Fat Tails in Human Judgment: Empirical Evidence and Implications for the Aggregation of Estimates and Forecasts

Miguel Sousa Lobo* Dai Yao†

September 18, 2024

Abstract

How frequent are large disagreements in human judgment? The substantial literature relating to expert assessments of real-valued quantities and their aggregation almost universally assumes that judgment errors follow a jointly normal distribution. We investigate this question empirically using 73 data sets from 4 different sources that include over 169,000 estimates and forecasts. We find incontrovertible evidence for excess kurtosis, that is, of fat tails. Despite the diversity of the analyzed data, varying in how much uncertainty there is in the quantity being assessed and varying in the level of expertise and of sophistication of those making the assessments, we find surprising consistency in the frequency with which an individual judgment is in large disagreement with the consensus. Fitting a generalized normal distribution to the data, we find most estimates for the shape parameter to be between 0.9 and 1.6 (where 1 is the double-exponential distribution, and 2 is the normal distribution). This has important implications, in particular for the aggregation of expert estimates and forecasts and for the construction of confidence intervals. We describe optimal Bayesian aggregation with fat tails, and propose a simple average-median average heuristic (AMA) that performs well for the range of empirically observed distributions.

*miguel.lobo@insead.edu, Associate Professor of Decision Sciences at INSEAD.

†dai@yaod.ai, Associate Professor at The Hong Kong Polytechnic University.

1 Introduction

Laplace first described in the second half of the 18th century a law of errors whereby the frequency with which an error is observed decreases in inverse proportion to the exponential of its square, now commonly known as the normal, or Gaussian distribution. He had previously described another law in which the frequency decreases in inverse proportion to the exponential of the magnitude of the error, what is now called the Laplace, or double-exponential distribution (Kotz et al., 2001). The current prevalent use of the normal distribution over other laws of errors can be understood on two accounts. First, the central limit theorem provides a compelling explanation for the empirical finding that the normal distribution naturally arises in many situations, when the observed uncertainty is the sum of a large number of independent sources of randomness. Second, the normal distribution's mathematical properties make it convenient for analytical and computational reasons: it is conjugate prior to itself, it has maximum entropy for a given variance, the sum of two normal random variables is normal, its multivariate form is preserved under linear transformation, the simple least-squares procedure solves for the maximum likelihood estimate, *etc.* However, the distribution of errors in any given application, be they measurement, estimation, or forecasting errors, should remain an empirical question. In particular, there isn't a strong *a priori* argument to believe that human judgment about uncertain quantities should follow a normal distribution. A perhaps plausible argument involves the judge or expert receiving a large number of different signals in a random fashion, and forming an estimate based on the average of these signals. This is, however, a contrived model, seemingly inconsistent with much research in the psychology of cognition and judgment (*e.g.*, Kahneman, 2011), and requires empirical testing.

While normality is often a good-enough approximation, normally distributed data is the exception rather than the rule. The assumptions of a large number of additive and weakly dependent random terms from which the central limit theorem is derived seldom hold entirely. A testament to the fact that tails are often fatter than in the normal case, that is that excess kurtosis is often present, is the vast literature on robust statistics (*e.g.*, Lye and Martin, 1993; Huber and Ronchetti, 2009). Much attention had been given to this issue in a number of application areas where a large amount of data is available, notably in finance (*e.g.*, Nelson, 1991; Theodossiou, 1998). However, in the substantial literature on the aggregation of expert estimates and forecasts, which has otherwise extensively explored different aspects of the problem, normality is widely assumed. Clemen and

Winkler (1993) already noted that “careful assessment of the distributions’ tails and careful choice of models may be critical [. . .] in aggregation models,” however little to no work exists regarding the shape of the tails of human judgment, and the issue is not mentioned in any of the widely cited reviews by Clemen (1989), Armstrong (2001), or Lawrence et al. (2006). In some cases, such as recent work on judgment aggregation incorporating estimates of the relative level of expertise of each judge, approaches that make no parametric assumptions on the distribution of judgment error are considered in addition to jointly normal models (see, in particular, Palley and Soll (2019)). While the resulting procedures are robust to the shape of the distribution, this does mean they have a “built-in disadvantage” (Palley and Satopää, 2023) relative to approaches with stronger prior assumptions.

Estimating the weight of the tails of a distribution and obtaining information about its fourth moment requires a large sample size. As we will detail, a few hundred samples are required before much can be said in this regard. As a consequence, in any given practical instance where a few judgments are to be aggregated, the shape of the tails cannot be inferred from the data at hand. This makes it impractical, and statistically inefficient, to use a model that allows for an arbitrary tail shape without a strong prior. We can, however, obtain such prior information from similarly-trained judges in comparable tasks. If the shape of the distribution is consistent across different data sources pertaining to similar tasks, that distribution should be used instead of a normal model.

Beyond testing for the presence of excess kurtosis in the distribution of deviations from consensus among judges, our goal is to develop a model with good fit that can be used to derive procedures for the aggregation of individual judgments. We do this working with a generalized normal distribution (GN) with three parameters, for location, scale, and shape. It directly models the decay of the tails with the shape parameter as the exponent of the inverse log-density, and includes the normal and double-exponential distributions as special cases. A number of alternative distribution families are available to model excess kurtosis, most notably Gosset’s Student’s t-distribution. Our choice of the GN is both based on empirical fit, as we find that the generalized normal provides a better fit to the data, and consistent with a modelling strategy which is neutral as to the data generation process. The derivation of Student’s t-distribution is tied to a specific data generation model, with its samples obtained by taking the ratio between the sample mean and the sample standard deviation of a set of independent and identically distributed normal samples. The excess kurtosis in Student's t-distribution is best understood as arising as a result of the occasional draw of a set of normal samples with unusually low variance, which is unrelated to the plausible explanations for excess kurtosis in human judgment. The generalized normal, on the other hand, allows for a functionally simpler family of distributions which directly models the structural characteristics to be estimated, with parameters that are more easily interpreted—in particular a shape parameter that directly specifies the rate of decay of the tails of the probability density.

Our investigation is restricted to estimates or forecasts of real-valued quantities, and to symmetric distributions. In the data sets we analyze, skewness is present in only a few cases: for quantities that are both restricted to be positive and also have a large coefficient of variation. We find that, in these cases, working with the logarithm allows us to retain a common, symmetrical model for all data sets. This is a conservative approach in that, for the transformed data sets, the kurtosis is always larger in the non-transformed data on account of a heavy upper tail.

To increase the validity of our findings, we use multiple data sets with estimates and forecasts from different sources, including economic and financial forecasts, involving different levels of uncertainty about the quantity assessed, and judges with different levels of expertise. If normality does not hold, as our findings establish, then the sample mean is not the optimal estimate of the distribution mean. Further, under an empirically unsuported normality assumption, the confidence interval based on Student's t distribution does not correctly reflect the information contained in the sample. It may either substantially over- or underestimate the uncertainty about the mean.

We discuss procedures for aggregating estimates when the distribution of deviations from the consensus is fat-tailed, including obtaining the appropriate posterior distributions and credible intervals. We provide benchmarks for the performance of different policies, based both on simulation and on empirical data for which the realized values are available. Where the common bias is moderate[1], the empirical benchmarks support the results from simulation and confirm the value of aggregation procedures adapted to the fat-tailed nature of forecasting errors.

Simple heuristic rules have been considered as practical alternatives to Bayesian models, such as the average (Makridakis and Winkler, 1983; Larrick and Soll, 2006), weighted averages (Winkler and

[1] We find the benchmarks based on the economic and financial forecasts in our data to be of little use: due to substantial correlation across judges, the common bias is the dominant source of error. When and how to model and control for such correlation is the subject of an extensive literature; extending that literature for fat-tailed data will require further work beyond our scope here.

Makridakis, 1983; Winkler and Clemen, 1992), and trimmed means (Yaniv, 1997). In addition to the Bayesian estimate, we also assess the performance of some of these heuristics with fat-tailed data and of a new one we propose, the average-median average heuristic (AMA). For distributions that are empirically consistent with judgment data, this new heuristic approximates the optimal Bayesian aggregation. It has robust performance in our judgment data sets, without much degradation for normal data.

The article is organized as follows. In §2 we describe the generalized normal distribution, and discuss the sample size needed to estimate the shape parameter, as well as estimation procedures. Our empirical analysis is in §3. We describe the 73 data sets used, from two panels of economists, from financial analysts, and from MBA student surveys, totaling 169,276 estimates and forecasts. We use quantile-quantile plots for a visual assessment of fit and, for each data set, we use information criteria to determine the model structure with best fit. Based on these models, we then provide estimates for the shape parameter of the generalized normal distribution. We discuss explanations for the fat-tailed nature of the data, including incentives and mixture of normals due to heterogeneity across judges. The results from our analysis suggest that these explanations only partly explain the excess kurtosis, leaving some attributable to the nature of the cognitive processes underpinning judgment. In §4, we consider the problem of combining estimates when the distribution of errors is fat tailed, to obtain point estimates and intervals. We describe the optimal policy, introduce the average-median average heuristic, and report benchmarks of the different policies based on the simulation of different distributional assumptions (normal, double-exponential, and intermediate) as well as based on empirical data sets, under both a linear and a quadratic penalty function. Finally, §5 provides some brief concluding remarks, including recommendations for priors for the shape parameter when dealing with judgment data.

All data, code, and associated documentation for the results and methods described here are available online as an accompanying replication package. This includes Matlab functions for generating pseudo-random numbers from a generalized normal distribution and for estimating the shape parameter of a generalized normal distribution. It also includes functions for the optimal Bayesian aggregation of estimates under generalized normal assumptions given a prior value of the shape parameter: computing a point estimate and a credible interval for the location parameter.

2 The Generalized Normal Distribution

2.1 Definition

Following Nadarajah (2005), we say that a random variable X has a generalized normal (GN) distribution with parameters $\theta = (u, s, p)$ if its probability density function is

$$f(x) = \frac{1}{2s\Gamma\left(1+1/p\right)} \exp\left\{ -\left|\frac{x-u}{s}\right|^p \right\}.$$

The location and scale parameters $u \in \mathbf{R}$ and $s \in \mathbf{R}_{>0}$ have the same units as X. The shape parameter $p \in \mathbf{R}_{>0}$ dictates the 'fatness' of the tails.[2] The first four central moments of the GN distribution are

$$\mathbf{E}X = u, \quad \operatorname{Var} X = \frac{\Gamma(3/p)}{\Gamma(1/p)} s^2, \quad \operatorname{Skewness} X = 0, \quad \operatorname{Kurtosis} X = \frac{\Gamma(1/p)\Gamma(5/p)}{\Gamma(3/p)^2}.$$

We sometimes denote a specific subset of distributions by indexing with the shape parameter: GN_p. With this notation, GN_2 is the normal or Gaussian distribution (with $\sigma = s/\sqrt{2}$), and GN_1 is the double-exponential or Laplace distribution. A uniform distribution is obtained as a limiting case for $p \to +\infty$ (if s scales with $\sqrt{\Gamma(1/p)/\Gamma(3/p)}$ to ensure a finite, non-zero variance).

From $\log f(x) \propto -|x-u|^p$, we see that the GN distribution is log-concave for $p \geq 1$. Since the product of two log-concave functions is log-concave, for observations with GN-distributed errors with $p \geq 1$ any log-concave prior on the location parameter guarantees a log-concave, and therefore unimodal, posterior. If the tails of the distribution of the observation errors are any 'fatter' than those of a double-exponential distribution, we can always construct an example where the posterior of the location parameter is not unimodal. While there is no *a priori* reason to expect the posterior of the location parameter to be log-concave, or even unimodal, where this is supported empirically it has substantial benefits in computational tractability and stability of estimates.

n	GN_1	$\text{GN}_{1.5}$	GN_2
20	0.82	1.34	1.96
50	0.52	0.85	1.24
100	0.36	0.60	0.88
200	0.26	0.43	0.62
500	0.16	0.27	0.39
1,000	0.12	0.19	0.28
2,000	0.08	0.13	0.20
5,000	0.05	0.08	0.12
10,000	0.04	0.06	0.09

Table 1: Expected half width of a 95% confidence interval for the shape parameter p under different distributions and sample sizes.

2.2 On Sample Size and the Need for a Prior for the Shape Parameter

From Bayes' rule, and short of a constant that only depends on the x, the joint posterior distribution of the parameters (u, s, p) given observations $x_1, x_2, \dots, x_n$ is

$$\log f(u, s, p|x) \ \propto\ -n \log s - n \log \Gamma\left(1 + 1/p\right) - \frac{1}{s^p} \sum_{i=1}^{n} |x_i - u|^p + \log f(u, s, p),$$

where $f(u, s, p)$ is the joint prior on the parameters. From this we will see that obtaining an accurate estimate of the shape parameter p, that is estimating the thickness of the tails of the distribution, is challenging. The joint posterior distribution of p and s is such that it is difficult to discriminate between models with a smaller scale parameter and fatter tails *vs.* models with a larger scale parameter and thinner tails. From the Fisher information matrix, the expected half width of a 95% confidence interval as a function of the sample size is given in Table 1.[3] From this we see that, for samples that follow a $\text{GN}_{1.5}$ distribution, on the order of 200 samples are needed to be able to discriminate from a normal distribution (the half width of the interval becomes less than 0.5, the difference between the values of the parameter in the two distributions). Obtaining a

[2] The GN distribution is sometimes called the generalized Gaussian distribution (GG or GGD). The parameters we denote by u, s, and p are sometimes denoted by μ, σ, and s. The scale parameter s is sometimes multiplied by a factor that depends on p to match the standard deviation (which has the drawback of making the integration factor more complex). The alternative parametrization $f(x) = K(c, p) \exp\{c\,|x - u|^p\}$ can also be found.

[3] We use here improper uniform priors for the parameters, and start from the expressions in Nadarajah (2005) in page 693 and integrate numerically as we cannot confirm the correctness of the final expression in page 694.

reasonably narrow confidence interval for the shape parameter p with, say, accuracy to one decimal, requires on the order of ten thousand observations. In the common problem of aggregating, say, ten or fewer expert forecasts, the sample contains little information about the shape of the tails of the distribution of the forecasting errors. We need to rely on prior knowledge, and for this we need an empirical understanding of human judgement.

2.3 Estimating the Shape Parameter

A simple estimation procedure is to find the shape parameter p such that the kurtosis of the generalized normal distribution matches that of the data. An approximation can be quickly computed, by numerical inversion of the formula for the kurtosis of the GN distribution given in §2.1. This approach has, however, limitations and sources of bias. Of greatest concern is the following. As we will describe below, different columns in our data, and sometimes rows as well, can have different means and variances. This is because they are associated with estimates about different quantities, or because they were made at different times. Because of this, the kurtosis-to-shape-parameter matching can only be done after normalizing each column, say to zero sample mean and unit sample variance. This column-wise normalization, especially if there are relatively few observations in each column, can lead to underestimation of tail thickness. The tails will be 'thinned out' because those columns in which observations with a large deviation from the mean happen to occur will be normalized by a larger factor, so that, after the normalization, the points further out in the tail are disproportionately attenuated. While it is possible to control for this bias by constructing a 'disattenuated' estimate, this loses the appeal of simplicity and is all the more difficult to do when there is heterogeneity across both rows and columns. We include in our results estimates based on matching the kurtosis for validation with a simple estimate, but also because in some cases it is difficult to obtain numerical convergence of the Bayesian estimate. To mitigate bias when computing these estimates we discard columns with 10 or fewer observations.

In addition to this simple procedure, we obtain a Bayesian estimate by computing the posterior distribution of the model parameters given the observations, $f(u, s, p|x)$. Given a data-generation model that specifies the distribution of the observations conditional on the model parameters $f(x|u, s, p)$ and a prior on the parameters $f(u, s, p)$, we evaluate by Markov-chain Monte Carlo simulation the posterior distribution of the parameters given the observed data

$f(u,s,p|x) \propto f(x|u,s,p)f(u,s,p)$. From this we obtain the posterior marginal of the shape parameter $f(p|x)$, and report its mean $p_{\text{Bayes}} = \mathbf{E}_{p|x}p$, and a 95% credible interval $[p_{\text{lower}}, p_{\text{upper}}]$ such that $F(p_{\text{lower}}|x) = 1 - F(p_{\text{upper}}|x) = 0.025$, where $F(p|x) = \int_{-\infty}^{p} df(\cdot|x)$.

3 Empirical Tests

3.1 Data

For increased validity of our empirical analysis, we use forecasts and estimates ranging widely in the level of expertise of those providing them, as well as in the degree of uncertainty about the quantity of interest. We use four sources of data, in three different domains: *economic forecasts*, *earnings forecasts*, and *trivia questions*.

1. *Economic Forecasts.* We collected economic forecasts from two sources: the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters (SPF), and the forecasts by a panel of economists collected by the Wall Street Journal (WSJ), all relating to the economy of the United States of America. The SPF forecasts are quarterly for the period of 1968 to 2021, while the WSJ forecasts are monthly for the period of 2003 to 2021. There are 338 professional forecasters who contributed to the SPF panel at least once, and 185 economists to the WSJ panel. They are associated with a variety of institutions. In many cases they work for financial-sector firms, but some panel members have different profiles such as economic consultants, in-house economists at large corporations, and academics. In each set, we collect in rows the forecasts by each forecaster, and in columns the forecasts for each period. When a forecaster did not provide a prediction for a particular number, we put an indicator for missing data. The SPF data includes mostly quarterly forecasts. The exception is for the three-month treasury bill rate, for which estimates of the annual average are also included, which we indicate by an appended 'y'. We construct two data sets for each forecast quantity. The first retains the final estimate from each professional forecaster in the panel prior to the realization of the uncertainty. The second retains forecasts made one year in advance. The three-month treasury bill rate annual average is again an exception, with forecasts made two years in advance also available. One- or two-year-ahead forecasts are indicated by an appended 'a' or 'aa'. Altogether we obtain 105,785 forecasts in the following 23 data sets: Moody's AAA corporate bond yield (BOND, BONDa); Moody's BAA corporate bond yield (BAABOND,

BAABONDa); Housing starts (HOUSING, HOUSINGa); Ten-year treasury bond rate (TBOND, TBONDa); Unemployment rate (UNEMP, UNEMPa); Nominal gross domestic product (NGDP, NGDPa); Nominal corporate profits after tax (CPROF, CPROFa); Nonfarm payroll employment (EMP, EMPa); Real personal consumption expenditures (RCONSUM, RCONSUMa); Three-month treasury bill rate (TBILL, TBILLa, TBILLy, TBILLya, TBILLyaa). The WSJ forecasts are on a monthly, quarterly, or annual basis, as indicated for each data set when necessary. An exception to this is non-farm payroll, where the forecasts are for the average of the next four quarters. Where possible, we construct up to three data sets for each forecast quantity: one retains the final estimate from each economist prior to the realization of the uncertainty; the other two retain forecasts made one and two years in advance. The data sets labeled with an appended 'a' or 'aa' are the one- or two-year-ahead forecasts. Altogether we obtain 38,829 forecasts in the following 24 data sets: Gross domestic product, quarterly and yearly (GDP, GDPa, GDPy, GDPya, GDPyaa); Non-farm payroll, average of next four quarters (NFARM); Consumer price index, monthly (CPI, CPIa, CPIaa); Rate on 10-year notes, monthly (R10Y, R10Ya, R10Yaa); Change in home price, yearly (CIHP, CIHPa, CIHPaa); Unemployment rate, monthly (UNEMP, UNEMPa, UNEMPaa); Rate on federal funds, quarterly (FEDFUNDS, FEDFUNDSa, FEDFUNDSaa); Housing starts, yearly (HOUSINGSTARTS, HOUSINGSTARTSa, HOUSINGSTARTSaa).

2. *Earnings Forecasts.* We collected from the Institutional Brokers' Estimate System (I/B/E/S) the estimates made by financial analysts of the future earnings per share (EPS) of publicly traded American companies. Analysts make forecasts on annual EPS of the companies they cover throughout the fiscal year, and revise the forecasts as quarterly EPS data are released. For each stock ticker and in each fiscal year, we retain the forecasts made in the 10th month of the fiscal year. This ensures that they are made after the third quarterly EPS is announced, but not too close to the end of the fiscal year. When an analyst makes multiple revisions to the forecast during this month, we retain the last revision. As before, in each set, we collect in rows the forecasts by each analyst, and in columns the forecasts for each fiscal year. When an analyst did not provide a particular forecast, we put an indicator for missing data. We selected companies for which there were a minimum 500 forecasts in the 10th month of the fiscal year, by more than 10 different analysts. This resulted in 13,004 forecasts for 21 companies, in the period from 1981 to 2021. We label the data sets according to the I/B/E/S ticker of the company. The are categorized in five different main business sectors:

information technology (AMZN, AAPL, HPQ, INTC, MSFT, MU, QCOM, XLNX, CHKP), energy (APC, APA, BHI, EOG, HAL, PDP, SLB), pharmaceutical (AMGN), media and entertainment (DIS), and manufacturing (BA, TXN, CAT).

3. *Trivia Questions.* In a required course in a business school MBA program, students were given judgmental exercises as part of a larger survey designed to illustrate cognitive and behavioral biases. We collected the survey responses from 3,895 students taking the course over a period of ten years, taught by six different faculty. The class was delivered in campuses in both Europe and Asia, and the students were of a wide range of nationalities, none of which accounted for more than 6% of the population. From each survey, we make use of three questions that required students to estimate or forecast a quantity. We collect in each column all responses to a particular question from students who completed the survey in the same week and with the same instructor, which allows us to test for, and if necessary control for, any context effects such as classroom discussion prior to the survey or different topics in the news at the time. A total of 11,685 estimates are included in five data sets as follows. Students were asked to estimate the number of member countries of the United Nations. Half of each class was in a low anchor condition, and half in a high anchor condition. The anchor was created by a hint in the text of the question (95 or 300 countries). We separate the responses with low and high anchor condition into two data sets (UNLO and UNHI). While a foreign exchange rate forecast is not usually classified as trivia, it can be considered as such when asked of non-specialists. Students were asked to forecast an exchange rate, typically one-year-ahead USD-EUR. Half of the students were put in a low anchor condition in a preceding question, which asked whether they thought the exchange rate would be above or below a stated number lower than the exchange rate at the time. The other half of the students were, in likewise fashion, anchored on a number higher than the exchange rate at the time. We separate the responses with low and high anchor condition into two data sets (FXLO and FXHI). Another question included in the surveys asked students to estimate a large number about which they had scant knowledge. The responses ranged over several orders of magnitude. Examples include the number of eggs produced in the USA in one year, the surface area of the Vatican, and the current market value of one day of the world's oil production. We group all these questions in one data set (SCALE).

3.2 Data-Generation Model

The observations are organized in rows i and columns j. Each row i corresponds to a judge. In the economic and financial forecasts this means the same person over all columns, but not so in the MBA survey responses. Each column j corresponds to a different quantity being estimated or forecast.

The data-generation model for the Bayesian estimates is as follows. The observations are assumed to be independently drawn from generalized normal distributions,

$$X_{ij} \sim \mathrm{GN}(u_{ij}, s_{ij}, p_{ij}).$$

We consider different model structures for the location, scale, and shape parameters. For each data set, we test these models for fit against the observations. We consider three alternatives for the location parameter, as follows.

- A single common location parameter for all observations, $u_{ij} = u$.
- Location parameters for each column, $u_{ij} = u_j$.
- Location parameters both for each column (the consensus estimate) and for each row (the judge bias), with the location parameter for the distribution of each observation equal to the sum of the corresponding column and row location parameters, $u_{ij} = u_j + v_i$.

For the scale parameter, we consider the following model structure alternatives.

- A single common scale parameter for all observations, $s_{ij} = s$.
- Scale parameters for each column, $s_{ij} = s_j$.
- Both column- and row-specific scale parameters, s_j and t_i. The s_j model the uncertainty due to the inherent difficulty in estimating or forecasting a particular quantity. The t_i model the uncertainty specific to a particular judge, which may be due to ability, effort, or access to information. When we model both column- and row-specific scale parameters, s_j and t_i, there appears to be no obvious default rule in the literature for combining the two scale parameters. We considered and tested four different rules: the geometric average $s_{ij} = \exp((\log s_j + \log t_i)/2) = \sqrt{s_j t_i}$, as well as rules based on the ℓ_1, ℓ_2, and ℓ_∞ norms, $s_{ij} =$

$(s_j + t_i)/2$, $s_{ij} = \sqrt{(s_j^2 + t_i^2)/2}$, and $s_{ij} = \max(s_j, t_i)$.[4] Based on information criteria, we found the geometric average to provide the best fit (in the data sets where we consider both column- and row-specific scale parameters; for more on the information criteria and model fit see §3.3).

Finally, for the shape parameter, we consider the following.

- A single common shape parameter for all observations, $p_{ij} = p$.
- Shape parameters for each column, $p_{ij} = p_j$.
- Shape parameters for each row, $p_{ij} = p_i$.

Not all 27 combinations of the three sets of three alternatives above make sense. We only test row-specific parameters when the identity of judge i is the same across all columns. Further, we only test column-specific scale parameters if also including column-specific location parameters, only test column-specific shape parameters if also including both column-specific location and scale parameters, and likewise for row-specific parameters.

The prior distributions are as follows, and mutually independent except where noted. The location parameters are assumed to be drawn from an improper uniform prior, $u \sim \text{Uniform}(-\infty, +\infty)$. When a single scale parameter is used, its logarithm is assumed to be drawn from an improper uniform prior, $\log s \sim \text{Uniform}(-\infty, +\infty)$, that is $f(s) \propto 1/s$. This ensures scale independence of results, that is, the estimates do not depend on the units used. In models with multiple s_i, a hierarchical prior is required to ensure a proper posterior. We use the hierarchical prior $\log s_i \sim \mathcal{N}(\mu_{\log s}, \sigma^2_{\log s})$, again with improper $\text{Uniform}(-\infty, +\infty)$ priors on $\mu_{\log s}$ and on $\log \sigma_{\log s}$ (Gelman, 2006). Finally, we set a uniform prior on the shape parameters $p \sim \text{Uniform}(0.1, 4)$. The

[4] These four rules correspond to different assumptions about the way in which the accuracy of the judge (due to skill, effort, or access to information) interacts with the uncertainty about the quantity to be estimated (due to inherent randomness, or difficulty of acquiring relevant information). The geometric average corresponds to assuming a multiplicative effect, that is, that the judge's shortcomings proportionally magnify the uncertainty inherent to the quantity. The rules based on the ℓ_1 and ℓ_2 norms imply additive assumptions, respectively, on the standard deviations and on the variances associated with the judge and with the number to be estimated. The maximum rule corresponds to the assumption that the ability to obtain an accurate estimate is limited by the highest of the two variances, associated with the judge and with the quantity to be estimated.

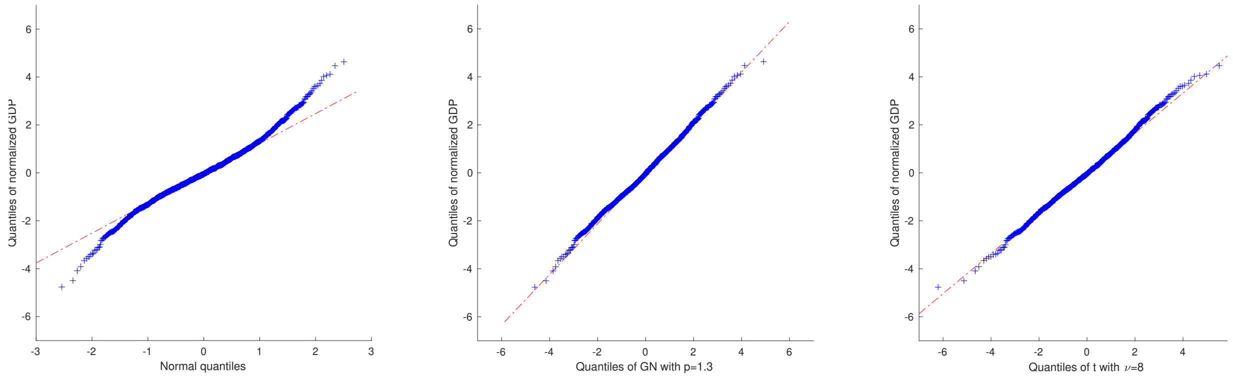

Figure 1: Quantile-quantile plots of USA gross domestic product growth forecasts by economists (after each quarter's forecasts were normalized to have zero mean and unit variance), against a normal distribution, against a generalized normal distribution with shape parameter $p = 1.3$, and against a Student's t-distribution with the number of degrees of freedom $\nu = 8$.

choice of bounds is arbitrary, as long as it excludes improper posteriors near zero and for large p, but is wide enough to not impact the estimates[5].

3.3 Model Fit and Selection

The MBA data sets had a couple of obvious outliers which appeared to be due to transcription errors. Some of these were manually removed, but we also applied a general rule to all data sets removing any point over a number of standard deviations away from the mean, using the sample mean and sample standard deviation for each column of data. The estimates are robust to using four, five or six standard deviations as the threshold[6]. We present results using five standard deviations, and report the comparison using different thresholds in Section **??**. To the extent that this rule may lead to false positives in the detection of errors in the data, our estimates of the thickness of the tails may be conservative.

Figure 1 shows quantile-quantile plots for a sample data set, the one-quarter-ahead forecasts of USA GDP from the panel of economists, which is broadly representative of the pattern seen in

[5] In determining how wide the interval needs to be to not impact the estimates, we were guided by the prior knowledge provided by the simple estimates based on the kurtosis, which we obtained prior to the construction of the Bayesian model.

[6] Correlations between estimates across all data sets using the different thresholds range from 0.97 to 0.99.

other data sets. The plots are against the quantiles of the normal distribution, of a generalized normal with shape parameter $p = 1.3$, and of a Student's t with degrees of freedom $\nu = 8$, with the parameters chosen by visual fit. The fat-tailed nature of the data relative to the normal distribution is evident, resulting in a poor fit. The fit appears adequate with both GN and Student's t, with GN having the better fit.

Regarding the alternatives for model structure described in §3.2, to determine the model with best fit, we consider both the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). As discussed below, we balance these two criteria in our choice of model. The BIC, which is more conservative in that it puts a stronger penalty on the number of parameters. As a general consideration, because we are interested in generalizable results rather than in explaining a particular data set, when in doubt we lean towards model parsimony to avoid overfitting. For some data sets we fit two models, one supported by each criterion, which also allows us to investigate the effect of controlling for heterogeneity across judges. The maximum-likelihood estimates are obtained by numerical optimization[7], with p constrained to the interval $[1, 2]$ for numerical robustness (if the optimization is not constrained away from the region where the log-likelihood is not concave in the location parameter, then the problem is not globally convex and the numerical optimization sometimes converges to a sub-optimal local optimum). The model fit results are consistent with our understanding of the data.

For the economic and financial forecasts, we consider models that allow for heterogeneity across judges. Table 2 gives detailed results for the one-quarter-ahead forecasts in GDP, which are representative of the results over the different economic and financial forecast data sets. Different location parameters for each period are always supported, and different scale parameters are supported in most cases. The AIC supports heterogeneity across forecasters, either in the location parameter only (different bias) or in both the location and scale parameters (different bias and different accuracy). The BIC does not support heterogeneity across forecasters. Both the BIC and the AIC always support a single shape parameter. In light of this, when feasible, we will provide two estimates of the shape parameter for economic data sets. The first estimate is based on a model with different location and scale parameters for each column, but common to all rows. The second estimate is based on a model that allows for different location and scale parameters both for each

[7] With the the Matlab Optimization Toolbox implementation of sequential quadratic programming.

Model Parameters	N	k	$\log f(x\|\theta^*)$	AIC	BIC
$\theta = (u, s, p)$	3873	3	-6557	13121	13140
$\theta = (u_1, \ldots, u_n, s, p)$	3873	72	-2115	4373	4824
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p)$	3873	141	-1076	2433	$3316^\ddagger$
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p_1, \ldots, p_n)$	3873	210	-1019	2457	3772
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s, p)$	3873	257	-1730	3973	5583
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, p)$	3873	326	-612	1875	3917
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, p)$	3873	511	-130	$1282^\dagger$	4481
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, p_1, \ldots, p_n)$	3873	580	-118	1397	5029
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, q_1, \ldots, q_m)$	3873	695	-86	1563	5915

Table 2: Model selection for USA's gross domestic product forecasts (GDP) from the WSJ panel of economists. The model with best fit according to AIC (†) has different location and scale parameters for each quarter and for each economist, and a common shape parameter. The model with best fit according to BIC (‡) has a different location and scale parameter for each quarter, and a common shape parameter.

column and for each row. As we note in subsequent discussion, this allows us to investigate whether the excess kurtosis can be explained by a mixture of normal distributions with different variances arising from heterogeneity across judges. However, to be able to control for both column and row effects requires that there be a large number of judges from each of which we have a large number of forecasts. This was only the case for the larger data sets, GDP, GDPa, and NFARM. For the earnings forecasts, while the model selection results are similar to those for the economic forecasts, obtaining posterior estimates of the parameter for these data sets is numerically challenging, as we will discuss. We therefore restrict their analysis to the estimates based on the kurtosis.

Within the trivia questions data sets, FXLO, FXHI, and SCALE share a similar pattern. Given that the foreign exchange forecasts were collected at different times and in some cases about different currency pairs, different columns of data have significantly different means. Similarly, the SCALE data set includes responses to a diverse set of questions, and different means are supported. Since we work with the logarithm of the estimates, the scale parameter is a measure of the coefficient of variation of the original data, that is the ratio of the standard deviation to the mean. This explains the somewhat weak evidence we find for different scale parameters for each column. The BIC supports different scale parameters in FXLO and SCALE, but not so in FXHI. It does not support different shape parameters. The only case where there is (weak) support for different shape parameters, is with the AIC in FXLO. Overall, given the weight of evidence, we model FXLO, FXHI, and SCALE with separate location and scale parameters for each column, and with a single shape parameter. For UNLO and UNHI, the estimates were collected over a period of ten years during which the number of member countries of the United Nations hardly changed. The demographic characteristics of the MBA student body were also stable. For these data sets, the BIC and AIC both support a single location, scale and shape parameters common to all columns, which is the model we use. That is, this confirms there is no significant impact from different instructors' approach to the class, nor from changes in the world news context.

3.4 Estimates of the Shape Parameter and Discussion

Summary statistics for the data and estimates and credible intervals for the shape parameter p are reported in Tables 3-7, where n is the number of columns and N the number of data. For the models that only include column effects, we deleted columns with less than 10 estimates. For the models that control for both column and row effects, we iteratively deleted columns and rows until each column and row had at least 8 estimates. The reported skewness and kurtosis are after normalizing each column of data to have zero mean and unit variance. The estimate p_{kurt} is obtained by the method of matching this normalized sample kurtosis to the kurtosis of the GN distribution. The Bayesian estimate p_{Bayes} is the mean of the posterior marginal distribution of p. The credible interval for p is from the 2.5% and 97.5% quantiles of this distribution.

The posterior distributions are obtained from our implementation of Markov-chain Monte Carlo integration with Metropolis-Hastings sampling. Two million trials are computed for each problem. In the first half of the chain, the covariance of the jump distribution is progressively adapted to match the covariance of the existing samples. The jump distribution is also scaled to keep the acceptance rate between 25% and 33%. In the second half of the chain, the jump distribution is fixed. The first half of the chain is discarded, and the second half subsampled every 200 trials, resulting in 5,000 samples of the posterior. These samples are checked for autocorrelation, and we only report results when it is below .25 with a 10-sample lag. In the cases where we report, the autocorrelation is usually close to zero.

In some cases, we could not obtain reliable numerical estimates of the posterior distribution of p. This was the case for data sets with very high kurtosis where the posterior distribution may not be log-concave, especially when this is compounded by having a large number of columns, and therefore a high-dimensional parameter space for the Markov chain. For these cases we present only the estimate based on matching the kurtosis. This is the case, notably, for the earnings estimates. In the data sets where we obtain both estimates, the simple estimate based on the kurtosis underestimates the weight of the tails, usually overestimating the shape parameter by around 0.1.

We verified our implementation by generating pseudo-random observations from a normal distribution (the null hypothesis) to create data sets of sizes similar to our empirical data sets. Running the estimation procedure on these data sets, the credible intervals included $p = 2$ with a frequency consistent with the interval's probability (no bias in rejecting the null).

As is true for concerns regarding correlation across judges and common bias, the issue of modeling skewness is somewhat problem specific, and not central to our analysis. We find that in all the data sets we used, skewness is only significant for quantities that are restricted to be positive and that also have a large coefficient of variation. That is, the distribution of the estimates is positively skewed only for positive quantities where the uncertainty is large relative to the quantity being estimated. After taking the logarithm of the estimates in these cases, skewness is not a concern in any of our data sets. The most extreme case is the data set of MBA survey items with scale uncertainty (SCALE), for which the skewness of the original data is 60.1, and the skewness of their logarithm is 0.42. While this may not be a generally applicable approach, working with the logarithm of the estimates in these cases allows us to work with a common modelling approach for all data sets. This is a conservative approach insofar as the non-transformed data always has higher kurtosis on account of its upper tail.

We find overwhelming evidence that, in the aggregate, human judgment is fat tailed. We also find a notable consistency across different tasks, across different degrees of uncertainty about the quantity being assessed, and across different levels of expertise on the part of the judges. Over all 36 data sets for which we obtained a Bayesian estimate, the average estimate of the shape parameter is 1.23. The 10th and 90th percentiles are 0.92 and 1.57, and the median 1.22. The minimum estimate

Data set	N	n	Skew.	Kurt.	p_{kurt}	k	p_{Bayes}	p_{low}	p_{high}
BOND	4548	157	-0.13	5.41	1.08	315	1.50	1.45	1.55
BONDa	4455	157	-0.05	3.41	1.68	315	1.45	1.36	1.54
BAABOND	1201	44	0.40	5.16	1.12	89	0.91	0.80	1.01
BAABONDa	1188	44	0.45	3.28	1.77	89	1.48	1.28	1.69
HOUSING	7651	208	0.11	4.44	1.27	417	1.21	1.15	1.26
HOUSINGa	7237	203	0.41	4.40	1.28	407	1.17	1.12	1.23
TBOND	4299	116	0.38	5.37	1.08	233	1.06	1.01	1.11
TBONDa	4132	116	0.34	3.64	1.55	233	1.40	1.31	1.50
UNEMP	8069	208	0.17	4.19	1.34	417	1.30	1.24	1.36
UNEMPa	7624	203	-0.06	3.84	1.47	407	1.55	1.49	1.62
NGDP	7972	208	-0.18	5.31	1.09				
NGDPa	7524	203	-0.47	5.38	1.08				
CPROF	6184	208	0.07	4.39	1.29				
CPROFa	5805	203	0.01	3.92	1.44				
EMP	2439	69	-0.23	7.17	0.89				
EMPa	2430	69	-0.40	4.87	1.17				
RCONSUM	5348	157	0.30	5.96	1.00				
RCONSUMa	5171	157	0.14	4.71	1.21				
TBILL	5357	157	0.76	5.78	1.03				
TBILLa	5200	157	0.45	4.15	1.36	315	1.17	1.10	1.24
TBILLy	1342	40	0.59	4.85	1.18	81	1.10	0.98	1.22
TBILLya	332	12	0.70	4.74	1.20				
TBILLyaa	277	12	0.20	3.79	1.49	25	1.26	0.95	1.64

Table 3: Estimates of the GN shape parameter p for the SPF panel of economic forecasters.

Data set	N	n	Skew.	Kurt.	p_{kurt}	k	p_{Bayes}	p_{low}	p_{high}
GDP	3873	70	0.17	4.5	1.26	141	1.23	1.15	1.31
GDPa	2354	42	-0.15	5.5	1.07	85	1.12	1.03	1.21
GDPy	765	14	0.67	5.8	1.02				
GDPya	855	14	0.38	5.0	1.15	29	1.22	1.07	1.38
GDPyaa	594	10	-0.03	5.8	1.03	21	1.12	0.96	1.29
NFARM	8925	170	0.25	4.8	1.18	341	1.11	1.07	1.13
CPI	1921	35	-0.02	5.5	1.07	71	0.93	0.82	1.05
CPIa	1883	33	-0.09	4.3	1.31	67	1.33	1.22	1.46
CPIaa	1030	20	-0.16	4.4	1.28	41	1.21	1.06	1.37
R10Y	1460	27	-0.13	5.1	1.10	55	1.08	0.94	1.23
R10Ya	1531	27	0.30	4.4	1.29	55	1.33	1.20	1.47
R10Yaa	951	19	0.24	3.3	1.73	39	1.55	1.34	1.80
CIHP	628	14	-0.31	4.4	1.29	29	1.23	1.03	1.45
CIHPa	676	14	-0.35	4.9	1.16	29	1.23	1.04	1.43
CIHPaa	447	9	-0.31	3.7	1.52	19	1.59	1.30	1.93
UNEMP	1870	34	0.18	4.9	1.18				
UNEMPa	1893	33	-0.07	5.5	1.07	67	1.04	0.93	1.14
UNEMPaa	1041	20	0.04	5.7	1.04	41	0.99	0.88	1.12
FEDFUNDS	1532	28	-0.85	13.4	0.64				
FEDFUNDSa	1609	28	1.07	9.1	0.77				
FEDFUNDSaa	1014	19	0.51	5.9	1.01				
HOUSINGSTARTS	734	14	0.10	10.6	0.71				
HOUSINGSTARTSa	798	14	0.34	5.2	1.12	29	0.92	0.79	1.04
HOUSINGSTARTSaa	445	8	0.02	4.5	1.27	17	1.18	0.96	1.42

Table 4: Estimates of the GN shape parameter p for the WSJ panel of economists.

Data set	N	n	Skew.	Kurt.	p_{kurt}
AMZN	647	22	1.99	9.7	0.75
AAPL	889	31	0.58	5.8	1.02
HPQ	514	26	-0.20	6.9	0.91
INTC	681	26	-0.37	6.0	1.01
MSFT	829	33	-0.65	6.1	0.99
MU	511	24	-0.76	5.8	1.02
QCOM	590	23	-1.68	10.2	0.73
XLNX	502	25	1.15	8.0	0.83
CHKP	530	22	-0.61	6.8	0.92
APC	668	30	-0.29	5.3	1.09
APA	603	28	-0.55	6.1	0.99
BHI	508	22	-0.80	4.8	1.20
EOG	541	22	-0.26	5.8	1.02
HAL	680	31	-0.41	5.9	1.02
PDP	511	22	-0.91	6.7	0.93
SLB	732	36	-0.01	7.0	0.90
AMGN	549	29	-0.15	5.2	1.12
DIS	627	33	0.22	5.0	1.15
BA	552	31	-0.13	5.3	1.09
CAT	514	28	-0.22	5.9	1.01
TXN	826	33	-0.10	6.1	0.99

Table 5: Estimates of the GN shape parameter p for earnings forecasts.

Data set	N	n	Skew.	Kurt.	p_{kurt}	k	p_{Bayes}	p_{low}	p_{high}
UNHI	2017	18	-0.91	5.3	1.09	3	0.95	0.86	1.04
UNLO	2022	18	-0.72	4.1	1.36	3	1.52	1.41	1.64
FXHI	2003	18	0.38	3.5	1.64	39	1.69	1.54	1.85
FXLO	2025	18	-0.27	4.1	1.36	39	1.35	1.22	1.49
SCALE	3618	34	0.42	3.9	1.43	71	1.66	1.55	1.77

Table 6: Estimates of the GN shape parameter p for trivia questions.

				Homogeneous judges				Heterogeneous judges			
Data set	N	m	n	k	p_{Bayes}	p_{low}	p_{high}	k	p_{Bayes}	p_{low}	p_{high}
GDP	3682	125	70	141	1.23	1.15	1.31	395	1.81	1.65	1.96
GDPa	2158	99	42	85	1.12	1.03	1.21	287	1.67	1.48	1.89
NFARM	8783	127	170	341	1.17	1.12	1.22	599	1.64	1.55	1.74

Table 7: Estimates of the GN shape parameter p for economic forecasts, controlling for heterogeneity across forecasters.

is 0.84, and the maximum 1.67. The estimates based on the kurtosis are broadly consistent with this. Over all 73 data sets, the average estimate of the shape parameter based on matching the kurtosis is 1.15, with 10^{th} and 90^{th} percentiles of 0.89 and 1.46. Over the 36 data sets for which we have a Bayesian estimate, the average estimate of the shape parameter based on the kurtosis is 1.30. We don't find evidence that the distributions of judgments from MBA students and from professional economists have fundamentally different shapes, with average estimates of 1.20 and 1.37. The earnings estimates do seem to be consistently fatter tailed, with an average estimate of 0.99. In all 73 data sets, normality is excluded at the 95% confidence level.

The data from the panel of economists might, *a priori*, be expected to be a stringent test for the hypothesis that judgment is fat tailed, given that the uncertainties involved are the subject of much study, econometric methods are benchmarked in the literature, and there are a large number of sources of information and leading statistics to draw from, so that some approximation of the conditions for the central limit theorem might be expected to hold.

We consider two explanations for the fat-tailed nature of the data that are external to the judges: heterogeneity and incentives. Judges are likely to be heterogeneous in the variance of their estimates, which may be due to some judges being better informed than others, to differences in skill, or simply to differences in effort. These differences can also arise, over time, for different estimates from the same judge, since access to information, skill, and motivation are not static. This heterogeneity can have different implications for the shape of the resulting mixture. If judges are more heterogeneous in their variances than in their means, the mixture will have fatter tails. If judges are more heterogeneous in their biases than in their variances, the mixture can have thinner tails. In the larger data sets from the panels of economists (GDP, GDPa, and NFARM), we explicitly control for such heterogeneity by allowing each economist to have a different location parameter (or bias), and a different scale parameter (or standard deviation around consensus plus bias). The pattern of results is not consistent with a mixture of normals on this dimension as the sole explanation. After controlling for heterogeneity the shape parameter is larger, but the distribution of each judge's estimates still has significantly fatter tails than a normal distribution.

The second explanation from external causes is response to incentives: some economic forecasters may be motivated to release estimates away from the consensus. This will arise if the perceived cost-benefit calculus of professional rewards is such that, in expected-utility, the prestige and associated rewards from being 'the only one who got it right' when everyone else was very wrong is greater than the penalty of being very wrong (Fang and Yasuda (2014) argue for evidence of such effects in the context of financial analysts). However, this explanation is unsatisfactory for the results from the MBA surveys where, since they were completed anonymously, such career incentives do not play a role. It is also conceivable that calculations of professional incentives, to the extent that they play a role in the panel of economists, should be less prominent for longer-term forecasts. For the same economic numbers, we don't find a pattern of one-year-ahead forecasts being consistently more or less fat tailed that the shorter-term forecasts.

The remaining explanations are internal to the judges, including the nature of cognitive processes. Fat tails might be a consequence of our propensity to rely on simple heuristics and susceptibility to biases, including anchoring and availability. Further, people may internalize and act according to the logic of high rewards from being 'the only one who got it right', or display 'pushing away' effects as identified by Rader et al. (2015).

Whatever the relative contribution of each of these explanations, for how to do the aggregation when the distribution of the available estimates is known to be is fat-tailed. Based on this empirical evidence, in a practical problem of aggregating expert estimates, we might consider using a prior for the shape parameter such as $p \sim \text{Uniform}(1.0, 1.5)$. However, the required procedure is significantly less complex if p is taken as fixed (which is to say, if we use single mass point as the prior). We next investigate the performance of such estimates under model mis-fit, and assess alternative policies and a proposed heuristic.

4 Combining Estimates and Forecasts with Fat-Tailed Errors

4.1 Problem Specification

Defining optimality of an aggregate point estimate requires an assessment of the economic cost of an incorrect estimate or forecast, as well as of risk preferences. From the point of view of the person doing the aggregation, prior beliefs combined with the observation of the individual judgments imply a posterior distribution for the location parameter. An estimate is then optimal in the sense that it minimizes the expected cost or expected disutility of the estimation error, with the expectation calculated over the posterior distribution of the location parameter. While a quadratic penalty function is widely used, likely as a legacy of least-squares and its computational simplicity, a linear penalty, or absolute deviation, may be a better default choice, with more sound economic justification in many problems. We consider here both linear and quadratic penalty functions, which we also refer to as mean absolute deviation (MAD) and root mean square (RMS).

While the maximum likelihood, or maximum-*a-posteriori* (MAP), estimate can also be computed from the posterior distribution of the location parameter (and, as shown below, in the Laplace and normal cases its computation is trivial), this may not be a good choice as it does not in general minimize the expected cost of the estimation error.

In many applications there is an interest not just in a point estimate but in the entire posterior distribution, to be used in a broader risk model, which should also include an assessment of the probability distribution for the common bias over all judges (there is a substantial literature on correlated experts, for which see the previously cited review articles). A credible interval can be computed from the posterior distribution to summarize the uncertainty about the location parameter given the observations at hand. Student's t-distribution corresponds to the normal case with improper uniform priors, but the general case requires numerical computation.

4.2 On Credible Intervals

Bayesian confidence intervals, or credible intervals, should be consistent both with the data and with our prior understanding of the distributional characteristics of judgment errors. Using a normal model when errors are fat tailed can lead to intervals that are either significantly wider or significantly narrower than appropriate. With a normal model, the interval depends only on the sample mean and variance, and is not impacted by the higher moments of the sample. With, say, shape parameter $p = 1$ (a double-exponential model), the credible interval will tend to be narrower when outliers are present, and wider if the sample at hand happens to have thin tails.

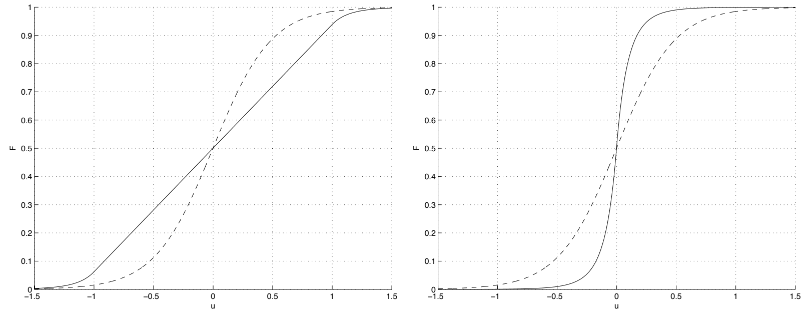

Figure 2: Posterior marginal cumulative distributions of the location parameter with a normal model of errors (dashed) and with a Laplace model of errors (solid), (a) with sample $x_a = [\ -1,$ $-1,\ -1,\ -1,\ +1,\ +1,\ +1,\ +1\ ]$, and (b) with sample $x_b = [\ -2,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ +2\ ]$.

By way of illustration of the importance of using the correct model, consider two different sets of eight observations, $x_a = [\ -1,\ -1,\ -1,\ -1,\ +1,\ +1,\ +1,\ +1\ ]$ and $x_b = [\ -2,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ +2$]. The samples x_a and x_b have the same sample mean, variance, and skewness, but their sample kurtosis differs by a factor of four. Figure 2 plots the cumulative posterior distributions. With the normal model, the location parameter's posterior distribution (Student's t) is the same whether x_a or x_b was observed. However, if observations are known to be fat tailed, the two sets of observations have very different implications for posterior beliefs. With a double-exponential model the interval widths differ by a factor of more than ten, depending on whether x_a or x_b was observed. For x_a the credible interval with the normal model is 43% of the width of the credible interval with the Laplace model, while for x_b this ratio is 459%.

Figure 3 plots the distribution of the ratio between the widths of the credible intervals, based on a normal model versus a double-exponential model, for a sample size of 10 and when the data are draw from a double-exponential distribution. Using a normal model for fat-tailed data leads to credible intervals that are, on average, overly wide. The frequent presence of outliers which would have very low likelihood under a normal distribution leads the normal model to overestimate the variance of the data. Informally, x_b in the example above is more representative of what usually happens than x_a is.

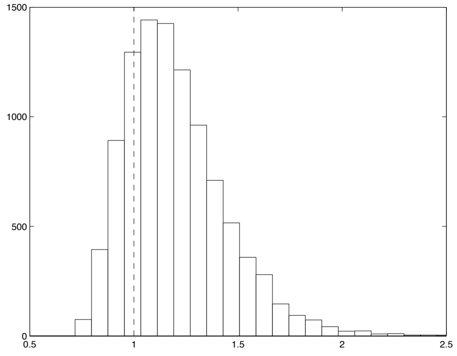

Figure 3: Ratio of the width of the credible interval based on a normal model to the width of the confidence interval based on a Laplace model, over the repeated draw of 10 independent Laplace observations, with weak prior knowledge of the variance. A model that correctly accounts for the fat-tailed nature of the data produces, on average, narrower credible intervals.

4.3 Optimal Point Estimates

When the x are drawn from a normal (or GN_2) distribution, and for any prior on s, the posterior of u can be show to be symmetric. This implies that the sample average $\overline{x}$ is the maximum-likelihood estimate of u, and also minimizes the expected value of any symmetric loss function, including MAD and RMS. Another estimate that is easily derived is the maximum likelihood when errors follow a double-exponential (or GN_1) distribution, which is the sample median.

The average of the observations, $\overline{x}$, is an unbiased estimator of, and given a large number of observations converges to, the distribution's mean u. However, it can be significantly suboptimal, especially with a limited number of observations as is often the case in practical problems of aggregating expert assessments. For a GN distribution other than the normal, the optimal estimate requires numerical integration. The MAD-optimal estimate minimizes $\int_u |\widehat{x} - u| \, df(u|x)$. From the first-order condition $-\int_{-\infty}^{\widehat{x}} df(u|x) + \int_{\widehat{x}}^{+\infty} df(u|x) = 0$, the optimal estimate $\widehat{x}$ satisfies $\int_{-\infty}^{\widehat{x}} df(u|x) = 0.5$, that is, it is the median of the posterior distribution of u. The RMS-optimal estimate minimizes $\int_u (\widehat{x} - u)^2 \, df(u|x)$. From the first-order condition $\int_u (\widehat{x} - u) \, df(u|x) = 0$, the optimal estimate is $\widehat{x} = \int_u u \, df(u|x)$, that is, the average of the posterior distribution of u. Table 8 summarizes the cases considered.

As to practical computation of the optimal estimate, while Monte Carlo integration is generally the preferred method for calculating posterior distributions, its advantages are less relevant for low-

Distribution of observation errors	Maximum likelihood	Minimum expected linear error (MAD)	Minimum expected quadratic error (RMS)
Normal (GN$_2$)	Sample average	Sample average	Sample average
Generalized normal (GN$_p$ with arbitrary p)	*Numerical maximization*	*Median of posterior, by numerical integration*	*Average of posterior, by numerical integration*
Double-exponential (GN$_1$)	Sample median	*Median of posterior, by numerical integration*	*Average of posterior, by numerical integration*

Table 8: Optimal estimates of the location parameter under different error models and cost functions.

dimensional problems (two dimensional in this case: the location and scale parameters). A gridding approach allows us to exploit the problem structure, computing only once for each grid point in u and in s the components of the posterior distribution that do not depend on the other parameter. Our implementation follows this gridding approach, which we found to be more computationally efficient.

4.4 The Average-Median Average Heuristic

From numerical experiments, we noticed that the optimal estimates for GN$_1$ and GN$_{1.5}$ models are often near the mid-point between the sample average and the sample median. Based on this, we propose the average-median average (AMA) heuristic,

$$\mathrm{AMA}\,x = \frac{1}{2}\left(\mathrm{Average}\,x + \mathrm{Median}\,x\right).$$

From Table 8, this is not an unreasonable heuristic. The average is the optimal estimate, in every sense, for shape parameter $p = 2$, that is for normally distributed errors. The median is optimal in the maximum likelihood sense for shape parameter $p = 1$, that is for errors that follow a double-exponential distribution. It is not then surprising that the mid-point between these two estimates would provide sensible estimates for intermediate values of the shape parameter.

4.5 Policy Benchmarks with Simulated Data

Table 9 reports policy benchmarks from simulation. We consider the cases where the samples are drawn from GN$_1$, from GN$_{1.5}$, and from normal distributions, and sample sizes (that is, number of

(a) GN_1 samples (fat tails)

Judges	Bayes GN_1	Bayes $\mathrm{GN}_{1.5}$	Average	Trimmed	Median	AMA
3	0%	3%	6%	—	4%	0%
5	0%	4%	12%	1%	5%	2%
10	0%	6%	20%	8%	1%	4%
20	0%	7%	26%	11%	2%	6%

(b) $\mathrm{GN}_{1.5}$ samples (intermediate tails)

Judges	Bayes GN_1	Bayes $\mathrm{GN}_{1.5}$	Average	Trimmed	Median	AMA
3	0%	0%	1%	—	11%	2%
5	1%	0%	2%	3%	12%	2%
10	3%	0%	3%	1%	8%	1%
20	5%	0%	4%	1%	11%	1%

(c) GN_2 samples (normal tails)

Judges	Bayes GN_1	Bayes $\mathrm{GN}_{1.5}$	Average	Trimmed	Median	AMA
3	2%	0%	0%	—	16%	4%
5	5%	1%	0%	6%	19%	5%
10	10%	2%	0%	3%	18%	5%
20	14%	3%	0%	3%	21%	6%

Table 9: Policy benchmarks from simulation with RMS cost: relative regret, or percentage loss relative to optimal policy.

judges) of 3, 5, 10, and 20. The numbers shown are the relative regret, or percentage loss relative to the optimal policy, with a quadratic cost (RMS, square root of the expected square deviation). We also computed these same cases with a linear cost (MAD, expected absolute deviation), which yielded very similar results (see Table A.1 in Online Appendix). The results were obtained by simulation over 10,000 trials for each case, resulting in mean standard errors of ±1% or less. With two of the policies computed by numerical integration, this required several hours of computation time. In addition to the policies already discussed, we include in our benchmarks the trimmed mean, a widely used heuristic, which we implement as the average after removing the lowest and highest data points (this is identical to the median when $n = 3$).

Assuming that the samples are normal and computing estimates with the sample average when they are drawn from a GN_1 distribution, leads to a regret of up to 26% in the cases tested. Conversely, assuming that the samples are drawn from a GN_1 distribution when they are drawn from a normal distribution results in regret up to 14% in the cases tested. The policy which is optimal for samples drawn from a $\mathrm{GN}_{1.5}$ distribution is robust to the shape of the distribution's tail, and performs well for both GN_1 and normal samples.

The average-median average heuristic (AMA) performs remarkably well across all cases. As expected, it performs well for fat-tailed data. They AMA heuristic also has a surprisingly good performance with normal data. Some insight into why this is the case can be had from the following. Consider X drawn from a normal distribution, a root-mean-square cost function, and a policy $g(\cdot)$ with relative regret r,

$$\varepsilon\left(g(X)\right) = \sqrt{\mathbf{E}_X \left(g(X) - \mu\right)^2} = (1 + r)\,\varepsilon\left(\overline{X}\right),$$

where

$$\varepsilon\left(\overline{X}\right) = \sqrt{\mathbf{E}_X \left(\overline{X} - \mu\right)^2} = \frac{\sigma}{\sqrt{n}}$$

is the loss with the average, μ and σ are the mean and standard deviation of X, and n is the sample size. The optimality of the average and orthogonality of errors implies that the norm of the difference between policies is $\sqrt{(1 + r)^2 - 1}\,\varepsilon\left(\overline{X}\right)$. The loss of the policy which consists of the mid-point between the sample average and policy $g(\cdot)$ is then, by the same argument,

$$\varepsilon\left(\frac{\overline{X} + g(X)}{2}\right) = \varepsilon\left(\overline{X} + \frac{g(X) - \overline{X}}{2}\right) = \sqrt{1 + \left(\frac{\sqrt{(1 + r)^2 - 1}}{2}\right)^2}\,\varepsilon\left(\overline{X}\right) = \frac{\sqrt{r^2 + 2r + 4}}{2}\,\varepsilon\left(\overline{X}\right).$$

The derivative of this expression with respect to r and evaluated at zero is 1/4. In Table 9(c) we see this relationship: the relative regret of the AMA is approximately one-fourth of that of the median.[8]

4.6 Empirical Policy Performance

We obtained empirical benchmarks by, where they are available, using the actual or realized values of the predicted quantities. We excluded the smaller data sets, and the size of some of the included data sets was reduced due to some realizations not being available since, for the economic data, we used final revised numbers which are available with a substantial lag. For the trivia questions we were only able to use the estimates of the number of member countries of the United Nations, with 192 as the reference value. For each column of data we ran 10,000 trials, and in each trial randomly sampled 10 estimates and then applied each policy. The results were averaged over all columns (predicted quantities) in each data set, based on which we compute a relative policy regret for each data set.

We found a large bias on all economic and earnings data sets, with over 90% of the forecasts to the same side of the eventual realization, which may be related to most of the forecasters being associated with sell-side firms since this was almost always towards the optimistic side. This large bias is the dominant source of forecasting error, reducing the performance difference between the policies. Without more detailed modeling work to account and control for bias and correlation across experts, which is outside of our scope here, these benchmarks are of limited value. When the average was used as the estimate, the average regret over all data sets was 2.1% in the economic forecasts and 11.4% in the earnings forecasts. With the median as the estimate, this was 0.3% and 0.6%, and with the AMA 0.7% and 4.5%.

The estimates of the number of member countries of the United Nations made by MBA students are less biased. Out of all data sets, UNHI has the lowest common bias relative to the standard deviation of the individual estimates. Here we obtain the empirical results that are most consistent with the results from simulation. The underperformance of the estimate based on the sample average is consistent with the previous results from simulation. For UNLO, the average regret is 7.7% with the average, 0.0% with the median, 3.1% with the AMA, and 1.4% with the theoretically

[8] The code file that performs this simulation is table_policies.m in the accompanying replication package.

optimal Bayesian estimate. For UNHI, the average regret is 9.2% with the average, 5.8% with the median, 0.1% with the AMA, and 0.0% with the Bayesian estimate.

UNLO					UNHI				
Judges	α_2	$\alpha_{1.3}$	Δ_2	$\Delta_{1.3}$	Judges	α_2	$\alpha_{1.3}$	Δ_2	$\Delta_{1.3}$
3	0.17	0.20	1.42	1.32	3	0.06	0.06	1.51	1.40
5	0.34	0.38	0.91	0.84	5	0.06	0.07	0.99	0.91
10	0.76	0.73	0.56	0.53	10	0.06	0.06	0.62	0.55
20	0.98	0.96	0.37	0.35	20	0.07	0.09	0.42	0.37

Table 10: Credible intervals with 3, 5, 10, and 20 judges. The α are the frequencies with which the true value is outside of the 95% interval. The Δ are the average widths of the intervals. Results are for a normal model and for a $\text{GN}_{1.3}$ model. For a larger number of experts the confidence interval becomes narrower and, due to unmodeled bias (or correlation across judges), excludes the true value with increasing frequency. Generally, the intervals based on the $\text{GN}_{1.3}$ model have smaller average width than the intervals based on the normal model, without higher α.

Table 10 illustrates the impact on the average width of the confidence interval of using a normal versus a $\text{GN}_{1.3}$ model. Note that for a larger number of experts the confidence interval becomes narrower and, in the data set with a larger unmodeled bias, excludes the true value with increasing frequency. As discussed in §4.2, the normal model leads to confidence intervals that are, on average, wider. The intervals based on the $\text{GN}_{1.3}$ model (that is, when correctly assuming that tails are fat) have, over the different cases considered, consistently smaller average width than the intervals based on the normal model, but this does not translate to a consistently higher proportion of cases where 192 falls outside of the confidence interval.[9]

5 Conclusion

Human judgment, as a complex process that does not meet the conditions for a central limit theorem to apply, cannot *a priori* be expected to follow a normal distribution. Further, when considering a collection of estimates or forecasts, other factors may contribute towards non-normality, including heterogeneity in skill and motivation across judges. Working with large data sets from multiple

[9] The code file that performs this simulation is empirical_benchmark.m in the accompanying replication package.

domains, we find overwhelming evidence that large prediction errors and large deviations from the consensus are far more frequent than would be expected under a normal distribution, which suggests moving away from the normality assumption in the aggregation of expert assessments and forecasts.

In our empirical analysis, the thickness of the tails of judgment shows a degree of consistency across different tasks, across different levels of expertise, and across different degrees of uncertainty about the quantity in question. While, in any given aggregation problem, the thickness of tails cannot be estimated from the small samples that are typically available, this consistency supports using prior assumptions about the shape of the distribution. A generalized normal model with shape parameter $p = 1.3$ provides a good fit for the judgmental estimation and forecasting data examined. For cases where heterogeneity across judges is expected to be low, or can be controlled for, we suggest $p = 1.5$. In the converse case, when heterogeneity is expected to be high and cannot be controlled for, we suggest $p = 1.1$. For less statistically sophisticated users, the average-median average heuristic (AMA) is a simple but adequate and robust alternative.

Further work is needed regarding correlation across judges in this context. Both the empirical distributional characteristics of common bias and the implications of fat tails for the aggregation of correlated estimates and forecasts are questions of theoretical interest and practical significance.

References

Armstrong, J. Scott. 2001. Combining forecasts. J. Scott. Armstrong, ed., *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers, 417–439.

Clemen, Robert T. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* **5**(4) 559–583.

Clemen, Robert T., Robert L. Winkler. 1993. Aggregating point estimates: A flexible modeling approach. *Management Science* **39**(4) 501–515.

Fang, Lily H, Ayako Yasuda. 2014. Are stars' opinions worth more? the relation between analyst reputation and recommendation values. *Journal of Financial Services Research* **46** 235–269.

Gelman, Andrew. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**(3) 515–33.

Huber, Peter J., Elvezio M. Ronchetti. 2009. *Robust Statistics*. 2nd ed. John Wiley & Sons, Inc.

Kahneman, Daniel. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Kotz, Samuel, Tomasz Kozubowski, Krzystof Podgórski. 2001. *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance (Progress in Mathematics)*. Birkhauser Verlag.

Larrick, Richard P., Jack B. Soll. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Mamagement Science* **52**(1) 111–127.

Lawrence, Michael, Paul Goodwin, Marcus O'Connor, Dilek Önkal. 2006. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting* **22**(3) 493–18.

Lye, Jenny N., Vance L. Martin. 1993. Robust estimation, nonnormalities, and generalized exponential distributions. *Journal of the American Statistical Association* **88**(421) 261–267.

Makridakis, Spyros, Robert L. Winkler. 1983. Averages of forecasts: Some empirical results. *Management Science* **29**(9) 987–996.

Nadarajah, Saralees. 2005. A generalized normal distribution. *Journal of Applied Statistics* **32**(7) 685–694.

Nelson, Daniel B. 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59**(2) 347–370.

Palley, Asa B, Ville A Satopää. 2023. Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. *Management Science* **69**(9) 5128–5146.

Palley, Asa B., Jack B. Soll. 2019. Extracting the wisdom of crowds when information is shared. *Management Science* **65**(5) 2291–2309.

Rader, Christina A, Jack B Soll, Richard P Larrick. 2015. Pushing away from representative advice: Advice taking, anchoring, and adjustment. *Organizational Behavior and Human Decision Processes* **130** 26–43.

Theodossiou, Panayiotis. 1998. Financial data and the skewed generalized t distribution. *Management Science* **44**(12) 1650–1661.

Winkler, Robert L., Robert T. Clemen. 1992. Sensitivity of weights in combining forecasts. *Operations Research* **40**(3) 609–614.

Winkler, Robert L., Spyros Makridakis. 1983. The combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)* **146**(2) 150–157.

Yaniv, Ilan. 1997. Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes* **69** 237–249.