Dovern, Jonas; Glas, Alexander; Kenny, Geoff

Article — Published Version

Testing for differences in survey-based density expectations: A compositional data approach

Journal of Applied Econometrics

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Dovern, Jonas; Glas, Alexander; Kenny, Geoff (2024) : Testing for differences in survey-based density expectations: A compositional data approach, Journal of Applied Econometrics, ISSN 1099-1255, Wiley, Hoboken, NJ, Vol. 39, Iss. 6, pp. 1104-1122, https://doi.org/10.1002/jae.3080

This Version is available at:
https://hdl.handle.net/10419/306086

Received: 4 June 2023 | Revised: 11 January 2024 | Accepted: 29 April 2024

DOI: 10.1002/jae.3080

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Testing for differences in survey-based density expectations: A compositional data approach

Jonas Dovern[1] | **Alexander Glas**[2] | **Geoff Kenny**[3]

[1]School of Business, Economics and Society, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nuremberg, Germany

[2]Pensions and Sustainable Financial Markets, ZEW – Leibniz Centre for European Economic Research, Mannheim, Germany

[3]European Central Bank, Frankfurt am Main, Germany

Correspondence
Jonas Dovern, School of Business, Economics and Society, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lange Gasse 20, 90403 Nuremberg, Germany.
Email: jonas.dovern@fau.de

Summary

We propose to treat survey-based density expectations as compositional data when testing either for heterogeneity in density forecasts across different groups of agents or for changes over time. Monte Carlo simulations show that the proposed test has more power relative to both a bootstrap approach based on the KLIC and an approach that involves multiple testing for differences of individual parts of the density. In addition, the test is computationally much faster than the KLIC-based one, which relies on simulations, and allows for comparisons across multiple groups. Using density expectations from the ECB Survey of Professional Forecasters and the US Survey of Consumer Expectations, we show the usefulness of the test in detecting possible changes in density expectations over time and across different types of forecasters.

KEYWORDS

compositional data, density forecasts, disagreement, survey forecasts

1 | INTRODUCTION

Expectations are central both for microeconomic decision-making and for macroeconomic dynamics. Hence, it is not surprising that a large body of literature studies the properties of expectations in various economic contexts. In recent years, the use of survey-based expectation data has become increasingly more common. This process is particularly strong in macroeconomics where more and more surveys are set up to study the macroeconomic expectations of private households (Conrad et al., 2022; Coibion et al., 2024; D'Acunto et al., 2021; Kim & Binder, 2023), firms (Andrade et al., 2022; Coibion et al., 2018, 2020; Dovern et al., 2023; Kumar et al., 2023), and professional forecasters (Glas & Hartmann, 2022; Rich & Tracy, 2021).

We observe two tendencies in this literature that we want to bring together in our paper. On the one hand, there is a growing focus on understanding the reasons for and the effects of heterogeneity of macroeconomic expectations at least since the seminal contribution by Mankiw et al. (2003). On the other hand, there is a tendency toward the analysis of probabilistic (density) expectations that offer a more complete picture of expectations relative to conventional point expectations (Manski, 2018). What is missing, so far, are major efforts to combine these two important aspects. Rich and Tracy (2021) use the Wasserstein distance as a measure of heterogeneity in experts' density forecasts. However, they do not compare distinct groups of forecasters and do not test for statistical differences. Mitchell and Hall (2005) and Clements (2018) use Diebold–Mariano-type tests based on the Kullback Leibler Information Criterion (KLIC) as suggested by Bao et al. (2007). Mitchell and Hall (2005) compare the density forecasts ("fan charts") for inflation in the United Kingdom reported by the Bank of England and the National Institute of Economic and Social Research and find

that the former tend to be more accurate than the latter. Clements (2018) tests whether experts' density forecasts outperform unconditional benchmark densities. While this approach could be used to test for (cross-sectional) expectation heterogeneity, both studies rely on time series variation in forecast performance to compare different density expectations.

Our paper's contribution is to suggest a method that can be used to test for heterogeneity of probabilistic expectations. Such probabilistic expectations are usually elicited by asking agents to assign probabilities to intervals of potential outcomes. Our central insight is that these vectors of probabilities are compositional data. Compositional data consist of vectors of proportions (here: probabilities) that are subject to the constraint that the sum of all elements must equal a fixed value (here: one); see Aitchison (1982, 1986). We propose to use tests that have been developed for compositional data to test for differences in probabilistic expectations across different groups of individuals or survey waves.

Using Monte Carlo simulations, we compare the proposed compositional approach to an alternative bootstrap-based approach that builds on the more traditional way of comparing (expectation) distributions using distance measures such as the KLIC. The results from the Monte Carlo simulations suggest that the tests designed for compositional data have higher power against alternatives that imply only moderate differences in density expectations between two subpopulations, especially when the sample size is relatively small. Moreover, our proposed test is much faster because it does not require simulations due to the fact that the distribution of the test statistic is known under the null hypothesis. In addition, the test allows for a joint comparison of multiple groups, whereas the KLIC-based approach can only be used to compare two groups at a time.

We then apply the method to five different research questions that have recently been discussed in the literature. Specifically, we test for heterogeneity of expectations and changes over time among different groups of professional macroeconomic forecasters (based on data from the European Central Bank's Survey of Professional Forecasters, SPF) and private households (based on data from the Federal Reserve Bank of New York's Survey of Consumer Expectations, SCE). We show that (i) professional forecasters quickly changed their short-term inflation density forecasts in response to the recent period of rising inflation rates, whereas long-term expectations reacted more gradually, (ii) in most survey waves, both inflation and GDP growth expectations differ significantly between experts that round their probability statements and those that do not, (iii) there is strong evidence against the hypothesis that private households' inflation expectations reported by men and women are equal, (iv) households learn about concepts they are initially unfamiliar with—such as inflation—via repeated survey participation as recently highlighted by Kim and Binder (2023), and (v) for a sizeable fraction of periods in our sample, households from different regions report significantly different density expectations for the future change of nationwide house prices.

First and foremost, our work relates to other studies that analyze heterogeneity of macroeconomic expectations across individuals or firms. For example, Malmendier and Nagel (2011, 2016) show that US households who experienced low stock market returns and/or high inflation rates during their lifetime tend to be more pessimistic with respect to future stock market developments and/or inflation than individuals with more moderate lifetime experiences. Similarly, Kuchler and Zafar (2019) find that local house price experiences affect households' expectations about future house price changes. In particular, the experience of volatile house prices leads to a higher dispersion of house price expectations. With respect to firm expectations, Kumar et al. (2015) find that the inflation expectations of firm managers in New Zealand are heavily dispersed, at odds with the notion of anchored or fully rational expectations. A common feature shared by these studies is that they focus on point forecasts. Our contribution is to provide methods that allow us to analyze heterogeneity of density expectations and, thus, to move beyond the analysis of heterogeneity of point expectations.

In terms of the methodology used, our work relates to—and borrows heavily from—the literature on compositional data. Aitchison (1986) and Filzmoser et al. (2018) offer comprehensive overviews of methodological aspects that are important when dealing with such data. The methods are widely applied in many disciplines, including geochemistry (e.g., Buccianti, 2018; Reimann et al., 2012), sedimentology (Weltje & von Eynatten, 2004), demography (Lloyd et al., 2012), and medicine (Braga & Feingenbaun, 2020; Kitano et al., 2020). In economics, methods for compositional data have been used, for instance, to analyze income or expenditure shares (Fry et al., 1996) and how time budgets are shared for different activities (Gupta et al., 2020). Our contribution is to show that these methods are also relevant and helpful when dealing with probabilistic expectations.

The rest of this paper is structured as follows. Section 2 briefly summarizes the basics of compositional data and describes the tests that we propose to use for the analysis of heterogeneity and temporal stability in probabilistic expectations. Section 3 presents the results from the Monte Carlo simulations that we use to assess the properties of the tests. Section 4 describes the applications of the proposed method. Section 5 concludes.

2 | METHODOLOGY

We consider survey-based probabilistic density expectations reported by individuals $i = 1, \dots, N$ at time $t = 1, \dots, T$ for some future (macroeconomic) outcome in period $t + h$ so that h indicates the forecast horizon. In practice, these density expectations are most commonly elicited by asking subjects to assign probabilities to a set of K different outcome intervals (or "bins"). The assigned values indicate the probabilities by which subjects expect the outcome to fall into the corresponding intervals. Hence, each probabilistic density expectation is reported in the form of a histogram that is represented by a vector $\mathbf{p}_{i,t,h} = (p_{i,t,h,1}, \dots, p_{i,t,h,K})'$ with nonnegative elements $p_{i,t,h,k}$ for $k = 1, \dots, K$. Since the union of all intervals covers the entire outcome space, a natural constraint (which is usually enforced by the survey design) is that $p_{i,t,h,1} + p_{i,t,h,2} + \dots + p_{i,t,h,K} = 1$.

We are interested in the following problem: Given two sets of density forecasts from different groups of forecasters, denoted as $g \in \{A, B\}$, we want to test—based on the sample of available histograms—the null hypothesis that individuals from both groups draw their probabilistic expectations from the same distribution. More formally, let $\left(\mu^g_{t,h,1}, \mu^g_{t,h,2}, \cdot; \cdot \mu^g_{t,h,K}\right)' = \boldsymbol{\mu}^g_{t,h} = \mathbf{E}\left(\mathbf{p}^g_{i,t,h}\right)$ denote the expected value of the vector of interval probabilities for any individual from group g.[1] Note that the definition and the number of intervals, K, need to be the same for both groups because the test is going to compare vectors that need to be of the same length. Usually, this is given automatically because expectations for both groups come from the same survey. Our null hypothesis then is H_0: $\boldsymbol{\mu}^A_{t,h} = \boldsymbol{\mu}^B_{t,h}$ against the alternative hypothesis that H_1: $\boldsymbol{\mu}^A_{t,h} \neq \boldsymbol{\mu}^B_{t,h}$. We want to test this hypothesis about the two population moments using two samples of observed histogram forecasts of size N_A and N_B (with $N_A + N_B = N$).[2]

In the following subsections, we first propose a test for analyzing differences of histogram forecasts across groups of individuals that we borrow from the literature on compositional data. We then describe two alternative approaches. The first alternative breaks down our null hypothesis into K interval-specific testable hypotheses and uses the Bonferroni correction to control for the size of the test of the primary null hypothesis. The second alternative is a bootstrap-based approach that is based on the traditional way of measuring the dissimilarity between two distributions using the KLIC.

2.1 | Compositional data approach

Treating expectations of the form considered in this paper as compositional data starts from the insight that the sum of all probabilities must equal one. With respect to the statistical modeling of the distribution of the vectors of probabilities $\mathbf{p}_{i,t,h}$, this implies that the sample space is not simply the K-dimensional space of nonnegative real numbers $\mathbb{R}^K_+$ but the so-called $K-1$-dimensional *simplex* defined by

$$\mathbb{S}^{K-1} = \{(p_{i,t,h,1}, \dots, p_{i,t,h,K}) : p_{i,t,h,1} \geq 0, \dots, p_{i,t,h,K} \geq 0;\ p_{i,t,h,1} + \dots + p_{i,t,h,K} = 1\}. \tag{1}$$

Failing to take account of this—by applying "standard" statistical methods—will lead to various problems, including problematic interpretation of the covariance of the interval probabilities (see Aitchison, 1986, chapter 3).

Instead, one needs to apply a proper one-to-one transformation that leads to a vector of random variables that one can handle more easily. Commonly, the *additive logratio transformation* is used, and we adopt this choice in our paper, too.[3] Choosing the Kth probability as the reference (without loss of generality because it does not matter which element of $\mathbf{p}_{i,t,h}$ is used), the transformed expectation data is given by

[1] In economic terms, this expectation denotes the consensus expectation in the population, that is, the average expectation that reflects common information after all idiosyncratic factors—such as private information, different priors, and different expectation-formation models—has been integrated out.

[2] Instead of working with the survey histograms, one could fit a parametric distribution to the survey probabilities, for example, a normal distribution (Giordani & Söderlind, 2003), the generalized beta distribution (Engelberg et al., 2009), or a skew t distribution (Ganics et al., 2024). However, choosing a particular parametric distribution is difficult due to the large degree of heterogeneity frequently observed in survey data. This is especially true for household surveys that also suffer from irregularities such as bimodal responses and disjointed sets of nonzero probabilities. A major advantage of the approaches discussed in this paper is that they do not require any parametric assumptions or fitting procedure.

[3] This choice implies that we require strictly positive probabilities to simplify the analysis. Hence, we replace all zero entries by very small (random) numbers in the applications and adjust the other entries accordingly to ensure that the unit constraint is still met. Given that rounding is an eminent feature of survey-based probabilistic expectations (Binder, 2017; Clements, 2021; Glas & Hartmann, 2022; Reiche & Meyler, 2022), treating zero entries as rounded approximations of small probabilities seems a reasonable modeling choice. Martín-Fernández et al. (2003) and Filzmoser et al. (2018) discuss various strategies of dealing with zeroes and missing values in compositional data.

$$\tilde{p}_{i,t,h,k} = \ln\left(\frac{p_{i,t,h,k}}{p_{i,t,h,K}}\right) \qquad \text{for } k = 1, \ldots, K-1. \tag{2}$$

This transformation makes the constraint that elements must add up to one obsolete. Instead, the sample space for the transformed object $\tilde{\mathbf{p}}_{i,t,h} = (\tilde{p}_{i,t,h,1}, \ldots, \tilde{p}_{i,t,h,K-1})'$ is $\mathbb{R}^{K-1}$. The above stated hypothesis test translates into a simple test of equality of the transformed population means, that is, H_0: $\tilde{\boldsymbol{\mu}}^A = \tilde{\boldsymbol{\mu}}^B$ versus H_1: $\tilde{\boldsymbol{\mu}}^A \neq \tilde{\boldsymbol{\mu}}^B$, suppressing the indices for the time period and expectation horizon for the remainder of this subsection to simplify the notation.

We assume that the corresponding $K-1 \times K-1$ population covariance matrix, $\boldsymbol{\Sigma}$, is the same in each subpopulation (an assumption that we could relax easily). $\boldsymbol{\Sigma}$ refers to within-group variation across survey participants. It can be interpreted, for instance, as reflecting heterogeneity in density expectations caused by idiosyncratic information or other information rigidities such as sticky information. Note that it is *not* a measure of uncertainty for individual density expectations.

In this setting, the null hypothesis can be implemented by a Hotelling test using the test statistic

$$\mathcal{T}^2 = \frac{N_A N_B}{N_A + N_B}\left(\bar{\tilde{\boldsymbol{p}}}_A - \bar{\tilde{\boldsymbol{p}}}_B\right)' \boldsymbol{S}^{-1}\left(\bar{\tilde{\boldsymbol{p}}}_A - \bar{\tilde{\boldsymbol{p}}}_B\right), \tag{3}$$

where

$$\bar{\tilde{\boldsymbol{p}}}_A = \frac{1}{N_A}\sum_{i=1}^{N_A} \tilde{\mathbf{p}}_i,$$

$$\bar{\tilde{\boldsymbol{p}}}_B = \frac{1}{N_B}\sum_{j=1}^{N_B} \tilde{\mathbf{p}}_j$$

and

$$\boldsymbol{S} = \frac{1}{N_A + N_B}\left(\sum_{i=1}^{N_A}\left(\tilde{\mathbf{p}}_i - \bar{\tilde{\boldsymbol{p}}}_A\right)\left(\tilde{\mathbf{p}}_i - \bar{\tilde{\boldsymbol{p}}}_A\right)' + \sum_{j=1}^{N_B}\left(\tilde{\mathbf{p}}_j - \bar{\tilde{\boldsymbol{p}}}_B\right)\left(\tilde{\mathbf{p}}_j - \bar{\tilde{\boldsymbol{p}}}_B\right)'\right)$$

denote the maximum likelihood estimates of the population parameters. The test statistic $\mathcal{T}^2$ in Equation (3) asymptotically follows a χ^2 distribution with $K-1$ degrees of freedoms.[4]

For finite samples, we can use a related statistic if we assume that $\tilde{\mathbf{p}}_{i,t,h}$ follows a multivariate normal distribution $\mathcal{N}(\tilde{\boldsymbol{\mu}}_{t,h}, \boldsymbol{\Sigma}_{t,h})$, implying that the original vector of probabilities $\mathbf{p}_{i,t,h}$ follows an *additive logistic normal distribution* according to the definition in Aitchison (1986, p. 113). In particular,

$$\mathcal{F} = \frac{N_A + N_B - K}{(N_A + N_B - 2)(K-1)}\mathcal{T}^2$$

follows an F distribution with $K-1$ and $N_A + N_B - K$ degrees of freedom in this case.

One advantage of treating histogram expectations as compositional data when testing for mean differences across groups is that we can easily extend the approach to allow for a joint comparison of more than two groups. This can be done by applying an ANOVA-type analysis to test the null hypothesis $H_0 : \tilde{\boldsymbol{\mu}}^1 = \tilde{\boldsymbol{\mu}}^2 = \ldots = \tilde{\boldsymbol{\mu}}^G$ against the alternative that at least one mean is different from the others. Another benefit is that the computations necessary for this approach are very fast, which is an advantage over the KLIC-based approach discussed below in Section 2.3.

2.2 | Multiple testing Bonferroni approach

The second approach for testing the null hypothesis described above deconstructs the histograms and compares the probabilistic expectations interval by interval. The primary null hypothesis implies for all $k = 1, \ldots, K$ that $H_0^k : \mu^A_{t,h,k} = \mu^B_{t,h,k}$ is true. The alternative in each case is $H_1^k : \mu^A_{t,h,k} \neq \mu^B_{t,h,k}$. For each k, we can use a standard two-sample t-test to test this. The primary null hypothesis is rejected if we can reject the implied null hypothesis for at least one of the bins. To ensure good small sample properties, we apply the test to the log probabilities, that is, $\ln(p_{i,t,h,k})$ for $k = 1, \ldots, K$.

[4]See Aitchison (1986, p. 125) for the suitable central limit theorem.

Because this approach gets us into a multiple-testing setup, we have to apply a correction to the significance level used for the individual t-tests to control the overall size of our testing approach. A common approach to do so is the Bonferroni correction that implies using a significance level of α/K for each individual hypothesis H_0^k, where α is the overall size that should be achieved.

Similar to the Hotelling test from the previous subsection, the approach described here can deal with more than two groups and does not require much computing power. A drawback of this approach is that the Bonferroni correction is known to be conservative ("undersized") when the individual test statistics are correlated, which—due to the compositional nature of our data—is the case in our context. This reduces the power of the testing approach.

2.3 | KLIC-based approach

The third approach for testing the primary null hypothesis starts from the fact that the KLIC is commonly used to compare probability distributions. The KLIC describes the expected value of the logarithmic difference between two sets of probability distributions (Mitchell & Hall, 2005). For discrete probability distributions, such as the histogram forecasts described above, the KLIC is defined as

$$\text{KLIC}\left(\overline{\boldsymbol{p}}_{t,h}^{A}, \overline{\boldsymbol{p}}_{t,h}^{B}\right) = \sum_{k=1}^{K} \bar{p}_{t,h,k}^{A} \ln\left(\frac{\bar{p}_{t,h,k}^{A}}{\bar{p}_{t,h,k}^{B}}\right). \tag{4}$$

In Equation (4), the elements of the vectors $\overline{\boldsymbol{p}}_{t,h}^{A} = (1/N_A)\sum_{i=1}^{N_A} \mathbf{p}_{i,t,h}$ and $\overline{\boldsymbol{p}}_{t,h}^{B} = (1/N_B)\sum_{i=1}^{N_B} \mathbf{p}_{i,t,h}$ represent the average (nontransformed) probability mass assigned to bin k based on all individuals in a particular group.

Under the null hypothesis defined above, the aggregate distributions of both groups are very similar for finite group sizes and asymptotically identical. In this case, the KLIC from Equation (4) is close to zero. The more $\bar{p}_{t,h,k}^{A}$ and $\bar{p}_{t,h,k}^{B}$ deviate from each other, the larger the value of KLIC $\left(\overline{\boldsymbol{p}}_{t,h}^{A}, \overline{\boldsymbol{p}}_{t,h}^{B}\right)$. To test whether KLIC $\left(\overline{\boldsymbol{p}}_{t,h}^{A}, \overline{\boldsymbol{p}}_{t,h}^{B}\right)$ is significantly different from zero and, hence, the null hypothesis should be rejected, we use a bootstrap approach. Specifically, we draw Z random samples of size N with replacement from the available expectation data. We then randomly assign N_A of the drawn histograms to group A and N_B drawn histograms to group B. For each bootstrap sample, we then calculate the KLIC as described in Equation (4). We conclude that KLIC $\left(\overline{\boldsymbol{p}}_{t,h}^{A}, \overline{\boldsymbol{p}}_{t,h}^{B}\right)$ is significantly different from zero whenever it exceeds the 95%-quantile of the Z bootstrapped KLIC values.[5]

The use of the KLIC for comparing histogram forecasts has several disadvantages. First, the KLIC can only be used to compare the probability distributions of two groups. In case the number of groups exceeds two, only pairwise comparisons can be carried out. Second, calculation of the bootstrapped KLICs is computationally intensive. Third, the results of the KLIC-based test are sensitive to the ordering of the two groups, that is, KLIC $\left(\overline{\boldsymbol{p}}_{t,h}^{A}, \overline{\boldsymbol{p}}_{t,h}^{B}\right) \neq$ KLIC $\left(\overline{\boldsymbol{p}}_{t,h}^{B}, \overline{\boldsymbol{p}}_{t,h}^{A}\right)$, although it should be noted that the test is correctly sized in both cases (if a particular order is chosen ex ante). Overall, these are important shortcomings relative to the previously discussed alternatives.

3 | MONTE CARLO SIMULATIONS

We now assess the properties of the testing approaches discussed in the previous section by means of Monte Carlo simulations. In particular, we compare the rejection frequencies of the Hotelling test, the multiple testing Bonferroni approach, and the KLIC-based test under the null hypothesis of equal subpopulation means and various altenatives.

[5]This approach is similar to Clements (2022) who proposes a test for heterogeneity in the revisions of GDP growth expectations in the US Survey of Professional Forecasters. In order to address potential issues due to small sample size, Clements (2022) simulates a set of imaginary SPF participants by randomly drawing from the set of forecast revisions reported in a given survey wave. While his bootstrap approach focuses on revisions of point forecasts, we randomly draw and reassign entire density forecasts. In an earlier study, Clements (2018) implements the above-mentioned alternative approach that tests for differences between average expectations of a group of forecasters and a model-based benchmark by exploiting variation across time. For our purpose, this alternative is not suitable for two reasons. First, many of the new household surveys with large cross-sections still have a short time dimension. Second, the alternative does not allow to analyze changes of group differences across time.

3.1 | Simulation setup

For the Monte Carlo evaluation, we calibrate our benchmark histograms to the one-year-ahead inflation histograms from the 2020Q1 wave of the SPF (see Section 4 for details on the survey). We first obtain an estimate of $\boldsymbol{\Sigma}$ by applying the additive logratio transformation in Equation (2) to the individual histograms and calculating the corresponding covariance matrix. To obtain $\tilde{\boldsymbol{\mu}}^A$, we fit a normal distribution to the aggregate SPF histogram. To do so, we first calculate the mean and standard deviation of the aggregate histogram by assuming that the probability mass in each bin is centered at the midpoint and use those parameters as starting values for the optimization.[6] We then calculate the probability mass for each bin using the fitted normal density and apply the additive logratio transformation to these probabilities.

For each scenario described below, we simulate $S = 2000$ artificial data sets of histograms.[7] We then apply the Hotelling test, the Bonferroni-adjusted t-tests, and the KLIC-based test as described in the previous section and calculate the rejection frequencies in each case. For the KLIC-based test, we set the number of bootstrap replications to $Z = 250$. Finally, we choose a nominal level of $\alpha = 0.05$.

In practice, the SPF histograms are relatively coarse, and many individual histograms do not closely resemble a normal distribution. For the one-year-ahead inflation expectations, almost two third of the SPF participants assign nonzero probability to at most five bins. Therefore, a possible concern could be that the choice of a Gaussian distribution for the simulations is not appropriate for individual survey responses and, thus, might yield a misleading impression of the tests' properties in real applications. To assess how deviations from normality affect the size and power of the tests, we conduct a second set of simulations with a data generating process (DGP) that mimics this data feature. In particular, we transform the simulated histograms into more coarse versions with nonzero probability assigned only to the (two to five) bins with the highest probabilities. For each individual histogram, we randomly draw the precise number of bins with nonzero probability with selection probabilities equal to the relative frequency of observations in the SPF with two to five bins (rescaled to sum to unity).[8] We refer to the two settings as the Gaussian setup and the truncated-probabilities setup below.

3.2 | Results for Gaussian setup

In a first step, we analyze whether the different tests are correctly sized for varying group size. For each group, we consider group sizes of 10, 25, 50, 75, 100, 200, and 500 individuals. While a group size of approximately 25 individuals seems to be a good description of surveys among professional forecasters, a group size of 500 individuals is more representative of a typical household survey.

Table 1 shows the rejection frequencies of all three tests under the null hypothesis. While Panel A shows the results for our baseline calibration of the covariance matrix that determines the within-group heterogeneity, Panels B and C present findings for lower/higher within-group heterogeneity. In these settings, we multiply $\boldsymbol{\Sigma}$ by a factor c, where c equals 0.5 (Panel B) or 5 (Panel C). For both the Hotelling test for compositional data and the KLIC-based approach (and for all sample sizes and levels of within-group heterogeneity), the empirical size is very close to the nominal size of 0.05. In contrast, the multiple testing Bonferroni approach is, as expected, undersized. With respect to the speed of the MC simulations, performing the KLIC-based test 2000 times takes a little more than 14 hours on a standard desktop computer while 2000 Hotelling tests take only nine seconds.

Next, we turn to an assessment of the power of the tests. Unless explicitly stated otherwise, we set the group sizes to $N_A = N_B = 25$. Because the Bonferroni approach is too conservative in the sense that it suffers from size distortions, we report size-adjusted power statistics for this approach.

We first consider shifts in the expected first moment of the histograms. Under H_0, all histograms have the same expected value. We then shift the expected value of the histograms for one group. We consider the following shifts: H_{1a}: 0.05, H_{1b}: 0.1, H_{1c}: 0.2, H_{1d}: 0.3, H_{1e}: 0.4, H_{1f}: 0.5, H_{1g}: 0.75, and H_{1h}: 1.00. Next, we change the population standard deviation of the

[6]Figure A.1 of the supporting information shows the aggregate SPF histogram (reporting densities instead of bin probabilities) based on the predictions reported by 46 survey participants. Mean and standard deviation based on the "mass-at-midpoint"-approach are 1.26 percentage points and 0.60 percentage point, respectively. The black line shows the fitted normal distribution, which has a mean of 1.25 percentage points and a standard deviation of 0.54 percentage point.

[7]This number was chosen because preliminary simulations had indicated that this is a number sufficiently large to give us stable estimates of the empirical sizes of tests and of the power curves. Lower values of S would have been possible, of course, to cut down on computational cost—at the expense of the precision of results.

[8]These frequencies are 9.9%, 20.0%, 16.4%, and 15.1% for the one-year-ahead inflation expectations in the SPF.

TABLE 1 Monte Carlo simulation results: Size analysis.

Group size ($N_A = N_B$)	10	25	50	75	100	200	500
Panel A: Gaussian setup, baseline within-group heterogeneity							
Compositional	0.055	0.052	0.056	0.048	0.046	0.046	0.048
Bonferroni	0.053	0.050	0.049	0.043	0.033	0.038	0.033
KLIC	0.062	0.050	0.063	0.049	0.054	0.047	0.049
Panel B: Gaussian setup, low within-group heterogeneity (relative to baseline)							
Compositional	0.055	0.048	0.048	0.042	0.051	0.057	0.045
Bonferroni	0.035	0.040	0.043	0.040	0.039	0.040	0.038
KLIC	0.047	0.049	0.042	0.051	0.047	0.066	0.050
Panel C: Gaussian setup, high within-group heterogeneity (relative to baseline)							
Compositional	0.052	0.049	0.051	0.050	0.058	0.054	0.050
Bonferroni	0.048	0.044	0.042	0.037	0.043	0.037	0.039
KLIC	0.038	0.053	0.047	0.052	0.047	0.052	0.055
Panel D: Truncated-probabilities setup, baseline within-group heterogeneity							
Compositional	0.050	0.047	0.044	0.038	0.049	0.046	0.045
Bonferroni	0.061	0.043	0.043	0.044	0.042	0.041	0.040
KLIC	0.048	0.046	0.052	0.046	0.053	0.045	0.049

Note: The panels show rejection frequencies for the Hotelling test for compositional data, the multiple testing approach, and the KLIC-based test under the null hypothesis of no expectation difference between two groups for varying group size. In the simulations, all tests are used with a nominal size of 0.05. Panel A presents results for the Gaussian setup with baseline within-group heterogeneity. Panels B and C show rejection rates when we multiply the baseline $\boldsymbol{\Sigma}$ by 0.5 and 5, respectively. Panel D presents our findings for the truncated-probabilities setup with baseline within-group heterogeneity.

histograms for one group by multiplying the standard deviation under H_0 by a factor unequal to one. We consider the following factors: H_{2a}: 1.05, H_{2b}: 1.1, H_{2c}: 1.2, H_{2d}: 1.3, H_{2e}: 1.4, H_{2f}: 1.5, H_{2g}: 1.75, H_{2h}: 2.00, H_{2i}: 2.50, and H_{2j}: 3.00. Next, we change the group sizes under the assumption of a moderate mean shift of 0.05 (i.e., under H_{1a}). Finally, we consider changes in the within-group heterogeneity by adjusting the covariance matrix $\boldsymbol{\Sigma}$ (again assuming a mean shift of 0.05 as in H_{1a}). In particular, we consider a range of settings for $c\boldsymbol{\Sigma}$, where c assumes the following values in the different alternative scenarios: H_{3a}: 0.5, H_{3b}: 0.75, H_{3c}: 1.0, H_{3d}: 1.25, H_{3e}: 1.5, H_{3f}: 1.75, H_{3g}: 2.0, H_{3h}: 2.5, H_{3i}: 3.0, and H_{3j}: 10.0.

The plot in the upper left of Figure 1 shows the rejection frequencies for the set of alternatives for which we shift the mean expectations of one group. Evidently, the performances of the three approaches are very different. While the Hotelling test for compositional data rejects the null hypothesis in about 70% of cases already for a small shift of 0.05, the KLIC-based test produces much lower rejection frequencies for small deviations from the null hypothesis. It matches the performance of the Hotelling test only for very large mean shifts of 0.75 or more. The size-adjusted power of the Bonferroni approach lies somewhere in between, matching the power of the Hotelling test for mean shifts of 0.3 or more.

The upper-right plot of Figure 1 shows analogous rejection rates for the second set of simulations that analyze the test performance against alternatives that deviate from the null hypothesis due to differences in the population standard deviation of the density expectations across groups. The alternatives range from moderate deviations (for which the ratio of the implied standard deviation of the two groups is 1.05) to extreme (ratio of 3). Again, the Hotelling test yields high rejection frequencies for all alternatives except H_{2a}. The Bonferroni approach here shows very similar (size-adjusted) power to the compositional approach. In contrast, rejection frequencies of the KLIC-based test are low for the alternatives that do not differ much from the null hypothesis; we observe rejection frequencies above 25% only for alternatives that are based on a ratio of the standard deviations of 1.5 or larger.

Next, we assess how the group size affects the rejection frequencies. We analyze this for the alternative H_{1a}, which implies a mean shift of 0.05 of mean expectations in one of the groups. The results in the lower-left plot show that increasing the sample size—even to numbers that would be common in household surveys—does not substantially increase the rejection frequency for the KLIC-based test. Rejection frequencies are much higher for the Hotelling test—for small sample sizes and increasingly so for larger sample sizes. Again, the Bonferroni approach is somewhere in-between, exhibiting very low (size-adjusted) power for small to medium sample sizes but catching up with the Hotelling test for large sample sizes of $N_A = N_B = 500$.

Finally, the lower-right plot of Figure 1 shows how the level of within-group heterogeneity affects rejection frequencies. It is evident that the KLIC-based test and the Bonferroni approach have no power against H_{1a} independently of the level of within-group heterogeneity. For the Hotelling test, we observe that—not surprisingly—it has good power against a

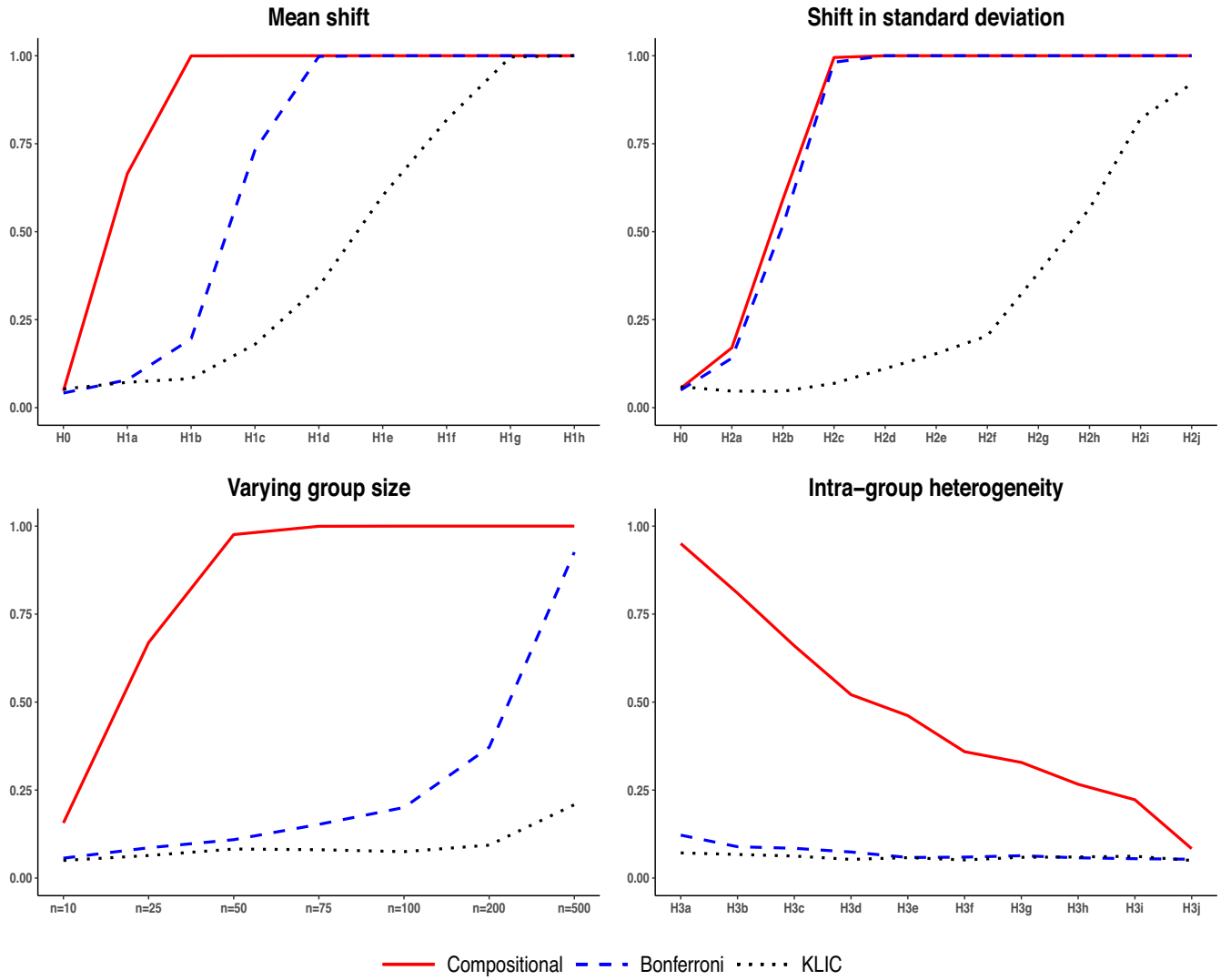

FIGURE 1 Monte Carlo simulation results: Power analysis for Gaussian setup. *Note*: The plots show rejection frequencies based on the Gaussian setup for the Hotelling test for compositional data (solid red lines), the multiple testing approach (dashed blue lines), and the KLIC-based test (dotted black lines) under various alternatives. The upper-left plot corresponds to alternative hypotheses with differences in means of density expectations (H_{1a}: 0.05, H_{1b}: 0.1, H_{1c}: 0.2, H_{1d}: 0.3, H_{1e}: 0.4, H_{1f}: 0.5, H_{1g}: 0.75, and H_{1h}: 1.00). The upper-right plot corresponds to alternative hypotheses with differences in the standard deviation of density expectations (H_{2a}: 1.05, H_{2b}: 1.1, H_{2c}: 1.2, H_{2d}: 1.3, H_{2e}: 1.4, H_{2f}: 1.5, H_{2g}: 1.75, H_{2h}: 2.00, H_{2i}: 2.50, and H_{2j}: 3.00). The lower-left plot corresponds to simulations with varying group size, assuming mean differences as in H_{1a}. The lower-right plot corresponds to simulations with varying within-group heterogeneity for which we multiply our baseline calibration for the covariance matrix by a factor c, where c assumes the following values: H_{3a}: 0.5, H_{3b}: 0.75, H_{3c}: 1.0, H_{3d}: 1.25, H_{3e}: 1.5, H_{3f}: 1.75, H_{3g}: 2.0, H_{3h}: 2.5, H_{3i}: 3.0, and H_{3j}: 10.0 (again assuming mean differences as in H_{1a}). In the simulations, all tests are used with a nominal size of 0.05.

small difference in the expected value of the histograms implied by H_{1a} when the heterogeneity within groups is small, but less so when it is high. Rejection frequencies decline considerably from almost 100% to around 10% over the scenarios considered in our simulations. Still, for any level of within-group heterogeneity, the rejection frequencies are substantially higher than those of the KLIC-based test and the Bonferroni approach.

The number of bins in the SPF questionnaire changes over time. To assess whether the power of the tests depends on the specifics of the survey design, we aggregate the bin probabilities for two adjacent bins in another set of simulations. Thereby, we effectively reduce the number of bins from twelve to six. Using the changed setting, we re-estimate rejection frequencies. Figure 2 shows that this procedure slightly reduces the power of the tests. However, the rejection frequencies of the Hotelling test remain high and dominate those of the other tests.

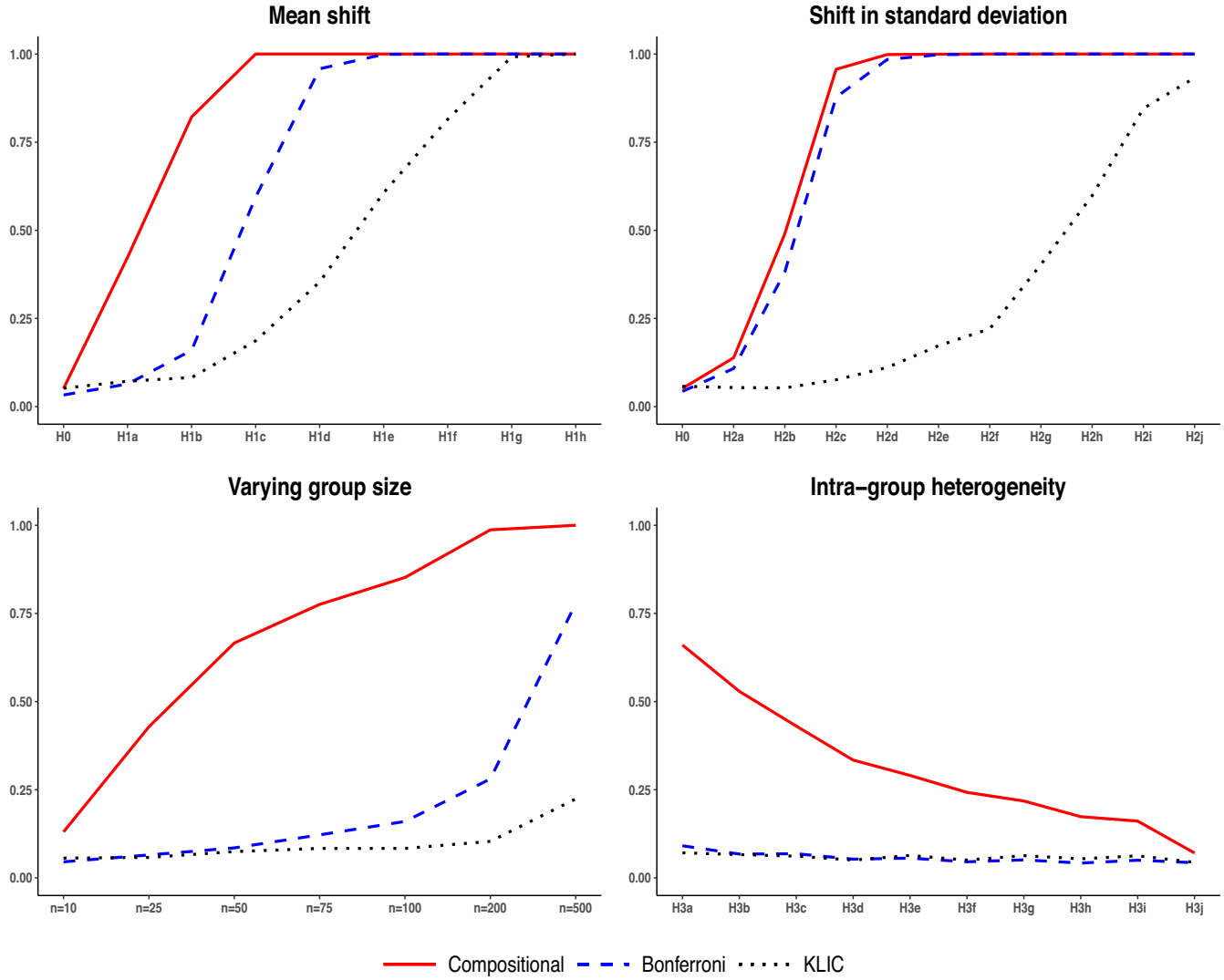

FIGURE 2 Monte Carlo simulation results: Power analysis for reduced number of bins. *Note*: The plots show rejection frequencies based on the Gaussian setup with the number of bins reduced by half for the Hotelling test for compositional data (solid red lines), the multiple testing approach (dashed blue lines), and the KLIC-based test (dotted black lines) under various alternatives. The upper-left plot corresponds to alternative hypotheses with differences in means of density expectations (H_{1a}: 0.05, H_{1b}: 0.1, H_{1c}: 0.2, H_{1d}: 0.3, H_{1e}: 0.4, H_{1f}: 0.5, H_{1g}: 0.75, and H_{1h}: 1.00). The upper-right plot corresponds to alternative hypotheses with differences in the standard deviation of density expectations (H_{2a}: 1.05, H_{2b}: 1.1, H_{2c}: 1.2, H_{2d}: 1.3, H_{2e}: 1.4, H_{2f}: 1.5, H_{2g}: 1.75, H_{2h}: 2.00, H_{2i}: 2.50, and H_{2j}: 3.00). The lower-left plot corresponds to simulations with varying group size, assuming mean differences as in H_{1a}. The lower-right plot corresponds to simulations with varying within-group heterogeneity for which we multiply our baseline calibration for the covariance matrix by a factor c, where c assumes the following values: H_{3a}: 0.5, H_{3b}: 0.75, H_{3c}: 1.0, H_{3d}: 1.25, H_{3e}: 1.5, H_{3f}: 1.75, H_{3g}: 2.0, H_{3h}: 2.5, H_{3i}: 3.0, and H_{3j}: 10.0 (again assuming mean differences as in H_{1a}). In the simulations, all tests are used with a nominal size of 0.05.

3.3 | Results for truncated-probabilities setup

We now turn to the Monte Carlo simulation for the alternative truncated-probabilities setup. Panel D of Table 1 shows that the tests are still appropriately sized for the alternative DGP.

Figure 3 presents the results for the power analysis. Clearly, the power of all tests is affected negatively when the data are not normally distributed. The upper-left plot shows that the rejection frequencies for the Hotelling and Bonferroni tests are much lower for small mean shifts and are now in a similar range as those for the KLIC-based test. In fact, rejection frequencies for the KLIC are slightly higher than those for the other tests for intermediate mean shifts, although the differences are relatively small. The upper-right plot shows that the power to detect shifts in the standard deviation is reduced for all three tests relative to the Gaussian setup, although the relative ranking remains the same. The lower-left

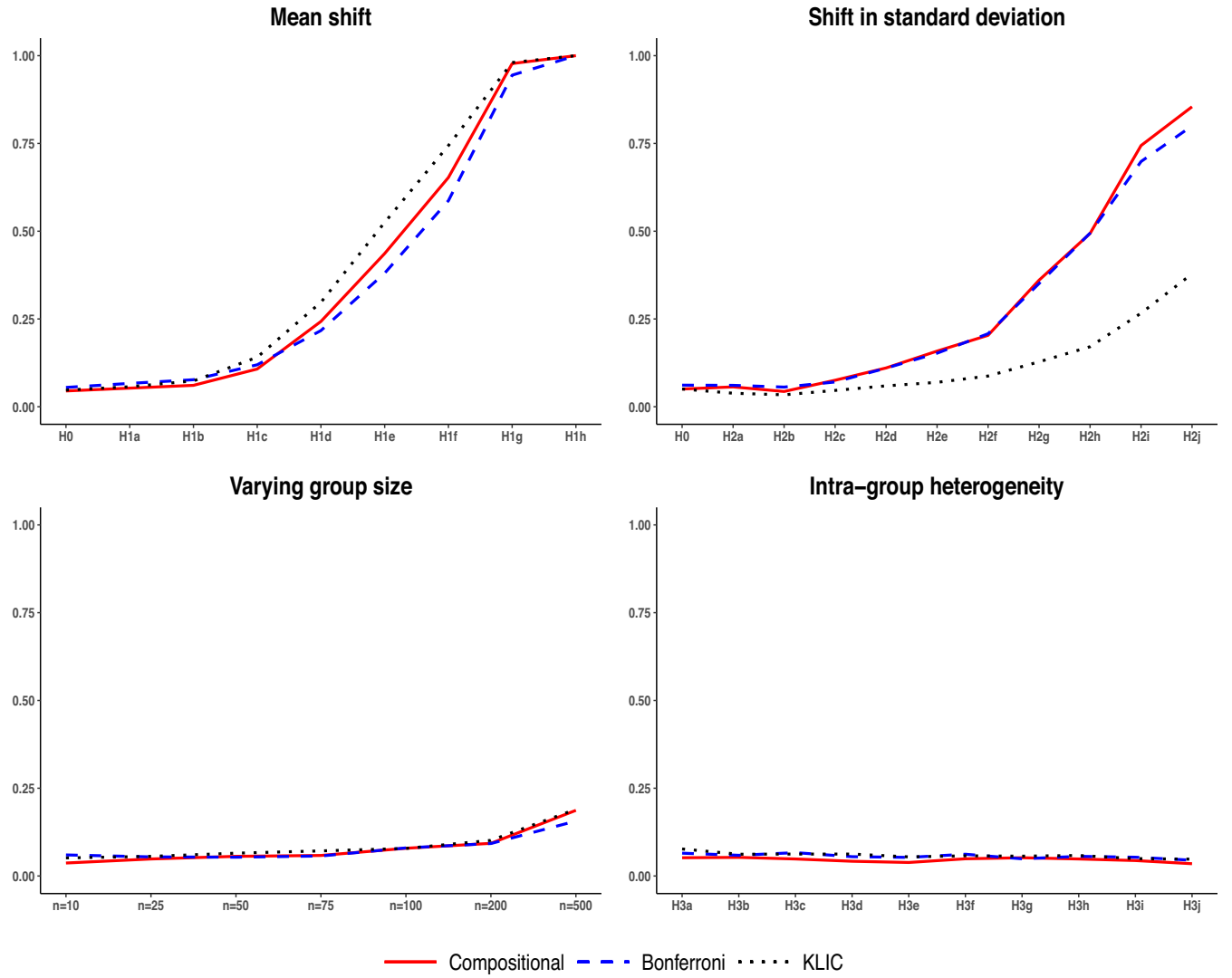

FIGURE 3 Monte Carlo simulation results: Power analysis for truncated-probabilities setup. *Note*: The plots show rejection frequencies based on the truncated-probabilities setup for the Hotelling test for compositional data (solid red lines), the multiple testing approach (dashed blue lines), and the KLIC-based test (dotted black lines) under various alternatives. The upper-left plot corresponds to alternative hypotheses with differences in means of density expectations (H_{1a}: 0.05, H_{1b}: 0.1, H_{1c}: 0.2, H_{1d}: 0.3, H_{1e}: 0.4, H_{1f}: 0.5, H_{1g}: 0.75, and H_{1h}: 1.00). The upper-right plot corresponds to alternative hypotheses with differences in the standard deviation of density expectations (H_{2a}: 1.05, H_{2b}: 1.1, H_{2c}: 1.2, H_{2d}: 1.3, H_{2e}: 1.4, H_{2f}: 1.5, H_{2g}: 1.75, H_{2h}: 2.00, H_{2i}: 2.50, and H_{2j}: 3.00). The lower-left plot corresponds to simulations with varying group size, assuming mean differences as in H_{1a}. The lower-right plot corresponds to simulations with varying within-group heterogeneity for which we multiply our baseline calibration for the covariance matrix by a factor c, where c assumes the following values: H_{3a}: 0.5, H_{3b}: 0.75, H_{3c}: 1.0, H_{3d}: 1.25, H_{3e}: 1.5, H_{3f}: 1.75, H_{3g}: 2.0, H_{3h}: 2.5, H_{3i}: 3.0, and H_{3j}: 10.0 (again assuming mean differences as in H_{1a}). In the simulations, all tests are used with a nominal size of 0.05.

plot shows that with the truncated-probabilities setup the low rejection frequencies for a small mean shift cannot be improved upon by increasing the group size to 500. Finally, the lower-right plot shows that the competitive edge of the Hotelling test in settings with low intragroup heterogeneity disappears for the truncated-probabilities setup.

In summary, the Hotelling test for compositional data and the KLIC-based test is appropriately sized while the Bonferroni approach is, as expected, undersized. For the Gaussian setup, the Hotelling test clearly outperforms the other two approaches in terms of power against all alternatives considered here. In particular, the Hotelling test detects significant group differences for much smaller differences in the expected value of density expectations relative to the KLIC-based approach and the Bonferroni approach, especially when group sizes and/or within-group heterogeneity are small. It also has much higher power compared with the KLIC-based approach against differences in the standard deviation of the density expectations across groups. While the rejection frequencies of the Hotelling test are considerably

lower for the case of the truncated-probabilities setup, the power of the other tests are not substantially higher for any of the considered alternatives in this setup. The results from the Monte Carlo simulations thus complement the conceptual advantages of our approach as described in Section 2.

4 | EMPIRICAL APPLICATIONS

In this section, we consider a range of applications for which the discussed tests can be useful. All applications deal with aspects of expectation heterogeneity that have recently been discussed in the literature. The data in the applications are either from the Survey of Professional Forecasters conducted by the European Central Bank for the euro area (as described in Bowles et al., 2007) or from the Survey of Consumer Expectations conducted by the Federal Reserve Bank of New York among US households (see Armantier et al., 2015). For all applications, we exclude those histograms from the sample that assign 100% probability to a single bin. Moreover, as the probability mass assigned to the exterior bins is zero in many cases, we choose the sixth bin as the reference category throughout.

4.1 | Response to rising inflation rates

After several years of low inflation rates, inflation in the euro area began to increase midway through 2021. In this section, we analyze whether the steady rise in inflation changed the inflation density expectations of SPF participants at different forecast horizons, that is, we test for differences in the aggregate densities across time, thereby shedding light on their temporal stability. Intuitively, one might expect to observe an immediate adjustment of short-term expectations while long-term expectations would not change if they were firmly anchored.

The SPF asks experts from financial and nonfinancial institutions to report predictions for several macroeconomic outcomes in the euro area, including one- and five-year-ahead inflation expectations. It has been conducted by the ECB since 1999 at a quarterly frequency. We focus on the density expectations that are elicited by asking panelists to state probabilities for a range of bins (e.g., the likelihood that inflation turns out to be between 1.5% and 1.9%) that jointly determine a histogram as an approximation of the underlying density expectations. An attractive feature of the SPF is that the bins have a constant width with the exception of the exterior bins, which are half-open. In particular, the intervals as defined in the SPF questionnaire have a width of 0.4 percentage point with a gap of 0.1 percentage point between bins.[9]

To analyze whether the increase in inflation had an effect on inflation expectations, we use 2021Q1 as a reference wave and compare the histogram forecasts from each subsequent wave (up until 2022Q2) to those reported in this reference wave. For all of these waves, the SPF bin definitions for inflation have remained identical and comprise $K = 12$ bins. To provide some descriptive evidence, we compute the first four moments of the one- and five-year-ahead aggregate density expectations from each wave, that is, $\overline{\boldsymbol{p}}_{t,h}$. Next, we formally test for each wave and horizon whether expectations have changed relative to the 2021Q1 wave by using the three testing approaches described in Section 2.

Panels A and C of Table 2 present the number of panelists in each wave along with the estimated moments (based on the "mass-at-midpoint" approach) of the aggregate probability distributions for the one- and five-year-ahead inflation expectations. As the test statistics along with the corresponding p-values for the three tests given in Panels B and D allow for an overall assessment only, these moments are informative about what particular features of the density expectations change and what features stay rather constant. They give an indication of whether the central tendency changes, whether expectational uncertainty changes, whether the (a-)symmetry of the distribution changes, and/or whether the expectations distribution became more leptokurtic or less leptokurtic.

For the one-year-ahead expectations, Panel A shows an upward shift in the histogram mean over time as well as an increase in the standard deviation for the 2022Q1 and 2022Q2 waves. As shown in Section 3, the Hotelling test should be able to detect such differences. For skewness and kurtosis, we do not observe any clear patterns. The upward shift in the mean is visible for virtually all of the underlying individual density expectations and suggests that SPF participants quickly reacted to the rising inflation rates. However, these changes are relatively small in magnitude from one period to the next relative to the heterogeneity of individual expectations. As a result, all tests do not reject the null hypothesis when

[9]One exception is a recent change in the survey design for expectations of GDP growth. In 2020Q2, bins with a width of two percentage points have been introduced.

TABLE 2 Differences in inflation expectations over time.

	2021Q1	2021Q2	2021Q3	2021Q4	2022Q1	2022Q2
Panel A: Histogram moments (one-year-ahead expectations)						
Group size	39	39	34	38	38	31
Mean	1.24	1.36	1.51	1.71	1.94	2.73
Standard deviation	0.78	0.78	0.74	0.79	0.98	1.06
Skewness	0.12	0.16	−0.19	0.11	0.29	−0.44
Kurtosis	3.85	4.38	4.40	3.90	3.15	2.82
Panel B: Distance measures (one-year-ahead expectations)						
Compositional	–	0.525	1.895	2.813	3.083	12.300
	–	(0.880)	(0.058)	(0.005)	(0.002)	(0.000)
Bonferroni	–	−1.685	2.536	3.759	3.599	−7.166
	–	(1.000)	(0.161)	(0.004)	(0.007)	(0.000)
KLIC	–	0.016	0.094	0.199	0.303	1.051
	–	(0.562)	(0.002)	(0.000)	(0.000)	(0.000)
Panel C: Histogram moments (five-year-ahead expectations)						
Group size	40	43	35	37	42	38
Mean	1.60	1.62	1.75	1.86	1.87	2.02
Standard deviation	0.83	0.80	0.88	0.89	0.89	0.85
Skewness	0.03	0.01	0.09	0.28	0.17	−0.05
Kurtosis	3.95	4.38	3.92	4.06	3.92	3.52
Panel D: Distance measures (five-year-ahead expectations)						
Compositional	–	0.247	0.504	1.157	1.625	4.701
	–	(0.993)	(0.894)	(0.334)	(0.111)	(0.000)
Bonferroni	–	1.063	0.985	1.958	−1.946	−3.505
	–	(1.000)	(1.000)	(0.647)	(0.662)	(0.009)
KLIC	–	0.004	0.017	0.053	0.060	0.140
	–	(0.868)	(0.380)	(0.080)	(0.058)	(0.000)

Note: Panel A presents moments (based on the "mass-at-midpoint" approach) for the aggregate one-year-ahead inflation expectations from the 2021Q1 to 2022Q2 waves of the SPF. Panel B shows the test statistics relative to the 2021Q1 wave along with corresponding *p*-values in parentheses. For the multiple testing approach, we report the largest test statistic across the twelve distinct bins and twelve times the corresponding *p*-value. Panels C and D present the results for the five-year-ahead inflation expectations.

comparing the 2021Q1 and 2021Q2 waves. In contrast, the tests detect significant differences in short-term forecasts when comparing subsequent waves to the reference period. Note that to make results comparable, we report the minimum of one and twelve times the smallest of the twelve *p*-value in case of the Bonferroni approach.[10]

We also observe an increase in the histogram mean of the five-year-ahead expectations, although the changes from one period to the next are clearly smaller than those for the one-year-ahead expectations. This likely reflects the fact that such expectations are more anchored and less impacted by price shocks that are perceived to have a large transitory component. In addition, we do not observe an increase in the standard deviation toward the end of the sample. As a result, the tests reject the null hypothesis only for the 2022Q2 wave relative to 2021Q1.

We conclude that the SPF participants quickly adapted their short-term inflation expectations in response to the recent inflation shock. Long-term expectations reacted less strongly and more gradually but also increased significantly relative to 2021Q1. This suggests that there was a deterioration in the degree to which medium term expectations were anchored.[11] This finding is consistent with the results in Binder et al. (2023) for US forecasters. The low *p*-values for the KLIC-based test are in line with our Monte Carlo simulation results for the truncated-probabilities setup. As seen in Figure 3, the power of the KLIC-based test can exceed that of the Hotelling test for moderate mean shifts in nonnormal settings.

[10]The displayed test statistics are the largest of the twelve bin-specific *t*-statistics.

[11]The ECB revised its inflation target in 2021. Comparing the five-year-ahead inflation expectations from the 2021Q3 wave (elicited just before the publication of the revised ECB strategy) with those from the 2021Q4 wave, we do not reject the null hypothesis. A possible explanation for this finding is that professional forecasters adjusted their density expectations to the new inflation target well in advance to the official announcement by the ECB.

4.2 | Different types of professional forecasters

A number of studies have recently observed that one can distinguish two types of survey-based forecasts based on the rounding behavior of the panelists (Binder, 2017; Clements, 2021; Glas & Hartmann, 2022; Reiche & Meyler, 2022). For density forecasts, Glas and Hartmann (2022) show that one type of panelist ("rounders") states interval probabilities that are multiples of five or ten and tends to assign positive probabilities to only a relatively small subset of the surveyed bins while another type of panelist ("nonrounders") reports probabilities that do not share a common divisor and tends to consider a larger number of bins, often reporting probabilities with higher precision (i.e., to at least one decimal point) for most, or indeed, all of the surveyed bins.

We test whether the reported expectations from rounders and nonrounders are indeed sampled from different populations. We define rounders as those panelists who report histograms containing probabilities of which more than half are multiples of five.

For this analysis, we look at the SPF density forecasts for inflation and real GDP growth and focus on the one-year-ahead forecasts from all available survey waves. Overall, the sample includes information from 108 panelists and $T = 94$ survey rounds, covering the period 1999Q1–2022Q2. The panel is unbalanced due to frequent dropouts and entries of new participants. The black lines in Figure A.2 of the supporting information show that, on average, 45–55 panelists report histogram forecasts for inflation and GDP growth each quarter (with declining trend).

Figure 4 shows the *p*-values for all tests and each survey wave; to ensure comparability of results, we again show the smallest *p*-value multiplied by the (time-varying) number of bins for the Bonferroni approach. The null hypothesis of equal expectations is rejected for most survey waves. We obtain the lowest rejection frequency of 75% of the survey waves for the KLIC-based test in the case of inflation expectations. In line with the evidence in Glas and Hartmann (2022), the rejections are driven primarily by the lower variances of the histograms reported by the rounders rather than differences in mean expectations (see Figure A.3 of the supporting information). We observe a few large *p*-values in the first half of the sample. The red lines in Figure A.2 of the supporting information show that only a small number of nonrounders are included in these particular waves, leading to very low power of the tests. Another spike is visible in 2009Q1. This can be explained by a pile-up of probabilities in the exterior bins due to the Great Recession, which partially masks the differences in the second moments between both groups (see Figure A.3 of the supporting information).

Referring back to the discussion of the Gaussian setup versus the truncated-probabilities setup in Section 3, the histograms reported by the nonrounders are more in line with the normality assumption than those of the rounders. As such, the SPF data can be thought of as a mixture of "well-behaved" and relatively coarse histograms. With that in mind, we briefly return to the previous application and now focus on the subsample of nonrounders only. Broadly speaking, we find that the nonrounders adjust their short-term expectations more slowly than the rounders (Table A.1 of the supporting information). In particular, the standard deviation of the aggregate histogram is essentially constant. As before, the tests

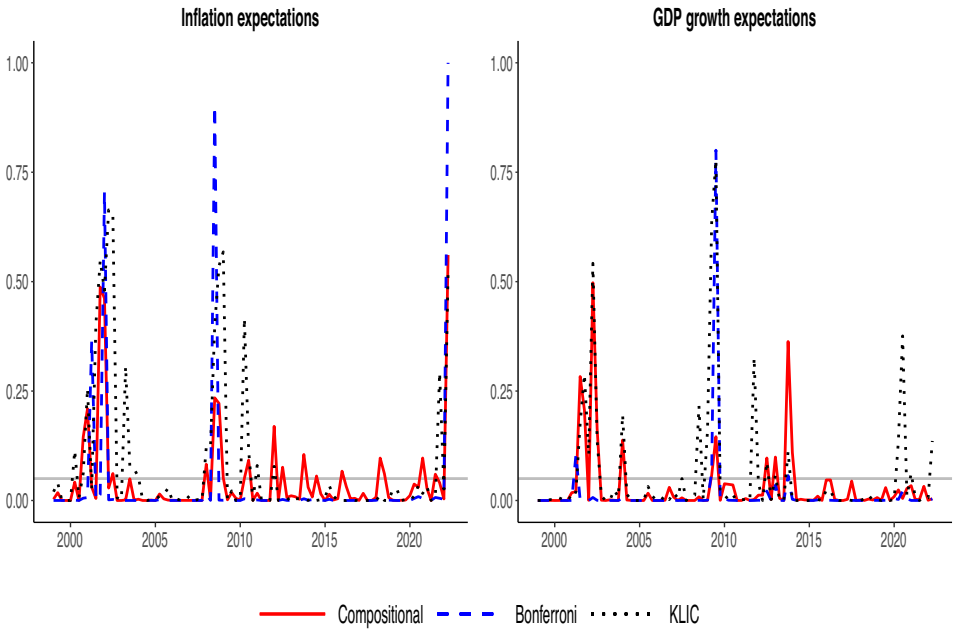

FIGURE 4 *p*-values for heterogeneity tests (SPF): Rounders versus nonrounders. *Note*: The plot shows the *p*-values from the Hotelling test for compositional data (solid red lines), the multiple testing approach (dashed blue lines), and the KLIC-based test (dotted black lines) for the analysis of differences in inflation expectations (left) and GDP growth expectations (right) between rounders and nonrounders. For the multiple testing approach, we report (the minimum of one and) the smallest *p*-value multiplied by the number of bins to make it comparable. The sample period is 1999Q1–2022Q2.

detect significant differences for the short-term expectations before such differences are evident for the long-term expectations. Interestingly, we observe that the KLIC-based test does not produce smaller *p*-values than the Hotelling test, unlike in Table 2. It is likely that this is because the histograms of the nonrounders are more in line with the normality assumption.

4.3 | Gender differences in household inflation expectations

A drawback of the SPF data is the relatively small cross-section. Figure 1 shows that a small group size negatively affects the power of all tests including the Hotelling test. Therefore, we now turn to data on expectations of private households with a larger number of individual survey responses. A potential disadvantage is that within-group heterogeneity may be larger for households than for experts. Moreover, it is likely that the data contain more frequent violations of the normality assumption than the SPF data.

First, we compare inflation expectations of men and women. Among others, D'Acunto et al. (2021) show that, on average, women expect higher inflation rates than men due to higher exposure to price changes for certain household items during grocery shopping. One may expect to find a similar divergence in the density forecasts reported by both genders. For US households, Armantier et al. (2021) show that female survey participants assign more probability mass to both exterior bins, resulting in higher uncertainty. Similarly, using survey data from German households, Conrad et al. (2022) find that the inflation histograms of women tend to be more dispersed than those of men.

Here, we tackle the question of gender differences in inflation expectations using the full density forecasts reported in the SCE, which is a monthly and representative survey among US households that asks questions about socioeconomic characteristics and macroeconomic expectations. The SCE has been conducted since June 2013. Each wave includes roughly 1300 households with a balanced relation between male and female household heads.[12]

We use density forecasts for the consumer price inflation rate over the next twelve months (*Q9* in the survey questionnaire).[13] Our sample includes responses from 18,066 households across $T = 103$ survey waves, covering the period from June 2013 to December 2021. The elicited histogram forecasts in the SCE are conceptually similar to those in the SPF. The specific design differs, however, in the sense that the width of the intervals is larger and varies across bins.[14]

All tests reject very strongly for all survey waves the null hypothesis of no differences in the density forecasts of men and women, with *p*-values way below 0.01 (results not shown). This is not surprising given the large differences in expectations across genders. The upper-left plot in Figure A.5 of the supporting information shows the aggregate histograms (pooled across households and survey waves) for men and women. In line with the studies discussed above, we observe that women assign more probability mass to the exterior bins and have higher mean expectations and variances. The latter can be seen more clearly in Figure A.6 of the supporting information, which presents the time series for the first four moments of the aggregate histograms of men and women. The figure also shows that the aggregate distribution of men is more symmetric and displays higher kurtosis. Given that the histograms reported by men and women strongly differ in terms of all four moments (and that a large sample size is available), it is not surprising that all tests reject the null hypothesis despite potentially large within-group heterogeneity.

4.4 | Panel conditioning effects

In a recent paper, Kim and Binder (2023) show that inflation expectations of first-time SCE participants differ from those of more experienced participants. In particular, first-time participants expect higher inflation rates (as measured by point forecasts) and report higher uncertainty (as measured by the interquartile range of their probabilistic expectations). This is interpreted as evidence of panel conditioning (or "learning-through-survey") effects. The idea is that respondents are initially quite unfamiliar with the concept of inflation and revise their expectation in subsequent waves after they collected information on the topic (including potentially from the survey questionnaire). Kim and Binder (2023) also document that the evidence of panel conditioning effects is weaker for personal income expectations. They explain this by the fact that respondents are more familiar with their own income situation already when entering the survey panel, and hence,

[12]Figure A.4 of the supporting information shows the sample size and the number of women per survey round.

[13]We obtain nearly identical results if we focus on long-term inflation expectations (*Q9c*). Results are available upon request by the authors.

[14]In particular, households are asked to assign probabilities to the following outcomes for future inflation (in percent): $(-\infty, -12]$, $(-12, -8]$, $(-8, -4]$, $(-4, -2]$, $(-2, 0]$, $(0, 2]$, $(2, 4]$, $(4, 8]$, $(8,12]$, $(12, +\infty)$

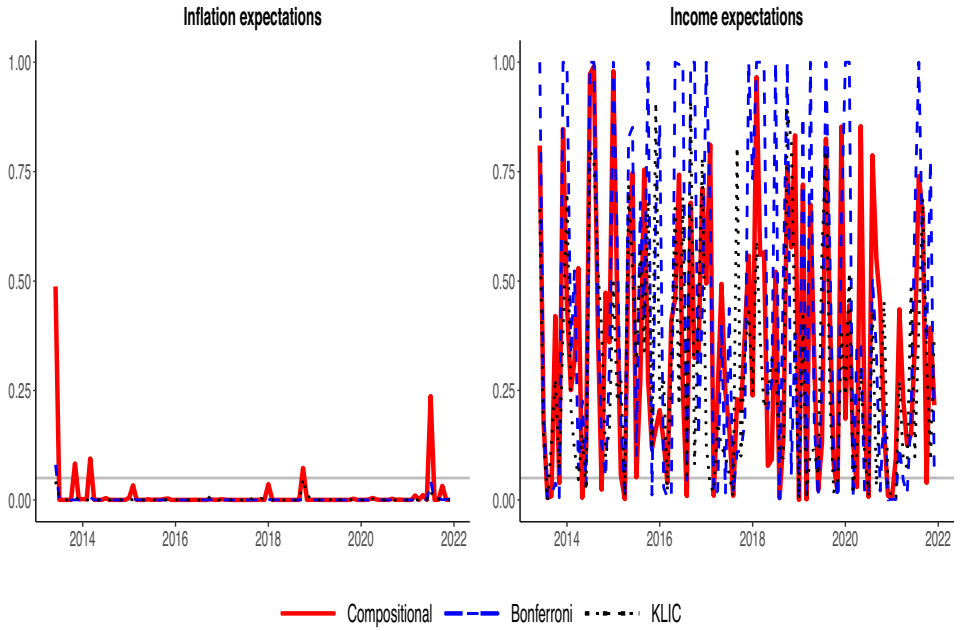

FIGURE 5 *p*-values for heterogeneity tests (SCE): Panel conditioning effects. *Note*: The plot shows the *p*-values from the Hotelling test for compositional data (solid red line), the multiple testing approach (dashed blue line), and the KLIC-based test (dotted black line) for the analysis of panel conditioning effects in inflation expectations (left) and personal income expectations (right). For the multiple testing approach, we report (the minimum of one and) the smallest *p*-value multiplied by the number of bins to make it comparable. The sample period is from June 2013 to December 2021.

the scope for learning is more limited. We supplement this evidence by testing for systematic differences in probabilistic inflation and income density expectations (*Q24*) between first-time survey participants versus more experienced participants. The sample is the same as the one used in the previous application. The left plot in Figure A.4 of the supporting information shows that the number of first-time respondents in each wave is around 150–200 with the exception of August 2013 where 740 new participants entered the SCE. The plots at the bottom of Figure A.5 of the supporting information show the aggregate distributions of both groups.

The left plot of Figure 5 shows clear evidence that first-time participants draw their probabilistic inflation expectations from a different distribution than more experienced panelists. Consistent with the findings of Kim and Binder (2023), Figure A.7 of the supporting information indicates a higher standard deviation for first-time respondents (and a noticeably lower kurtosis). In contrast, the right plot of Figure 5 shows only limited evidence of panel conditioning effects for the income expectations. Indeed, Figure A.8 of the supporting information shows that the differences in histogram moments are much less pronounced for income expectations. In sum, our findings confirm the evidence of panel conditioning effects documented by Kim and Binder (2023) and that such effects are much less pronounced for expectations of variables like personal income growth that households are likely to be very familiar with.

4.5 | Regional differences in national house price expectations

In this final application, we demonstrate that the KLIC-based approach ceases to be feasible in setups where expectations of more than two groups need to be compared. The choice is motivated by the finding of Kuchler and Zafar (2019) that differences in local house price dynamics tend to translate into dispersed forecasts of future nationwide house prices changes, a finding that appears inconsistent with full information rational expectations.

We test for differences of house price expectations across households from the 50 US states and Washington D.C.—or, alternatively, from four broader regions ("West," "Midwest," "Northeast," and "South"; see Figure A.4 for the number of households from each region). In particular, we test the hypothesis that the density expectations from all states (regions) are from the same population in an ANOVA framework. The data are again from the SCE, and we focus on expectations for the change of average house prices nationwide (*C1*). The upper right plot in Figure A.5 of the supporting information shows the aggregate histograms for the different regions. Figure A.9 of the supporting information shows the time series of the moments of the aggregate histograms. The figures do not reveal clear evidence of differences with one exception: The mean of the aggregate histogram for the "West" region is noticeably higher than those for the other regions in the first couple of survey waves.

The plots in Figure 6 present the results based on regions (left plot) and states (right plot). Evidently, there is more time variation in the *p*-values than for the differences in inflation expectations in the two previous applications. For the Hotelling test, we reject the null hypothesis of no differences in the house price expectations across regions (states) for

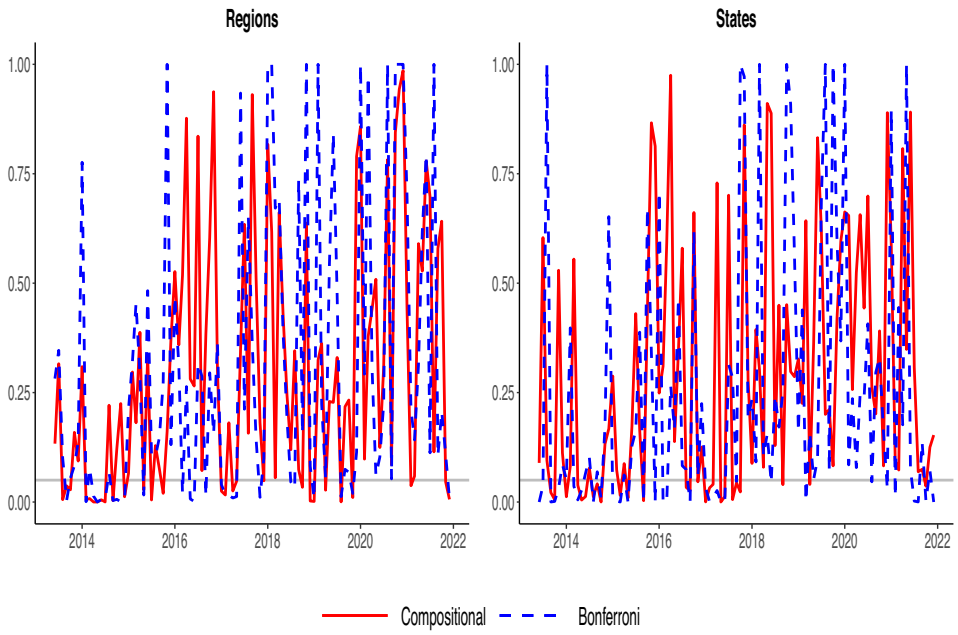

FIGURE 6 *p*-values for heterogeneity tests (SCE): Local differences in house price expectations. *Note*: The plot shows the *p*-values from the Hotelling test for compositional data (solid red line) and the multiple testing approach (dashed blue line) for the analysis of differences in house price expectations across regions (left) or states (right). For the multiple testing approach, we report (the minimum of one and) the smallest *p*-value multiplied by the number of bins to make it comparable. The sample period is from June 2013 to December 2021.

28 (26) of the 103 survey waves. The evidence for the multiple testing approach is similar. The periods with significantly different house price expectations are distributed without any obvious systematic pattern, although particularly for the state-level analysis we observe more differences in the beginning of the sample between 2013 and 2015. This is likely due to the higher mean expectations for the "West" region during this period.

5 | CONCLUSION

We propose a new test for heterogeneity and differences in density expectations. This test builds on the insight that probabilistic survey forecasts are compositional data. For normally distributed data, our Monte Carlo simulations show the superior performance of this test relative to a more traditional bootstrap-based approach using the KLIC as a distance measure between two densities and an approach that involves multiple testing for differences of individual parts of the density. The novel test has high power especially when intragroup heterogeneity is relatively low. For settings that mimic more closely the coarse density expectations observed in many surveys, all tests have very similar power. However, the novel test is always much faster compared to the KLIC-based test because it does not rely on simulations. In addition, it has the additional advantage that it allows for comparisons across more than two groups.

In five applications, we analyze survey-based density expectations of professional forecasters and households. First, we show that the short-term inflation expectations of experts adjusted rapidly in response to rising inflation rates in the euro area after 2021. Long-term expectations were not fully anchored but changed less strongly and more gradually. Second, we find that for most periods, short-run inflation and growth expectations significantly differ between forecasters that round their probability statements and those that do not. Third, we find very strong evidence against the hypothesis that inflation expectations of male and femal household heads are equal, confirming earlier results in the literature based on point forecasts. Fourth, we show that the inflation expectations of households who just entered the survey panel differ from those of more experienced participants. Finally, consistent with a role for local developments and information sets influencing subjective expectations data for aggregate outcomes, we show that for a sizeable fraction of periods in our sample, households from different regions report significantly different density expectations for the future change of nationwide house prices.

Our analysis shows that it is beneficial to treat survey-based density expectations as compositional data. This might be relevant also in other contexts where such survey data are used.

Our results could be extended by using the panel structure of most expectation surveys. So far, we have analyzed each survey wave as separate data samples, but one could also jointly analyze the full sample of expectation data. For instance, by adopting a dynamic model for the compositional expectation data. Furthermore, one could, of course, replace the KLIC by other distance measures—such as the Wasserstein distance (Dobrushin, 1970), the Jensen–Shannon divergence (Lin, 1991), or symmetric versions of the KLIC—to generate alternative benchmark tests.

Lastly, Bassetti et al. (2022) propose a nonparametric alternative for estimating the underlying density expectations from a cross-section of available survey-based histogram forecasts. In principle, that would also be an approach upon which one could build an analysis of heterogeneity across groups. However, for the SPF, the number of available histograms per group is rather small—likely too small for the application of this nonparametric method. We leave a comparison of our method with nonparametric approaches for future research.

ACKNOWLEDGEMENTS

We thank the editor, Michael McCracken, and three anonymous referees for very helpful comments. Our research has also been improved through valuable comments and suggestions by Christian Conrad, Malte Knüppel, and Fabian Krüger. We are responsible for any remaining errors. This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at https://doi.org/10.15456/jae.2024132.1553917996.

DATA AVAILABILITY STATEMENT

We use data from the European Central Bank' Survey of Professional Forecasters (SPF) and the Federal Reserve Bank of New York's Survey of Consumer Expectations (SCE). The data are available from https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/all_data.en.html and https://www.newyorkfed.org/microeconomics/databank.html, respectively. Replication data are available from https://journaldata.zbw.eu/dataset/testing-for-differences-in-survey-based-density-expectations-a-compositional-data-approach.

ORCID

Jonas Dovern https://orcid.org/0000-0003-0890-8809
Alexander Glas https://orcid.org/0000-0003-2229-1112
Geoff Kenny https://orcid.org/0000-0002-4627-1928

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*(2), 139–177. http://www.jstor.org/stable/2345821

Aitchison, J. (1986). *The statistical analysis of compositional data*, Monographs on Statistics and Applied Probability: Springer Netherlands. https://books.google.de/books?id=RHKmAAAAIAAJ

Andrade, P., Coibion, O., Gautier, E., & Gorodnichenko, Y. (2022). No firm is an island? How industry conditions shape firms' expectations. *Journal of Monetary Economics*, *125*, 40–56. https://www.sciencedirect.com/science/article/pii/S0304393221000568

Armantier, O., Bruine de Bruin, W., Topa, G., van der Klaauw, W., & Zafar, B. (2015). Inflation expectations and behavior: Do survey respondents act on their beliefs? *International Economic Review*, *56*(2), 505–536. https://onlinelibrary.wiley.com/doi/abs/10.1111/iere.12113

Armantier, O., Kosar, G., Pomerantz, R., Skandalis, D., Smith, K., Topa, G., & van der Klaauw, W. (2021). How economic crises affect inflation beliefs: Evidence from the Covid-19 pandemic. *Journal of Economic Behavior & Organization*, *189*, 443–469. https://www.sciencedirect.com/science/article/pii/S0167268121001839

Bao, Y., Lee, T.-H., & Saltoğ, B. (2007). Comparing density forecast models. *Journal of Forecasting*, *26*(3), 203–225. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.1023

Bassetti, F., Casarin, R., & Negro, M. D. (2022). A Bayesian Approach to Inference on Probabilistic Surveys. (*Staff Reports 1025*): Federal Reserve Bank of New York https://ideas.repec.org/p/fip/fednsr/94495.html

Binder, C. (2017). Measuring uncertainty based on rounding: New method and application to inflation expectations. *Journal of Monetary Economics*, *90*, 1–12.

Binder, C., Janson, W., & Verbrugge, R. (2023). Out of bounds: Do SPF respondents have anchored inflation expectations? *Journal of Money, Credit and Banking*, *55*(2-3), 559–576. https://onlinelibrary.wiley.com/doi/abs/10.1111/jmcb.12968

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., & Rautanen, T. (2007). The ECB Survey of Professional Forecasters (SPF) - A Review After Eight Years' Experience. (*Occasional Paper Series 59*): European Central Bank https://ideas.repec.org/p/ecb/ecbops/200759.html

Braga, L., & Feingenbaun, D. (2020). Assessing global Covid-19 cases data through compositional data analysis (CoDa). medRxiv.

Buccianti, A. (2018). Water chemistry: Are new challenges possible from CoDA (compositional data analysis) point of view? *Handbook of mathematical geosciences*: Springer, pp. 299–311.

Clements, M. P. (2018). Are macroeconomic density forecasts informative? *International Journal of Forecasting*, *34*(2), 181–198. https://www.sciencedirect.com/science/article/pii/S0169207017301310

Clements, M. P. (2021). Rounding behaviour of professional macro-forecasters. *International Journal of Forecasting*, *37*(4), 1614–1631. https://www.sciencedirect.com/science/article/pii/S0169207021000546

Clements, M. P. (2022). Individual forecaster perceptions of the persistence of shocks to GDP. *Journal of Applied Econometrics*, *36*(3), 640–656. https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2884

Coibion, O., Georgarakos, D., Gorodnichenko, Y., Kenny, G., & Weber, M. (2024). The effect of macroeconomic uncertainty on household spending. *American Economic Review*, *114*(3), 645–677.

Coibion, O., Gorodnichenko, Y., & Kumar, S. (2018). How do firms form their expectations? New survey evidence. *American Economic Review*, *108*(9), 2671–2713. https://www.aeaweb.org/articles?id=10.1257/aer.20151299

Coibion, O., Gorodnichenko, Y., & Ropele, T. (2020). Inflation expectations and firm decisions: New causal evidence. *The Quarterly Journal of Economics*, *135*(1), 165–219. https://doi.org/10.1093/qje/qjz029

Conrad, C., Enders, Z., & Glas, A. (2022). The role of information and experience for households' inflation expectations. *European Economic Review*, *143*, 104015. https://www.sciencedirect.com/science/article/pii/S001429212100283X

D'Acunto, F., Malmendier, U., Ospina, J., & Weber, M. (2021). Exposure to grocery prices and inflation expectations. *Journal of Political Economy*, *129*(5), 1615–1639. https://doi.org/10.1086/713192

Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, *15*(3), 458–486.

Dovern, J., Müller, L., & Wohlrabe, K. (2023). Local information and firm expectations about aggregates. *Journal of Monetary Economics*, *138*, 1–13. https://www.sciencedirect.com/science/article/pii/S0304393223000326

Engelberg, J., Manski, C. F., & Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, *27*(1), 30–41. http://www.jstor.org/stable/27639017

Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis*: Springer.

Fry, J. M., Fry, T. R. L., & McLaren, K. R. (1996). The stochastic specification of demand share equations: Restricting budget shares to the unit simplex. *Journal of Econometrics*, *73*(2), 377–385.

Ganics, G., Rossi, B., & Sekhposyan, T. (2024). From fixed-event to fixed-horizon density forecasts: Obtaining measures of multihorizon uncertainty from survey density forecasts, forthcoming. *Journal of Money, Credit and Banking*. https://onlinelibrary.wiley.com/doi/abs/10.1111/jmcb.13105

Giordani, P., & Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, *47*(6), 1037–1059. https://www.sciencedirect.com/science/article/pii/S0014292102002362

Glas, A., & Hartmann, M. (2022). Uncertainty measures from partially rounded probabilistic forecast surveys. *Quantitative Economics*, *13*(3), 979–1022.

Gupta, N., Rasmussen, C. L., Holtermann, A., & Mathiassen, S. E. (2020). Time-based data in occupational studies: The whys, the hows, and some remaining challenges in compositional data analysis (CoDA). *Annals of Work Exposures and Health*, *64*(8), 778–785.

Kim, G., & Binder, C. (2023). Learning-through-survey in inflation expectations. *American Economic Journal: Macroeconomics*, *15*(2), 254–278. https://www.aeaweb.org/articles?id=10.1257/mac.20200387

Kitano, N., Kai, Y., Jindo, T., Tsunoda, K., & Arao, T. (2020). Compositional data analysis of 24-hour movement behaviors and mental health in workers. *Preventive Medicine Reports*, *20*, 101213.

Kuchler, T., & Zafar, B. (2019). Personal experiences and expectations about aggregate outcomes. *The Journal of Finance*, *74*(5), 2491–2542. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12819

Kumar, S., Afrouzi, H., Coibion, O., & Gorodnichenko, Y. (2015). Inflation targeting does not anchor inflation expectations: Evidence from firms in New Zealand. *Brookings Papers on Economic Activity*, *46*, 151–225.

Kumar, S., Gorodnichenko, Y., & Coibion, O. (2023). The effect of macroeconomic uncertainty on firm decisions. *Econometrica*, *91*(4), 1297–1332.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Lloyd, C., Pawlowsky-Glahn, V., & Egozcue, J. (2012). Compositional data analysis in population studies. *Annals of the Association of American Geographers*, *102*(6), 1251–1266.

Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, *126*(1), 373–416. https://doi.org/10.1093/qje/qjq004

Malmendier, U., & Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics*, *131*(1), 53–87. https://doi.org/10.1093/qje/qjv037

Mankiw, N. G., Reis, R., & Wolfers, J. (2003). Disagreement about inflation expectations. *NBER Macroeconomics Annual*, *18*, 209–248.

Manski, C. F. (2018). Survey measurement of probabilistic macroeconomic expectations: Progress and promise. *NBER Macroeconomics Annual*, *32*(1), 411–471.

Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, *35*, 253–278.

Mitchell, J., & Hall, S. G. (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR fan charts of inflation. *Oxford Bulletin of Economics and Statistics*, *67*, 995–1033. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0084.2005.00149.x

Reiche, L., & Meyler, A. (2022). Making Sense of Consumer Inflation Expectations: The Role of Uncertainty. (*Working Paper Series 2642*): European Central Bank. https://ideas.repec.org/p/ecb/ecbwps/20222642.html

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., & Team, TGEMASP (2012). The concept of compositional data analysis in practice. Total major element concentrations in agricultural and grazing land soils of Europe. *Science of the Total Environment*, *426*, 196–210.

Rich, R., & Tracy, J. (2021). A closer look at the behavior of uncertainty and disagreement: Micro evidence from the Euro area. *Journal of Money, Credit and Banking*, *53*(1), 233–253. https://onlinelibrary.wiley.com/doi/abs/10.1111/jmcb.12728

Weltje, G., & von Eynatten, H. (2004). Quantitative provenance analysis of sediments: Review and outlook. *Sedimentary Geology*, *171*(1-4), 1–11.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Dovern, J., Glas, A., & Kenny, G. (2024). Testing for differences in survey-based density expectations: A compositional data approach. *Journal of Applied Econometrics*, *39*(6), 1104–1122, https://doi.org/10.1002/jae.3080