021). reliabilitydiag: Reliability diagrams using isotonic regression. R package version 0.2.0, `https://cran.r-project.org/package=reliabilitydiag`.

Dümbgen, L. and Lüthi, L. (2022). Honest confidence bands for isotonic quantile curves. Preprint, arXiv:2206.13069.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B*, 78:505–562.

Ehm, W. and Ovcharov, E. Y. (2017). Bias-corrected score decomposition for generalized quantiles. *Biometrika*, 104:473–480.

El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic ordering constraint: The k-sample case. *Journal of the American Statistical Association*, 100:252–261.

Engel, E. (1857). Die vorherrschenden Gewerbszweige in den Gerichtsämtern mit Beziehung auf die Productions- und Consumtionsverhältnisse des Königreichs Sachsen. *Z. Stat. Bureau Königl. Sächs. Min. des Innern*, 8–9:153–82.

Engels, B. (2015). *XNomial: Exact goodness-of-fit test for multinomial data with fixed probabilities*. R package version 1.0.4 at `https://CRAN.R-project.org/package=XNomial`.

Ferro, C., Mitchell, K., and Bashaykh, H. (2020). Measuring forecast calibration. Workshop slides, VALPRED (Aussois, 12 March 2020), `empslocal.ex.ac.uk/people/staff/ferro/Presentations/FerroVALPRED2020.pdf`.

Fissler, T. and Holzmann, H. (2022). Measurability of functionals and of ideal point forecasts. Preprint, arXiv:2203.08635.

Fissler, T. and Pesenti, S. M. (2022). Sensitivity measures based on scoring functions. Preprint, `https://ssrn.com/abstract=4046894`.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44:1680–1706.

Fissler, T. and Ziegel, J. F. (2019). Order-sensitivity and equivariance of scoring functions. *Electronic Journal of Statistics*, 13:1166–1211.

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, Cambrige.

Galbraith, J. W. (2003). Content horizons for univariate time-series forecasts. *International Journal of Forecasting*, 19:43–55.

Galbraith, J. W. and Van Norden, S. (2012). Assessing gross domestic product and inflation probability forecasts derived from Bank of England fan charts. *Journal of the Royal Statistical Society Series A*, 175:713–727.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55:665–676.

Gibbons, J. D. and Pratt, J. W. (1975). P-values: Interpretation and methodology. *The American Statistician*, 29:20–25.

Gneiting, T. (2011a). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762.

Gneiting, T. (2011b). Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27:197–207.

Gneiting, T. (2017). When is the mode functional the Bayes classifier? *Stat*, 6:204–206.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69:243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Gneiting, T. and Resin, J. (2021). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams and coefficient of determination. Preprint, arXiv:2108.03210v3.

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., and Schienle, M. (2023). Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and Its Application*, 10. In press, DOI:10.1146/annurev-statistics-032921-020240.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science*, 33:568–594.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Gupta, C., Podkopaev, A., and Ramdas, A. (2020). Distribution-free binary classification: Prediction sets, confidence intervals and calibration. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Heinrich, C. (2021). On the number of bins in a rank histogram. *Quarterly Journal of the Royal Meteorological Society*, 147:544–556.

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, 66:1295–1305.

Henzi, A. and Ziegel, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*, 109:647–663.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society Series B*, 83:963–993.

Hirji, K. F. (1997). A comparison of algorithms for exact goodness-of-fit tests for multinomial data. *Communications in Statistics - Simulation and Computation*, 26:1197–1227.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting—with applications to risk management. *Annals of Applied Statistics*, 8:595–621.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society Series B*, 76:3–27.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

Hui, L. and Belkin, M. (2021). Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22:679–688.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, Chichester, second edition.

Jordan, A. I., Mühlemann, A., and Ziegel, J. F. (2022). Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals. *Annals of the Institute of Statistical Mathematics*, 74:489–514.

Julio, J. M. (2006). The fan chart: The technical details of the new implementation. Banco de la República Colombia Bogotá, Borradores de Economía, 468.

Jungnickel, D. (2015). *Optimierungsmethoden: Eine Einführung*. Springer Spektrum, Berlin, Heidelberg.

Keich, U. and Nagarajan, N. (2006). A fast and numerically robust method for exact multinomial goodness-of-fit test. *Journal of Computational and Graphical Statistics*, 15:779–802.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133:237–293.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33:270–281.

Koehler, K. J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, 75:336–344.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R. and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94:1296–1310.

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26:159–190.

Kotze, T. J. V. W. and Gokhale, D. V. (1980). A comparison of the Pearson-X^2 and log-likelihood-ratio statistics for small samples by means of probability ordering. *Journal of Statistical Computation and Simulation*, 12:1–13.

Krüger, F. and Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, 39:972–983.

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*.

Kvålseth, T. (1985). Cautionary note about R^2. *American Statistician*, 39:279–285.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Levi, D., Gispan, L., Giladi, N., and Fetaya, E. (2022). Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22:5540.

Li, C., Pavlu, V., Aslam, J., Wang, B., and Qin, K. (2020). Learning to calibrate and rerank multi-label predictions. In *Machine Learning and Knowledge Discovery in Databases*.

Lindon, M. and Malek, A. (2020). Anytime-valid inference for multinomial count data. Preprint, arXiv:2011.03567.

Malloy, M. L., Tripathy, A., and Nowak, R. D. (2021). Optimal confidence sets for the multinomial parameter. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2173–2178.

Marshall, A. W., Olkin, I., and Arnold, B. C. (2011). *Inequalities: Theory of Majorization and Its Applications*. Springer Series in Statistics. Springer, New York, second edition.

Mason, S. J., Galpin, J. S., Goddard, L., Graham, N. E., and Rajartnam, B. (2007). Conditional exceedance probabilities. *Monthly Weather Review*, 135:363–372.

McDonald, J. H. (2009). *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, second edition.

Menzel, U. (2013). *EMT: Exact multinomial test: Goodness-of-fit test for discrete multivariate data*. R package version 1.1 at `https://CRAN.R-project.org/package=EMT`.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50:885–900.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14:24–49.

Murota, K. (2003). *Discrete Convex Analysis*. SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Murota, K. and Shioura, A. (2003). Quasi M-convex and L-convex functions - quasiconvexity in discrete optimization. *Discrete Applied Mathematics*, 131:467–494.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105:803–816.

Murphy, A. H. (1996). General decomposition of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Monthly Weather Review*, 124:2353–2369.

Murphy, A. H. and Epstein, E. S. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117:572–581.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4:133–142.

Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I – A discussion of principles. *Journal of Hydrology*, 10:282–290.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *Proceedings of Computer Vision and Pattern Recognition (CVPR) Conference Workshops*.

Noceti, P., Smith, J., and Hodges, S. (2003). An evaluation of tests of distributional forecasts. *Journal of Forecasting*, 22:447–455.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11:1833–1874.

Orjebin, E. (2014). A recursive formula for the moments of a truncated univariate normal distribution. Working paper, `https://people.smp.uq.edu.au/YoniNazarathy/teaching_projects/studentWork/EricOrjebin_TruncatedNormalMoments.pdf`.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 38:796–809.

Pérez, T. and Pardo, J. A. (2003). On choosing a goodness-of-fit test for discrete multivariate data. *Kybernetes*, 32:1405–1424.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P.,

Franses, P. H., Frazier, D. T., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Trapero Arenas, J. R., Wang, X., Winkler, R. L., Yusupova, A., and Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38:705–871.

Pinson, P. and Hagedorn, R. (2012). Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorological Applications*, 19:484–500.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. J. (2017). On fairness and calibration. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*.

Pohle, M.-O. (2020). The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. Preprint, arXiv:2005.01835.

Radlow, R. and Alf, E. F. J. (1975). An alternate multinomial assessment of the accuracy of the χ^2 test of goodness of fit. *Journal of the American Statistical Association*, 70:811–813.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481.

Rahmann, S. (2003). Dynamic programming algorithms for two statistical problems in computational biology. In *Algorithms in Bioinformatics. WABI 2003. Lecture Notes in Computer Science*, volume 2812, pages 151–164. Springer, Berlin, Heidelberg.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2022). Game-theoretic statistics and safe anytime-valid inference. Preprint, arXiv:2210.01948.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85.

Resin, J. (2020). *ExactMultinom: Multinomial Goodness-of-Fit Tests*. R package version 0.1.2 at `https://CRAN.R-project.org/package=ExactMultinom`.

Resin, J. (2021a). *CalSim: The calibration simplex*. R package version 0.5.2 at `https://CRAN.R-project.org/package=CalSim`.

Resin, J. (2021b). *Replication code for Gneiting and Resin (2021)*. `https://github.com/resinj/replication_GR21`.

Resin, J. (2022). A simple algorithm for exact multinomial tests. *Journal of Computational and Graphical Statistics*. In press, DOI:10.1080/10618600.2022.2102026.

Robertson, T. and Wright, F. T. (1980). Algorithms in order restricted statistical inference and the Cauchy mean value property. *Annals of Statistics*, 8:645–651.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. (2022). Mitigating bias in calibration error estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139:3921–3927.

Rüschendorf, L. and de Valk, V. (1993). On regression representations of stochastic processes. *Stochastic Processes and their Applications*, 46:183–198.

Sahoo, R., Zhao, S., Chen, A., and Ermon, S. (2021). Reliable decisions with threshold calibration. In *Advances in Neural Information Processing Systems*.

Satopää, V. and Ungar, L. (2015). Combining and extremizing real-valued forecasts. Preprint, arXiv:1506.06405.

Satopää, V. A. (2021). Improving the wisdom of crowds with analysis of variance of predictions of related outcomes. *International Journal of Forecasting*, 37:1728–1747.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801.

Schmidt, K. D. (2011). *Maß und Wahrscheinlichkeit*. Springer, Heidelberg, revised edition.

Sen, B., Banerjee, M., and Woodroofe, M. (2010). Inconsistency of bootstrap: The Grenander estimator. *Annals of Statistics*, 38:1953–1977.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, SIAM Classics edition.

Song, H., Diethe, T., Kull, M., and Flach, P. (2019). Distribution calibration for regression. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Steinwart, I., Pasin, C., Williamson, R., and Zhang, S. (2014). Elicitation and identification of properties. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 35:1–45.

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A., and Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, 354:1240–1241.

Stoyanov, J. (2000). Krein condition in probabilistic moment problems. *Bernoulli*, 6:939–949.

Strähl, C. and Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, 11:608–639.

Taggart, R. (2022). Point forecasting and forecast evaluation with generalized Huber loss. *Electronic Journal of Statistics*, 16:201–231.

Tarekegn, A. N., Giacobini, M., and Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Tate, M. W. and Hyer, L. A. (1973). Inaccuracy of the X^2 test of goodness of fit when expected frequencies are small. *Journal of the American Statistical Association*, 68:836–841.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106:7183–7192.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17:168–192.

Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B. (2021). A practical guide to selecting nodels for exploration, inference, and prediction in ecology. *Ecology*, 102:e03336.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3:1–13.

Tsyplakov, A. (2011). Evaluating density forecasts: A comment. Preprint, DOI:10.2139/ssrn.1907799.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: Proper scoring rules and moments. Preprint, DOI:10.2139/ssrn.2236605.

Vaicenavicius, J., Widmann, D., C., A., Lindsten, F., Roll, J., and Schön, T. B. (2019). Evaluating model calibration in classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., and Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Van Eeden, C. (1958). *Testing and Estimating Ordered Parameters of Probability Distributions*. PhD thesis, University of Amsterdam, Netherlands.

Wakimoto, K., Odaka, Y., and Kang, L. (1987). Testing the goodness of fit of the multinomial distribution based on graphical representation. *Computational Statistics & Data Analysis*, 5:137–147.

Wallis, K. F. (2003). Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting*, 19:165–175.

West, E. N. and Kempthorne, O. (1972). A comparison of the chi^2 and likelihood ratio tests for composite alternatives. *Journal of Statistical Computation and Simulation*, 1:1–33.

Wilks, D. S. (2013). The calibration simplex: A generalization of the reliability diagram for three-category probability forecasts. *Weather and Forecasting*, 28:1210–1218.

Wilks, D. S. (2019). Indices of rank histogram flatness and their sampling properties. *Monthly Weather Review*, 147:763–769.

Wolffram, D., Resin, J., Kraus, K., and Jordan, A. I. (2022). *Replication package for "Model Diagnostics and Forecast Evaluation for Quantiles"*. DOI:10.5281/zenodo.6546490.

Wright, F. T. (1984). The asymptotic behavior of monotone regression estimates. *Canadian Journal of Statistics*, 12:229–236.

Yu, B. and Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences of the United States of America*, 117:3920–3929.

Zhang, M.-L. and Zhou, Z.-H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18:1338–1351.

Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26:1819–1837.

Zhao, S., Ma, T., and Ermon, S. (2020). Individual calibration with randomized forecasting. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Ziegel, J. F. (2016). Contribution to the discussion of "Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings" by W. Ehm, T. Gneiting, A. Jordan and F. Krüger. *Journal of the Royal Statistical Society Series B: Methodological*, 78:505–562.