Human Judgment is Heavy Tailed: Empirical Evidence and Implications for the Aggregation of Estimates and Forecasts

Miguel Sousa Lobo* Dai Yao†

July 8, 2010

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 errors follow a jointly normal distribution. We investigate this question empirically using 17 datasets that include over 20,000 estimates and forecasts. We find incontrovertible evidence for excess kurtosis, that is, of fat tails. Despite the diversity of the analyzed datasets as regards to the degree of uncertainty about the quantity being assessed and to the level of expertise and sophistication of those making the assessments, we find consistency in the frequency with which an expert is in large disagreement with the consensus. Fitting a generalized normal distribution to the data, we find values for the shape parameter ranging from 1 to 1.6 (where 1 is the double-exponential distribution, and 2 the normal distribution). This has important implications, in particular for the aggregation of expert estimates and forecasts. We describe optimal Bayesian aggregation with heavy tails, and propose a simple average-median average heuristic that performs well for the range of empirically observed distributions.

1 Introduction

Laplace first described in the second half of the 18^{th} century the 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

*miguel.lobo@insead.edu

†dai.yao@insead.edu

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. On the one hand, the central limit theorem provides a compelling explanation for the empirical finding that the normal distribution naturally arises in a number of situations, when the observed uncertainty is the sum of a large number of independent sources of randomness. On the other hand, 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 (and, more generally, its multivariate form is preserved under linear transformation), the simple least-squares procedure solves for the maximum likelihood estimate, *etc.* However, in any given application the distribution of errors, be they measurement, assessment, 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 follows 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. Even if plausible, this is a somewhat contrived model, and surely requires testing, especially given the centrality of estimates and forecasts of sales, costs, and such as inputs into managerial decision-making.

While normality is often a 'good-enough' approximation, it seems fair to say that normally distributed data is the exception rather than the rule, as the assumptions of additive and weakly dependent random terms from which the central limit theorem derives seldom hold. A testament to the fact that tails are often heavier than in the normal case, that is that excess kurtosis is frequently 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. While Clemen and Winkler (1993) noted that "careful assessment of the distributions' tails and careful choice of models may be critical [...] in aggregation models," little to no work exists regarding the shape of the tails of human judgment, and the issue is not mentioned in either of the two widely cited reviews by Clemen (1989) and by Armstrong (2001), nor in the more recent by Lawrence et al. (2006) (see also Bunn (1989) and the associated *Special Issue on Combining Forecasts*).

Estimating the weight of the tails of a distribution and obtaining information about its fourth moment requires a large amount of data. As we will detail, a few hundred data 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 obtain such prior information from the behavior in comparable tasks of similarly-trained experts. The distributional characteristics can then be estimated from data pooled together from different sources. If the shape of the distribution is consistent across different data sources pertaining to similar tasks, and provides a good fit to the data, this should be used in lieu of a normal prior.

We investigate the shape of the distribution of deviations from consensus among judges, working with a generalized normal distribution with three parameters, for location, scale, and shape. It includes the normal and Laplace distributions as special cases. We restrict our investigation to estimates or forecasts of real-valued quantities, and for quantities that are restricted to be positive work with their logarithm. To increase the validity of our findings, we use multiple datasets with estimates and forecasts from different sources, involving different levels of uncertainty about the quantity assessed, and judges with different degrees of expertise. If, as our findings strongly suggest, normality does not hold, the sample mean is not the optimal estimate of the distribution mean. Further, the confidence interval based on, say, a Student's t posterior distribution does not correctly reflect the information contained in the sample, and it may either substantially over- or underestimate the uncertainty about the mean. We find that, somewhat counter-intuitively, using a normal model with a weak prior on the variance will tend to overestimate posterior uncertainty about the mean of heavy-tailed data (this may not be the case, however, when stronger prior knowledge of the variance is available).

We discuss alternative procedures for aggregating estimates when the distribution of deviations from the consensus is heavy tailed, including obtaining the appropriate posterior distributions and confidence intervals. We provide benchmarks for the performance of different policies, based both on simulation and on the empirical data for which the realized values are available. We find the benchmarks based on most of the economic forecasts to be of little use: due to substantial correlation across judges, the common bias becomes the dominant source of error in these datasets (adequately modeling and controlling for such correlation with heavy-tailed data will require further investigation beyond our scope here). However, where the common bias is moderate, the empirical benchmarks support the results from simulation.

Simple heuristic rules, such as the simple average (Makridakis and Winkler, 1983; Larrick and Soll, 2006), weighted averages (Winkler and Makridakis, 1983; Winkler and Clemen, 1992), and trimmed means (Yaniv, 1997), have been considered as practical alternatives to Bayesian models.

In addition to the Bayesian estimate, we also assess the performance of some of these heuristics with heavy-tailed data and of a new one we propose, the average-median average heuristic, that has robust performance for such data 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 datasets used, from a panel of economists and from MBA student surveys, totaling about twenty thousand estimates and forecasts. We include some quantile-quantile plots for a visual assessment of fit and, for each dataset, we use information criteria to determine the model structure with best fit. We then present estimates for the shape parameter of the generalized normal distribution based on these models. This includes a discussion of alternative explanations for the heavy-tailed nature of the data, including that heterogeneity across judges results in a mixture of normals, and of why we believe our analysis rules them out. §4 considers the problem of combining estimates when the distribution of errors is heavy tailed. It describes the optimal policy, introduces the average-median average heuristic, and reports benchmarks of the different policies based on the simulation of different distributional assumptions (normal, intermediate, and heavy tails) as well as based on the empirical datasets, under both a linear and a quadratic cost function. §5 has some brief concluding remarks, including recommendations for priors for the shape parameter when dealing with related judgmental data.

All data, code, and associated documentation for the results and methods described here are available at `http://`. This includes Matlab functions for generating GN-distributed pseudo-random numbers, for estimating the shape parameter of a GN distribution, and for computing point estimates and confidence intervals for the location parameter of a GN distribution given a prior value of the shape 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 and s have same units as X. The shape parameter p dictates the 'thickness' of the tails.[1] From the first four moments of the GN distribution (Nadarajah, 2005), we have

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

We sometimes denote a specific subset of distributions by indexing with the shape parameter, so that $\text{GN}_{p=2}$ is the normal, or Gaussian distribution (with $\sigma = s/\sqrt{2}$), and $\text{GN}_{p=1}$ is the Laplace, or double-exponential distribution. We obtain a uniform distribution 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. In fact, if the tails of the distribution of observation errors are any 'heavier' than those of a Laplace 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, should this be supported empirically it has substantial benefits in computational tractability and stability of estimates.

2.2 On Sample Size

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

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

where $f(u, s, p)$ is the joint prior on the parameters. From this we can see that obtaining an accurate estimate of the shape parameter, that is of 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, and models with a larger scale parameter and thinner tails. From the Fisher information matrix, the expected half width of the 95% confidence interval for p as a function of sample size is given in Table 1.[2] For $\text{GN}_{p=1}$ about 20

[1] 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, however, 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.

[2] 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.

samples are needed to be able to discriminate from a normal distribution and for $\mathrm{GN}_{p=1.5}$ about 200 samples are needed, although, in both cases, a wide confidence interval will still result. Obtaining a reasonably narrow confidence interval for p (for, say, accuracy to one decimal) requires on the order of a thousand observations or more.

As an example, we drew $x_1, x_2, \ldots, x_{50}$ independently from a zero-mean and unit-variance normal distribution. Figure 1 represents the likelihood function (the posterior distribution of the parameters with an improper uniform prior). Note the difficulty in estimating the shape parameter p, and its confounding with the scale parameter s. A Bayesian 95% confidence (or credible) interval for p, in this instance, is $[1.3, 3.9]$.

In the common problem of aggregating, say, ten or fewer expert forecasts, a sample of such size yields little information about the shape of the tails of the distribution of forecasting errors. With the GN model, for instance, even with n in the order of 50 to 100, it is not possible to meaningfully estimate p. We need to rely on prior knowledge of p, for which we need an understanding of human judgement. If the distributional characteristics of human judgment are predictable over similar tasks, prior knowledge of p can be incorporated into the aggregation of estimates and forecasts, or other statistical procedures.

2.3 Parameter Estimation and Data Models

A simple estimation procedure consists of finding the shape parameter p such that the kurtosis of the generalized normal distribution matches that of the data. An approximation is quickly computed, by numerical inversion of the formula for the kurtosis of the GN distribution given in §2.1. This approach has, however, a number of limitations and sources of bias. Of greatest concern is the following. As described in §3.1, we organize the data in matrix format. Each dataset contains estimates or forecasts of different uncertain quantities, and the multiple estimates of each of these quantities (obtained from different judges) are grouped in a column. For some of the datasets there are likewise multiple forecasts by each judge, which we group in a row. By the nature of the data, then, different columns (and sometimes different rows as well) may have different means and variances, so that the kurtosis-matching procedure can only be implemented 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

n	$\text{GN}_{p=1}$	$\text{GN}_{p=1.5}$	$\text{GN}_{p=2}$
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 95% confidence interval for shape parameter p under different distributions and sample sizes.

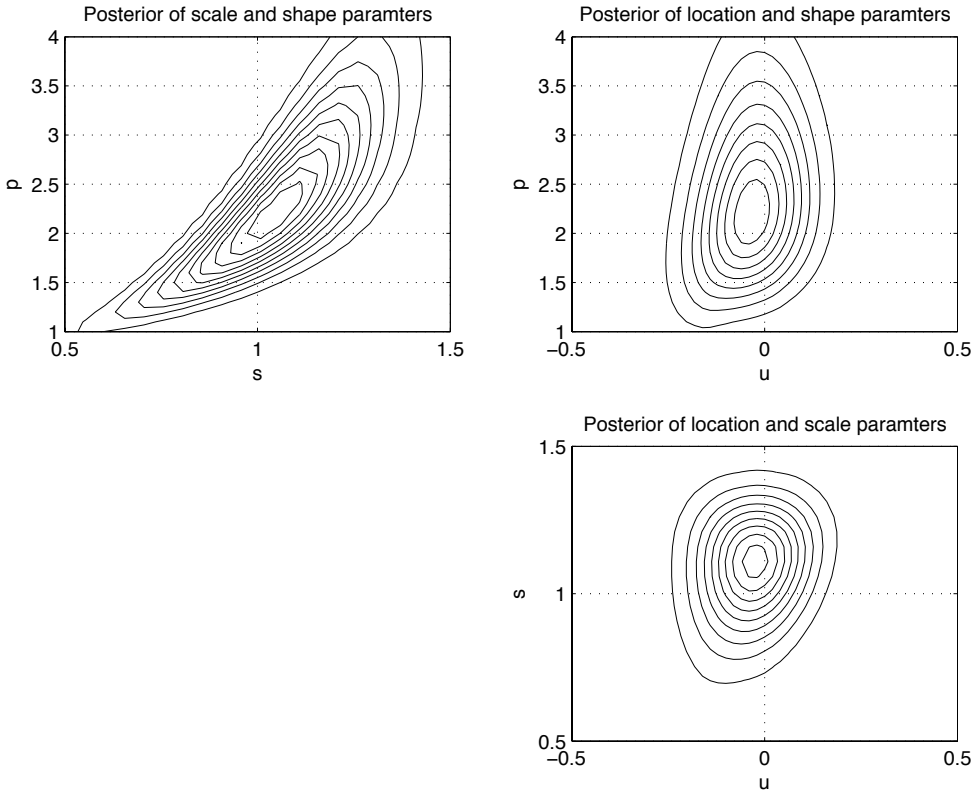

Figure 1: Two-way marginals of the joint posterior distribution of the location (u), scale (s) and shape (p) parameters given 50 normally distributed random samples.

possible to construct an estimate that controls for this effect, the approach then loses the appeal of simplicity – and this is all the more difficult to do when there is heterogeneity across both rows and columns. Nevertheless, and for reference, we include in our results estimates based on matching the kurtosis. To mitigate bias when computing these estimates we discard columns with 10 or fewer observations.

We attach more importance to Bayesian estimates based on 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 $f(u, s, p)$, the posterior distribution of the problem parameters given the observed data is $f(u, s, p|x) = f(x|u, s, p)f(u, s, p)/f(x) \propto f(x|u, s, p)f(u, s, p)$, which we evaluate by Markov-chain Monte Carlo simulation. From this we obtain the posterior marginal of the shape parameter $f(p|x)$, and report its mean $p_{\text{Bayes}} = \mathbf{E}_{p|x}p = \int p\, df(p|x)$, and a 95% confidence (or credible) interval $[p_{\text{lower}}, p_{\text{upper}}]$ such that $F(p_{\text{lower}}|x) = 1-F(p_{\text{upper}}|x) = 0.025$, where $F(q|x) = \int_{-\infty}^{q} df(p|x)$.

3 Empirical Tests

3.1 Data

For increased validity of our empirical analysis, we use multiple sources of data. The estimates and forecasts used range widely in the level of expertise of those providing them, as well as in the amount of uncertainty about the quantity of interest. We analyze a total of 20,033 estimates and forecasts in 17 datasets, about 10 types of quantities and from 2 different sources.

1. *Economists.* We collected from the Wall Street Journal's web site economic forecasts for the period of 2003 to 2009. The 89 economists that were present in the panel at some point over this time are associated with a variety of institutions. Most of these are financial-sector firms, but some panel members have different profiles such as economic consultants, in-house economists at large corporations, and academics. Most of the estimates are on a quarterly basis, but some annual estimates are also included. We construct two datasets for each forecasted quantity. The first retains the final estimate from each economist in the panel prior to the realization of the uncertainty, while the second retains forecasts made two years in advance (exception to this are non-farm payroll and the personal consumption expenditures price index, where early forecasts are not available; the datasets labeled with an appended 'a' are the advance forecasts). This leads to a total of 12 datasets. In each dataset, we collect in each row the estimates from each economist, and in each column the estimates for each

forecasted quantity, with an indication for missing data when an economist did not provide a particular estimate. A total of 8,348 forecasts are used, for the following quantities, all relating to the economy of the United States of America.

(a) *Gross domestic product (GDP and GDPa).*

(b) *Non-farm payroll (NFARM).*

(c) *Unemployment rate (UNEMP and UNEMPa).*

(d) *Consumer price index (CPI and CPIa).*

(e) *Rate on 10-year notes (R10Y and R10Ya).*

(f) *Personal consumption expenditures price index (PCEPI).*

(g) *Change in home price (CIHP and CIHPa).*

2. *Business School Students.* INSEAD MBA students are given judgmental exercises in the class Uncertainty, Data and Judgment, which is part of the core curriculum. These exercises are part of a larger survey designed to illustrated cognitive and behavioral biases. From each survey, we made use of an average of three questions that required students to estimate or forecast a quantity. The surveys were administered between 2000 and 2009 by six different faculty. We collect in each column of data all responses from students who were taught by the same instructor and completed the survey in the same week. A total of 11,685 estimates are included in five datasets as follows.

   (a) *Number of countries in the United Nations (UNLO and UNHI).* Students were asked to estimate the number of member countries of the United Nations. Half of the students were given a low anchor by including in the survey question the statement: 'To make your estimate, I would suggest that you start with a value of 95.' The other half were given a high anchor of 300 countries in similar fashion. We separate the responses in low- and high-anchor condition into two datasets.

   (b) *Foreign exchange rate (FXLO and FXHI).* Students were asked to forecast an exchange rate (typically USD-EUR) at the end of the year. Half of the students were given a low anchor in the preceding question by being asked whether they thought the exchange rate would be above or below a stated number which was below the exchange rate at the time. The other half were given a high anchor in similar fashion. Again we separate the responses in low- and high-anchor condition into two datasets.

(c) *High-uncertainty items (SCALE).* Another question included in the surveys asked students to estimate large numbers 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 dataset.

In all datasets, when the quantity of interest is restricted to be positive, we work with the logarithm of the data. As discussed below, this mitigates issues of skewness in the data.

3.2 Data-Generation Model

The observations are organized in rows i and columns j. Each row i corresponds to a judge. In the datasets from the panel of economists each i means the same person over all columns. In the other datasets, since only one estimate is available from each respondent, the row index has no significance. Each column j corresponds to a different quantity being estimated or forecasted. In some cases the quantity being estimated is the same for different j, such as for the estimates of the number of member countries of the United Nations, but estimates made at different times and in different contexts are grouped under different columns (when the same question is included in surveys administered by different faculty or at different times).

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, and for each dataset test for fit against the observations, different model structures for the location, scale, and shape parameters. 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 and for each row, 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 likewise consider the following.

- 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 inherent difficulty in estimating or forecasting a particular quantity. The r_i model the uncertainty due to the limitations in information or cognition specific to a particular judge. When we model both column- and row-specific scale parameters, s_j and t_i, there appears to be no obvious default rule 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 $s_{ij} = \max(s_j, t_i)$.[3] Based on an information criterion, we found the geometric average to provide the best fit (in the panel of economists datasets where we model 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, $p_{ij} = p_j$.

Not all 27 combinations of the three sets of three alternatives above make sense. We only test row-specific parameters for the datasets where the identity of judge i is the same across all columns (those from the panel of economists). Further, we only test column-specific scale parameters if also including column-specific location parameters, and we only test column-specific shape parameters if also including both column-specific location parameter and column-specific shape 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

[3]Consider that each row is a different judge, and that each column a different quantity to be estimated. These four rules correspond to different assumptions about the way in which the accuracy of the judge (due to knowledgeability, skill, 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 on the standard deviations and on the variances associated with the judge and with the uncertainty. The maximum rule corresponds to the assumption that the ability to obtain an accurate estimate is limited by the highest of the variances associated with either the judge or the uncertainty.

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. For instance, the data can never rule out that, for a particular row, $s_i = 0$, with the location parameters such that $u_j = x_{ij}$. An improper prior term in $1/s_i$ would then lead to an improper posterior (not integrable at $s_i = 0$). 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 $\log \sigma_{\log s}$ (see also Gelman (2006)). Finally, we set a uniform prior on the shape parameters $p \sim \text{Uniform}(0.5, 4)$. The choice of bounds is somewhat arbitrary, as long as it is wide enough to not impact the estimates, but excluding improper posteriors near zero and for large p.

3.3 Model Fit and Selection

Several of the datasets we used had a couple of obvious outliers which, most often, appeared to be due to transcription errors. Some of these were manually removed, but we also applied a general rule removing any point over five standard deviations away from the mean (using the sample mean and sample standard deviation for each column of data). To the extent that this rule may lead to false positives in the detection of outliers in heavy-tailed data, our estimates of the thickness of the tails may be conservative.

Figures 2 and 3 show quantile-quantile plots for two sample datasets (one-quarter-ahead forecasts of USA GDP from the panel of economists, and foreign exchange rate forecasts with low-anchor condition from the MBA surveys). The plots to the left are against the quantiles of the normal distribution. The heavy-tailed nature of the data is apparent. The plots to the right are against the quantiles of a generalized normal distribution (with the shape parameter p here chosen by visual fit). These two examples illustrate one of the cases of better apparent fit, and an example of the cases with less-good fit. Overall, the quantile-quantile plots over all datasets suggest an adequate 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,

$$AIC = -2 \max_{\theta} \log f(x|\theta) + 2k,$$

and the Bayesian information criterion,

$$BIC = -2 \max_{\theta} \log f(x|\theta) + k \log N,$$

where N is the number of data and k is the number of (non-zero) parameters in the model. We put more emphasis on the BIC, which is more conservative in that it puts a stronger penalty on

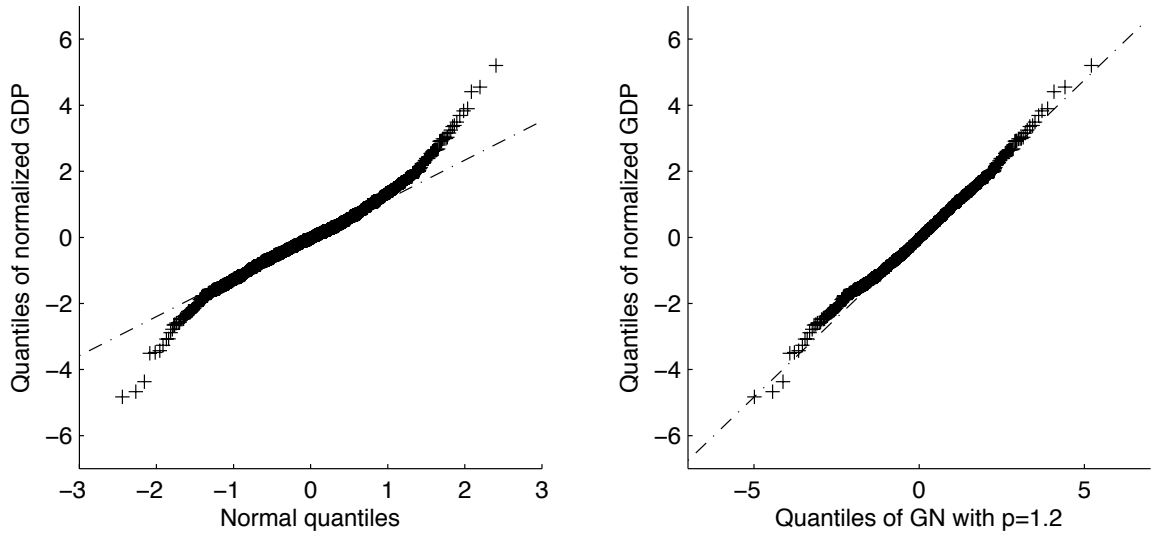

Figure 2: Quantile-quantile plots of USA gross domestic product growth forecasts (GDP, after each quarter's forecasts were normalized to have zero mean and unit variance), against a normal distribution and against a generalized normal distribution with shape parameter $p = 1.2$.

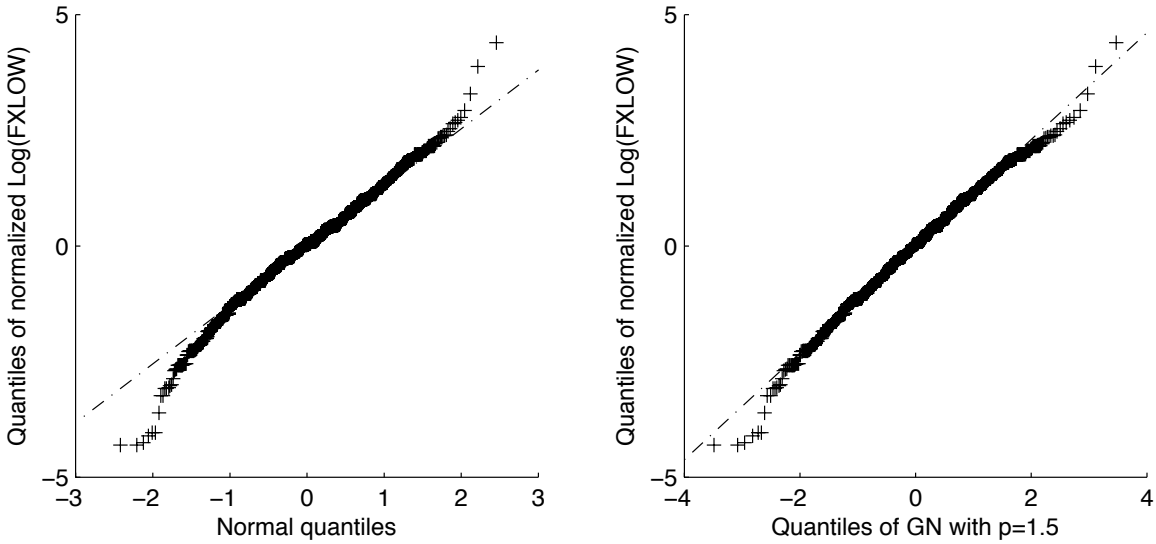

Figure 3: Quantile-quantile plots of logarithm of foreign currency exchange rate forecasts in low-anchor condition (FXLOW, after the forecasts from each MBA class were normalized to have zero mean and unit variance), against a normal distribution and against a generalized normal distribution with shape parameter $p = 1.5$.

the number of parameters. As we are interested in generalizable results rather than in explaining a particular dataset, it is preferable to err on the side of model parsimony in order to avoid overfitting. The maximum-likelihood estimates are obtained by numerical optimization using the sequential quadratic programming implementation in the Matlab Optimization Toolbox, 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, it sometimes converges to a sub-optimal local optimum).

Tables 2 to 6 present detailed results for some sample datasets. The model fit results are consistent with our understanding of the data. The foreign exchange forecasts were collected at different times (and in some cases about different currencies), so that different columns of data generally have significantly different means. Since we work with the logarithm of the estimates, the scale parameter is a measure of the coefficient of variation of the original data, the ratio of the standard deviation to the mean, which explains the limited evidence for different scale parameters for each column of data, that is, for each MBA class. The BIC supports different scale parameters for each column in the low-anchor condition dataset (FXLO), but not so in the high-anchor condition (FXHI). It does not support different shape parameters for each column. The dataset of disparate 'scale uncertainty' items (SCALE) follows a similar pattern. We chose to model all three datasets with separate location and scale parameter for each column and with a single shape parameter.

While the AIC generally leads to similar models with these four datasets, we chose to present the detailed results for the foreign exchange forecasts with low-anchor condition (FXLO) because, out of all the datasets, this is the only instance where there is some support for heterogeneity in the shape parameter. Given the weight of evidence over all other datasets and the fact that even in this case the evidence for it is limited, we do not estimate models allowing for heterogeneity in the shape parameter.

Like the foreign exchange forecasts, the estimates of the number of member countries of the United Nations were also collected over a period of ten years. However, the estimated quantity didn't meaningfully change in that period and the characteristics of the MBA student body are stable over time. Both the BIC and AIC support single location, scale and shape parameters common to all columns, which is the model we used for both datasets (low-anchor and high-anchor condition, UNLO and UNHI).

In the datasets with the forecasts from the panel of economists, we consider additional models that allow for heterogeneity across judges. Tables 5 and 6 give detailed results for the one-quarter-ahead forecasts of GDP and CPI, which are representative of this group of datasets. In different datasets, the BIC always supports different location parameters for each quarter of data, and

Model Parameters	N	k	$\log f(x\|\theta^*)$	AIC	BIC
$\theta = (u, s, p)$	2025	3	524	-1043	-1026
$\theta = (u_1, \ldots, u_n, s, p)$	2025	20	1426	-2813	-2700
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p)$	2025	37	1558	-3041	$-2833^\dagger$
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p_1, \ldots, p_n)$	2025	54	1615	$-3122^\ddagger$	-2819

Table 2: Model selection for foreign currency exchange rate forecasts in low-anchor condition (FXLO). The model with best fit according to the Bayesian information criterion (†) has a different location and scale parameter for each MBA class, and a common shape parameter.

Model Parameters	N	k	$\log f(x\|\theta^*)$	AIC	BIC
$\theta = (u, s, p)$	2003	3	860	-1713	-1696
$\theta = (u_1, \ldots, u_n, s, p)$	2003	20	2062	-4084	$-3972^\dagger$
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p)$	2003	37	2099	$-4124^\ddagger$	-3917
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p_1, \ldots, p_n)$	2003	54	2112	-4117	-3814

Table 3: Model selection for foreign currency exchange rate forecasts in high-anchor condition (FXHI). The model with best fit according to the Bayesian information criterion (†) has a different location parameter for each MBA class, and common scale and shape parameters.

Model Parameters	N	k	$\log f(x\|\theta^*)$	AIC	BIC
$\theta = (u, s, p)$	2022	3	406	$-805^\ddagger$	$-788^\dagger$
$\theta = (u_1, \ldots, u_n, s, p)$	2022	20	417	-795	-682
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p)$	2022	37	436	-798	-590
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p_1, \ldots, p_n)$	2022	54	447	-787	-484

Table 4: Model selection for estimates of number of countries in the United Nations in low-anchor condition (UNLO). The model with best fit according to the Bayesian information criterion (†) has common location, scale and shape parameters for all data.

Model Parameters	N	k	$\log f(x\|\theta^*)$	AIC	BIC
$\theta = (u, s, p)$	1930	3	-2837	5680	5697
$\theta = (u_1, \ldots, u_n, s, p)$	1930	38	-582	1241	$1452^{\dagger}$
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p)$	1930	73	-491	1128	1535
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p_1, \ldots, p_n)$	1930	108	-454	1124	1725
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s, p)$	1930	122	-378	1000	1679
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, p)$	1930	157	-286	887	1761
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, p)$	1930	241	-140	$762^{\ddagger}$	2103
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, p_1, \ldots, p_n)$	1930	276	-107	767	2303
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, q_1, \ldots, q_m)$	1930	324	-100	848	2651

Table 5: Model selection for USA's gross domestic product forecasts (GDP). The model with best fit according to the Bayesian information criterion (†) has a different location parameter for each quarter, and common scale and shape parameters. The model with best fit according to the Akaike information criterion (‡) has different location and scale parameters for each quarter and for each economist, and a common shape parameter.

Model Parameters	N	k	$\log f(x\|\theta^*)$	AIC	BIC
$\theta = (u, s, p)$	960	3	-793	1591	1606
$\theta = (u_1, \ldots, u_n, s, p)$	960	20	-56	153	250
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p)$	960	37	60	-46	$134^{\dagger}$
$\theta = (u_1, \ldots, u_n, s_1, \ldots, s_n, p_1, \ldots, p_n)$	960	54	80	-51	212
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s, p)$	960	100	20	160	647
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, p)$	960	117	223	$-211^{\ddagger}$	358
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, p)$	960	197	280	-166	793
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, p_1, \ldots, p_n)$	960	214	291	-154	887
$\theta = (u_1, \ldots, u_n, v_1, \ldots, v_m, s_1, \ldots, s_n, t_1, \ldots, t_m, q_1, \ldots, q_m)$	960	276	335	-118	1225

Table 6: Model selection for USA's consumer price index forecasts (CPI). The model with best fit according to the Bayesian information criterion (†) has a different location and scale parameter for each quarter, and a common shape parameter. The model with best fit according to the Akaike information criterion (‡) has different location and scale parameters for each quarter, a different location parameter (or bias) for each economist, and a common shape parameter.

sometimes different scale parameters. It does not support heterogeneity across forecasters. The AIC does support heterogeneity across economists, either in the location parameter only, or in both the location and scale parameters. Both the BIC and the AIC always support a single shape parameter. In light of this, when feasible, we provide two estimates for each of the panel of economists datasets. The first is based on a model with different location and scale parameters for each column, but common to all rows. The second is based on a model that allows for different location and scale parameters both for each column and for each row. As noted in the following discussion, this also allows us to investigate whether the tail thickness of the empirical distribution can be explained as a mixture of normal distributions with different variances, arising from the heterogeneity of judges. However, controlling for both column and row effects requires that there be a sufficiently large number of judges for which we have enough data, which was only the case for the larger datasets (GDP, GDPa, and NFARM).

3.4 Estimates of Shape Parameter and Discussion

Summary statistics for the data and estimates and confidence intervals for the shape parameter p are reported in Tables 7 and 8, 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 11 estimates. For the models that control for both column and row effects, we iteratively deleted columns and rows until each column and row has at least 8 estimates. The reported skewness and kurtosis are after normalizing each column of data for 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 Bayesian confidence interval for p is from the 2.5% and 97.5% quantiles of this distribution.

The posterior distributions are obtained from our Matlab 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 that 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 (which we require to be below .25 with a 10-sample lag, although in most cases it is effectively zero).

We verified our implementation by generating pseudo-random observations from a normal distribution to create datasets of size similar to our empirical datasets. Running the estimation

procedure on these datasets, we did not detect bias, and the confidence intervals included $p = 2$ with a frequency consistent with the confidence level.

As is true for concerns regarding correlation across judges and common bias, the issue of modeling skewness is somewhat problem specific. However, we find that in all the datasets we used the only major source of skewness is, for quantities that are restricted to be positive, when the coefficient of variation is large, that is, when the uncertainty is large relative to the quantity being assessed. By working with the logarithm of the estimates in those cases, skewness is not a substantial concern in any of our datasets. In the extreme case, for the dataset of MBA survey items with scale uncertainty (SCALE), the skewness of the original data is 60.1, while the skewness of their logarithm is 0.42.

We find overwhelming evidence that human judgment is heavy tailed. We also find remarkable consistency across different tasks, different degrees of uncertainty about the quantity being assessed, and different levels of expertise on the part of the judges. Over all 15 datasets, the minimum estimate is 0.95, and the maximum 1.69. The 10^{th} and 90^{th} percentiles are 1.14 and 1.61, and the median 1.35. The estimates based on the kurtosis and on the Bayesian posterior are also broadly consistent. The average of the estimates over all datasets is 1.36 based on the kurtosis, and 1.37 based on the Bayesian posterior. We don't find evidence that the distributions of judgments from MBA students and from professional economists have fundamentally different shapes. For the MBA datasets the average estimate is 1.43, and for the panel of economists datasets the average is 1.34. In all but the smallest dataset (CIHPa), 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 heavy 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. We consider, however, two alternative explanations for the heavy-tailed nature of the data, heterogeneity and incentives. While neither explanation changes the implications for how the aggregation of estimates should be performed, they need to be addressed to make valid inferences about human judgment in regards to the frequency of large deviations from the consensus.

Heavy-tailed distributions in different domains, from finance to biology to engineering, are frequently modeled as a mixture of normal distributions with different variances. Judges are likely to be heterogeneous in their propensity to deviate from the consensus. This may be due to some judges being better informed than others, to different cognitive abilities, or simply to heterogeneity in disposition. Note that heterogeneity across judges can go either way regarding the shape of the resulting mixture. If judges are more heterogeneous in their variances than in their means,

Dataset	N	n	Skew.	Kurt.	p_{kurt}	k	p_{Bayes}	p_{low}	p_{high}
GDP	1933	36	-0.03	5.3	1.10	75	1.15	1.05	1.25
GDPa	1292	24	-0.36	4.8	1.18	51	1.35	1.22	1.50
NFARM	1733	34	-0.10	4.1	1.38	71	1.45	1.32	1.59
CPI	965	18	-0.35	4.4	1.29	39	1.34	1.17	1.51
CPIa	593	11	-0.20	4.2	1.33	25	1.31	1.10	1.55
R10Y	807	15	-0.26	3.9	1.44	33	1.38	1.18	1.59
R10Ya	485	9	-0.63	3.7	1.51	21	1.47	1.20	1.78
PCEPI	208	4	0.45	5.5	1.06	11	1.14	0.85	1.47
CIHP	187	4	-0.35	3.9	1.44	11	1.33	0.93	1.82
CIHPa	145	3	-0.19	3.3	1.77	9	1.54	0.95	2.39
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 7: Estimates of shape parameter for data from panel of economists and from MBA students.

				Homogeneous judges				Heterogeneous judges			
Dataset	N	m	n	k	p_{Bayes}	p_{low}	p_{high}	k	p_{Bayes}	p_{low}	p_{high}
GDP	1894	74	36	75	1.20	1.09	1.32	225	1.54	1.38	1.72
GDPa	1218	65	24	51	1.37	1.22	1.53	183	1.70	1.46	1.99
NFARM	1703	63	34	71	1.44	1.31	1.58	199	1.31	1.15	1.48

Table 8: Estimates of shape parameter for data from panel of economists, controlling for heterogeneity across analysts.

the mixture will have heavier tails. If judges are more heterogeneous in their mean than in their variances, the mixture will have thinner tails. In the larger datasets from the panel 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 variance from the consensus). The pattern of results is not consistent with the mixture of normals hypothesis. After controlling for heterogeneity, the shape parameter is larger (closer to normal) in two of the datasets, but smaller in one. Without controlling for heterogeneity, the average estimate of the shape parameter is 1.34, while with independent location and scale parameters for each judge the average shape parameter is 1.52. That is, we still find that the distribution of each judge's estimates has significantly heavier tails than a normal distribution.

The second alternative explanation regards 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-value terms, the prestige from being the one of the few who 'got it right' when everyone else was very wrong is greater than the penalty of being very wrong (Fang and Yasuda (2009) argue for evidence of such effects in the context of financial analysts). However, this argument is inconsistent with our results from the MBA surveys where, since they were completed anonymously, incentives should not play a role. It is also conceivable that professional incentives, to the extent that they play a role in the panel of economists data, are less prominent for longer-term forecasts. We only find a small different in the average of the estimates of the shape parameter in the one-quarter-ahead forecasts (GDP, CPI, R10Y, CIHP) and in the two-years-ahead forecasts (GDPa, CPIa, R10Ya, CIHPa), 1.30 versus 1.42.

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 given (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, as well as assess alternative policies and a proposed heuristic.

4 Combining Estimates and Forecasts with Heavy-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, the prior beliefs and the observations of the available individual judgments imply a posterior distribution of beliefs for the location parameter. An estimate is then optimal

in the sense that it minimizes the expected cost or 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 a wider range of problems, such as, for instance, costs of over- or under-stocking due to inaccurate demand forecasts. 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 is seldom 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, which is then fed into a broader risk model (likely after including 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 Bayesian confidence interval can be computed from the posterior distribution to summarize the uncertainty about the location parameter given the observations at hand. Note that in the normal case with non-informative priors this can be done directly from Student's t distribution.

4.2 On Confidence Intervals

Confidence 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 heavy tailed can lead to confidence intervals that are either significantly wider or significantly narrower than appropriate. With a normal model, the confidence interval depends only on the sample mean and variance, and is not impacted by the higher moments of the sample. Relative to this interval, with, say, a Laplace model, the confidence interval will tend to be narrower when outliers are present, and wider if the sample at hand happens to have thin tails.

By way of illustration, 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 4 plots the cumulative posterior distributions. With the normal model the posterior distribution of the location parameter is identical whether x_a or x_b was observed. However, if observations are known to be heavy tailed, the two sets of observations have very different implications for posterior

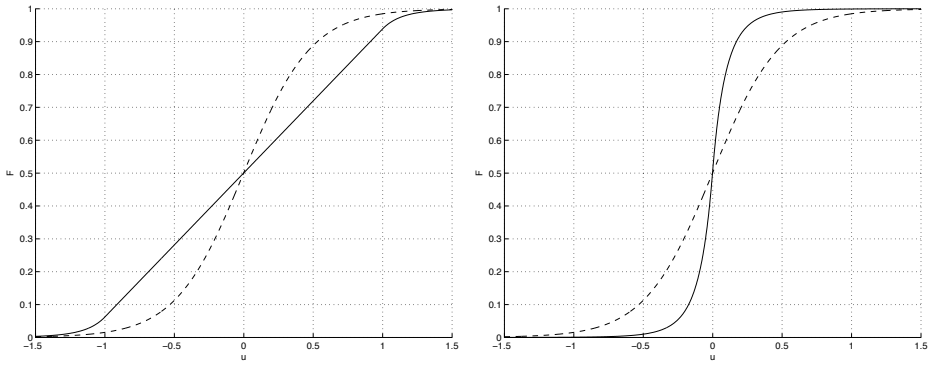

Figure 4: Posterior marginal cumulative distributions of the location parameter with a normal model of errors (dashed) and with a Laplace model of errors (solid), (a) with sample $x_a = [\ -1,\ -1,\ -1,\ -1,\ +1,\ +1,\ +1,\ +1\ ]$, and (b) with sample $x_b = [\ -2,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ +2\ ]$.

beliefs. While with a normal model the confidence intervals are the same for x_a and x_b, with a Laplace model the interval widths for an 80% confidence level differ by a factor of more than five. For x_a, the width of the 80% confidence interval with the normal model is 0.58 times the width of the same interval with the Laplace model. For x_b, this ratio is 3.08.

Figure 5 plots the distribution of the ratio between the widths of 95% confidence intervals, based on a normal model versus a Laplace model, for a sample size of 10 and when the data follow a Laplace distribution. Using a normal model for heavy-tailed data leads to confidence intervals that are, on average, overly wide. The frequent presence of points in the tails 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 heavy-tailed data than is x_a.

While we do not address correlation across judges here, we will see in §4.6 that, for our empirical data, the narrower intervals that result when the shape of the tails is correctly modeled do not appear to lead to the realized value being more frequently excluded from the confidence interval when unmodeled bias is present. Finally, note that these findings apply when a non-informative or weak prior is used for the variance, and can be reversed given better prior knowledge of the variance.

4.3 Optimal Policy for Point Estimates

When the x are drawn from a normal (or $\mathrm{GN}_{p=2}$) distribution, with an non-informative improper uniform prior, the posterior of u given the data and s is $\log f(u|x,s) \propto \sum_{i=1}^{n} (x_i - u)^2 = n\,(\overline{x} - u)^2 +$

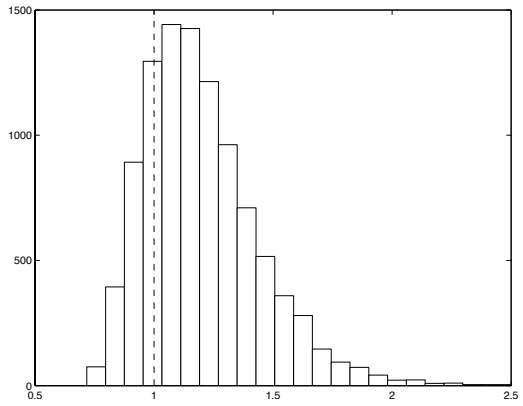

Figure 5: Ratio of the width of the 95% confidence interval based on a normal model to the width of the 95% confidence interval based on a Laplace model, with non-informative priors in both cases. The distribution is over 10 independent Laplace observations. With weak prior knowledge of the variance, a model that correctly accounts for the heavy-tailed nature of the data produces, on average, narrower confidence intervals than a normal model.

$\sum_{i=1}^{n} x_i^2 - n\overline{x}^2$, where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample average (in fact, $f(u|x,s)$ is $\mathcal{N}(\overline{x}, s/\sqrt{n})$). From $f(u|x) = \int_s f(u|x,s)df(s|x)$, and since $f(u|x,s)$ is symmetric around $\overline{x}$ for all s, we have that $f(u|x)$ is likewise symmetric. This implies that not only is the sample average $\overline{x}$ the maximum-likelihood estimate of u, it also minimizes the expected value of any symmetric loss function, including MAD and RMS.

Another estimate that is easily computed is the maximum likelihood when errors follow a Laplace (or double-exponential, $\mathrm{GN}_{p=1}$) distribution, which consists of the sample median. This is seen by noting that the posterior log-likelihood of u given the data and s is $\log f(u|x,s) \propto \sum_{i=1}^{n} |x_i - u|$, which, for all s, has a maximum at the median of the sample x.

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, the location parameter u. However, it can be significantly sub-optimal, especially with a limited number of observations as is usually the case in practical problems of aggregating expert assessments. For a GN distribution other than the normal, the optimal estimate has to be computed by 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 9 summarizes the cases considered.

Distribution of observation errors	Maximum likelihood	Minimum expected linear error (MAD)	Minimum expected quadratic error (RMS)
Normal ($\mathbf{GN}_{p=2}$)	Sample average	Sample average	Sample average
Generalized normal (arbitrary p)	*Numerical maximization*	*Median of posterior, by numerical integration*	*Average of posterior, by numerical integration*
Laplace ($\mathbf{GN}_{p=1}$)	Sample median	*Median of posterior, by numerical integration*	*Average of posterior, by numerical integration*

Table 9: Optimal estimates the location parameter under different error models and cost functions.

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-dimensional problems (two dimensional in this case: the location and scale parameters). On the other hand, 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 substantially more computationally efficient.

4.4 The Average-Median Average Heuristic

From numerical experiments, we observed that the optimal estimates for $\mathrm{GN}_{p=1}$ and $\mathrm{GN}_{p=1.5}$ models are often near to the mid-point between the average and the median. Based on this, we propose the average-median average (AMA) heuristic,

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

The policy benchmarks presented below indicate that this heuristic does in fact perform well for heavy-tailed data.

In our benchmarks, we also find that the AMA heuristic 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. Then, the loss of the policy average is

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

The proof of this result is from the optimality of the average and orthogonality of errors, which 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 average between the average and policy $g(\cdot)$ is then, by the same argument,

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

Note, in Table 11(c), that the loss of the AMA follows this relationship to the loss of the median: since the regret of the median relative to the average with normal data and RMS cost is not overly large, the regret of the AMA is approximately one-fourth of the regret of the median.

4.5 Policy Benchmarks from Simulation

Tables 10 and 11 report policy benchmarks for linear and quadratic cost functions. For each cost function, we consider the cases where the data are $\mathrm{GN}_{p=1}$, $\mathrm{GN}_{p=1.5}$, and normal, as well as sample sizes (that is, number of judges) of 3, 5, 10, and 20. The results were obtained by simulation over 10,000 trials for each case, resulting in mean standard errors of $\pm 1\%$ or less. Benchmarking the 24 cases, with two of the policies computed by numerical integration, required one day 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 data are $\mathrm{GN}_{p=1}$ when they are normal results in a substantial regret, up to 16% in the cases tested. The converse case, assuming that the data are normal (that is, using the average) whey they are $\mathrm{GN}_{p=1}$, leads to even higher regret, up to 27% in the cases tested. The policy which is optimal for $\mathrm{GN}_{p=1.5}$ data is fairly robust to the shape of the distribution's tail, and performs well for the $\mathrm{GN}_{p=1}$ and normal cases. The average-median average heuristic (AMA) performs remarkably well in all cases.

4.6 Empirical Policy Performance

We perform empirical benchmarks by using the actual or realized values of the quantities estimated or forecasted, where we were able to reliably determine them. The size of some of the datasets is

(a) $GN_{p=1}$ samples (heavy tails)

Judges	Bayes $GN_{p=1}$	Bayes $GN_{p=1.5}$	Average	Trimmed	Median	AMA
3	0%	4%	9%	—	2%	1%
5	0%	5%	15%	2%	3%	3%
10	0%	7%	22%	10%	1%	6%
20	0%	7%	27%	18%	1%	7%

(b) $GN_{p=1.5}$ samples (intermediate tails)

Judges	Bayes $GN_{p=1}$	Bayes $GN_{p=1.5}$	Average	Trimmed	Median	AMA
3	2%	0%	1%	—	9%	1%
5	3%	0%	2%	2%	11%	1%
10	4%	0%	3%	1%	8%	1%
20	6%	0%	4%	2%	9%	1%

(c) $GN_{p=2}$ samples (normal tails)

Judges	Bayes $GN_{p=1}$	Bayes $GN_{p=1.5}$	Average	Trimmed	Median	AMA
3	6%	1%	0%	—	16%	4%
5	9%	2%	0%	6%	19%	5%
10	13%	3%	0%	3%	18%	5%
20	16%	3%	0%	1%	21%	5%

Table 10: Policy benchmarks from simulation with linear cost function (relative regret, or percentage loss relative to optimal policy, in expected absolute deviation, or 'mean absolute deviation', MAD).

(a) $GN_{p=1}$ samples (heavy tails)

Judges	Bayes $GN_{p=1}$	Bayes $GN_{p=1.5}$	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%	25%	16%	2%	6%

(b) $GN_{p=1.5}$ samples (intermediate tails)

Judges	Bayes $GN_{p=1}$	Bayes $GN_{p=1.5}$	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%	2%	10%	1%

(c) $GN_{p=2}$ samples (normal tails)

Judges	Bayes $GN_{p=1}$	Bayes $GN_{p=1.5}$	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%	1%	21%	6%

Table 11: Policy benchmarks from simulation with quadratic cost function (relative regret, or percentage loss relative to optimal policy, in root of expected square deviation, or 'root mean square', RMS).

somewhat reduced due to some realizations not being available since, for the economic data, we use final revised numbers which are available with a substantial lag. We also exclude the smaller datasets. For the estimates of the number of member countries of the United Nations we use the reference value 192. This was not the correct answer in the earliest years in which the surveys were administered, but running the benchmarks using 189 to 191 as the reference value yielded very similar results. As before, we work with the logarithm of the data for quantities that are restricted to be positive (hence, to be precise, the reference value used for the UNLO and UNHI datasets is $\log(192)$).

Table 12 provides some summary statistics for the datasets included, as well as benchmarks for the average, median, and average-median average. For each column of data we conduct 10,000 trials, where in each trial we randomly sample 10 estimates and then apply each policy. The results are averaged over the columns.

Note the large positive bias on all economic datasets, with typically on the order of 90% or more of the forecasts above the eventual realization, which may be related to most of the economists on the panel being associated with sell-side firms. This large bias is the dominant source of forecasting error, with only a small performance difference between the policies, albeit with a small advantage for the average, on the order of 1%. A partial exception is the 10-year rate (which is less biased, with 20% of the forecasts below the realization), where the average underperforms by about 5%. 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 use.

The estimates made by MBA students of the number of member countries of the United Nations are less biased. 59% of the estimates were less or equal to 192 among the respondents anchored to 95 countries (UNLO), and 32% among those anchored to 300 countries (UNHI). The underperformance of the average here is on the order of 10%, consistent with the previous results from simulation.

We provide further details for these two datasets where bias is not the overriding source of error in Tables 13 and 14. Table 13 adds the optimal Bayesian estimate with a $\text{GN}_{p=1.3}$ model, as well as including, for each dataset and policy, the same four cases used in the simulation-based benchmarks of 3, 5, 10, and 20 judges. The average underperforms relative to the $\text{GN}_{p=1.3}$-optimal policy and to the AMA by around 5% to 10%.[4]

Table 14 illustrates the impact on the average width of the confidence interval of using a normal

[4]The numbers on Tables 13 and 14 are different from the corresponding cases in Table 12 due to a different sampling procedure. Instead of sampling from each column separately, we sample from the entire dataset (that is, as a single column). Also, due to the large number of cases and the computational requirements of the Bayesian estimates, we only conduct 1,000 trials which results in somewhat higher margins or error.

Dataset	N	n	$\widehat{I}\{x_{ij} \leq x_j^0\}$	Cols. where AMA bests average (MAD)	Cols. where AMA bests average (RMS)	Relative regret (MAD) Average	Relative regret (MAD) Median	Relative regret (MAD) AMA	Relative regret (RMS) Average	Relative regret (RMS) Median	Relative regret (RMS) AMA
GDP	1730	32	0.13	0.34	0.44	0%	2%	1%	0%	2%	1%
GDPa	1294	24	0.04	0.25	0.29	0%	2%	1%	0%	2%	1%
NFARM	1393	27	0.12	0.26	0.26	0%	1%	0%	0%	1%	0%
CPI	866	16	0.03	0.25	0.25	0%	2%	1%	0%	2%	1%
CPIa	596	11	0.01	0.27	0.27	0%	1%	1%	0%	1%	0%
R10Y	705	13	0.20	0.62	0.62	4%	0%	2%	5%	0%	2%
R10Ya	485	9	0.10	0.22	0.22	0%	2%	1%	0%	2%	1%
UNLO	2025	18	0.59	0.94	0.94	12%	0%	6%	8%	0%	3%
UNHI	2018	18	0.32	0.89	0.78	15%	0%	2%	11%	2%	0%

Table 12: Empirical policy benchmarks based on resampling 10 judges from different datasets. N is the total number of data, n the number of columns (each possibly subject to a different unmodeled common bias across experts). $\widehat{I}\left\{x_{ij} \leq x_j^0\right\}$ is the number of observations below the realized value, and summarizes the overall bias in each dataset. We report the fraction of columns where the average performance of the AMA heuristic was superior to that of the average (based on resampling 10 judges from each column), as well as the overall relative regret with the average, median, and AMA. In datasets with less bias, the AMA heuristic outperforms the average. When the unmodeled bias is high (as is the case for all datasets from the panel of economist, only somewhat less so for R10Y), this dominates the estimation error and the difference between policies is small.

UNLO (MAD)

Judges	Average	Median	AMA	Bayes
3	6.1%	0.0%	2.5%	3.4%
5	8.4%	0.0%	3.8%	3.4%
10	10.3%	0 .0%	5.1%	3.1%
20	10.2%	0.0%	5.1%	2.5%

UNLO (RMS)

Judges	Average	Median	AMA	Bayes
3	2.8%	0.1%	0.0%	1.2%
5	4.1%	0.0%	0.8%	0.5%
10	7.7%	0.0%	3.1%	1.4%
20	8.1%	0.0%	3.6%	1.0%

UNHI (MAD)

Judges	Average	Median	AMA	Bayes
3	7.4%	2.1%	0.0%	1.1%
5	9.5%	3.8%	0.5%	0.0%
10	10.6%	2.6%	0.0%	0.3%
20	11.1%	5.7%	0.0%	3.3%

UNHI (RMS)

Judges	Average	Median	AMA	Bayes
3	5.6%	5.4%	0.0%	1.8%
5	7.7%	7.4%	0.6%	0.0%
10	9.2%	5.8%	0.1%	0.0%
20	7.7%	10.5%	0.0%	1.9%

Table 13: Relative regret based on resampling from the UNLO and UNHI datasets. The Bayesian estimate is based on a $\mathrm{GN}_{p=1.3}$ model.

UNLO

Judges	α_2	$\alpha_{1.3}$	Δ_2	$\Delta_{1.3}$
3	0.17	0.20	1.42	1.32
5	0.34	0.38	0.91	0.84
10	0.76	0.73	0.56	0.53
20	0.98	0.96	0.37	0.35

UNHI

Judges	α_2	$\alpha_{1.3}$	Δ_2	$\Delta_{1.3}$
3	0.06	0.06	1.51	1.40
5	0.06	0.07	0.99	0.91
10	0.06	0.06	0.62	0.55
20	0.07	0.09	0.42	0.37

Table 14: Confidence intervals with 3, 5, 10, and 20 judges based on resampling from the UNLO and UNHI datasets. The α are the frequencies with which the true value (192) is not included in the 95% confidence interval. The Δ are the average widths of the intervals. Results are for a normal model (α_2 and Δ_2) and for a $\mathrm{GN}_{p=1.3}$ model ($\alpha_{1.3}$ and $\Delta_{1.3}$). 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. Note that the intervals based on the $\mathrm{GN}_{p=1.3}$ model (that is, assuming heavier tails) have, over the different cases, consistently smaller average width than the intervals based on the normal model, but that this does not translate to consistently higher α.

versus a $GN_{p=1.3}$ model. Note that for a larger number of experts the confidence interval becomes narrower and, where unmodeled bias is significant, excludes the true value with increasing frequency. As expected given §4.2, the normal model leads to confidence intervals that are, on average, wider. The intervals based on the $GN_{p=1.3}$ model (that is, when correctly assuming that tails are heavy) 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.

5 Conclusion

People are not normal: human judgment has fat tails, in the sense that large deviations from the consensus are far more frequent than predicted by a normal distribution. In our empirical analysis, we find the thickness of the tails of judgment to show a degree of consistency across different tasks, different levels of expertise, and different degrees of uncertainty about the quantity in question. We believe that this makes a persuasive argument for moving away from the normality assumption in the aggregation of expert assessments and forecasts.

Optimal aggregation of estimates and forecasts should incorporate prior knowledge about the distributional characteristics of human judgment. While the thickness of tails cannot be estimated from the small samples that will typically be available in any given problem, the consistency we observe in the data we have analyzed supports the use of prior knowledge. A generalized normal model, where the probability density of an error is proportional to the inverse of the exponential of the magnitude of the error to the power of p (with $p = 2$ the normal distribution, and $p = 1$ the Laplace, or double-exponential distribution) seems to provide a good fit for the judgmental estimation and forecasting data examined. If a generalized normal model is used, we suggest a shape parameter $p = 1.3$. 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, and in the teaching of business students, the average-median average heuristic (applied to the logarithm of the estimates when dealing with quantities that are restricted to be positive) is an adequate, robust alternative.

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

Acknowledgements

We would like to thank Enrico Diecidue, Theodoros Evgeniou, Michele Hibon, Ioana Popescu, and Ilia Tsetlin for supplying us with data from MBA judgmental exercises.

Theodoros Evgeniou, Ilia Tsetlin, and Robert Winkler provided many detailed comments and suggestions on an early version of this paper, we are very grateful for their generosity. We also thank Enrico Diecidue, Lily Fang, and Spyros Makridakis for helpful discussions.

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.

Bunn, D. 1989. Forecasting with more than one model. *Journal of Forecasting* **8**(3) 161–166.

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. 2009. Are stars' opinions worth more? The relation between analyst reputation and recommendation values. AFA 2006 Boston Meetings Paper; EFA 2007 Ljubljana Meetings Paper. Available at SSRN: http://ssrn.com/abstract=687491.

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.

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.

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.