# Internet Appendix to "Robust Measures of Earnings Surprises"*

This internet appendix has two parts. The first section discusses both the theoretical and empirical motivations for the use of mixture-normal distributions in generating individual forecast errors. The second section discusses numerical calculations and extensions of our robust measure of earnings surprises.

## I. Mixture-Normal Distribution for Individual Forecast Errors

Our mixture-normal model featuring a fraction $\omega_0$ of unbiased and a fraction $(1-\omega_0)$ of biased forecasts can be motivated from a number of theories of professional forecast bias. In Lim (2001), analysts with sources of information, i.e. working at a prestigious brokerage house, do not issue biased forecasts. Those without such sources of information will bias upward their forecasts to gain access to management. Applying the Scharfstein and Stein (1990) model of career concerns and herding to analyst forecasts, analysts without career concerns will issue unbiased forecasts while those with career concerns will issue biased forecasts. In the Laster, Bennett, and Geoum (1999) model of biased macro-forecasters due to clienteles, forecasters catering to intensive users will issue unbiased forecasts, while those catering to occasional users will issue biased forecasts. These theories can all be expanded so that the degree of bias can vary depending on other parameters. But the crux of these models very much feature unbiased versus bias forecasters, which then provides a theoretical motivation for our use of a mixture-normal distribution to generate individual forecast errors. The connection of bias to the precision of forecasters (i.e. the talent of the forecaster) depends on the particular model of bias. For instance, in Laster, Bennett, and Geoum (1999) and Scharfstein and Stein (1990), even talented forecasters (i.e. those with high initial precision) can be biased depending on their clienteles or career concerns. In Lim (2001), forecasters will lower initial precision of forecasts are more likely to issue biased forecasts so as to increase their precision. As a result, we allow for a number of parameters to drive our data-generating process.

We next show that such a mixture-normal distribution better fits the distribution of individual forecast errors than the normal distribution. We implement the Shapiro and Wilk (1965) test of the null hypothesis that a sample $x_1, \cdots, x_n$ came from a normally distributed population. The test statistic is $\frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, where $x_{(i)}$ is the $i$-th order statistic, $\bar{x}$ is the sample mean, and the constants $a_i$'s are based on the expected values and covariance matrix of the order statistics of i.i.d. normal random variables. We use this test to show that one-component normal distribution (with possibly different mean and variance for each firm-year pair) is not sufficient to characterize analyst forecast errors. Note that one should use firm-year pairs instead of pulling all the forecast errors together to run normality tests because the magnitude of errors and biases can vary across firms and over time.

To run this test, we require a minimum of 10 forecasts per firm-year pair, and those below the threshold are dropped. If the forecasts for each firm-year pair were from normal distribution, we should expect the p-values of the Shapiro and Wilk (1965) tests to be uniformly distributed between 0 and 1. However, as shown in Figure IA.1, the p-values based on real data is far from being uniformly distributed. For reference we also simulated forecasts from one-component normal distributions that match the number of forecasts, average value, and standard deviation of forecast errors for each firm-year pair, and reran the tests. The p-values based on this hypothetical dataset are much closer to uniform distribution as expected.

*Chin-Han Chiang and Wei Dai and Jianqing Fan and Harrison Hong and Jun Tu, 2018, Internet Appendix to "Robust Measures of Earnings Surprises", Journal of Finance vol no., pages, http://www.afa.jof.org/supplements.asp. Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the authors of the article.

The evidence suggests that using one-component normal is not sufficient to characterize the real data. While there are many ways to introduce more model complexity, using two-component mixture normal model is a simplified way to capture the heterogeneity in the forecasts, an approach we took in this paper. Furthermore, the marginal benefit of adding more components into the mixture seems limited. More specifically, we will fit the forecast errors for each firm-year pair to one-component normal, and two- and three-component mixture normal models and compare overall model fit using BIC. For the one-component normal, we need to estimate the mean and variance of the forecast error $U \sim \mathcal{N}(\mu, \sigma^2)$. For the two-component normal, forecast error $U$ for each firm-year pair follows $\mathcal{N}(\hat{S}, \sigma_F^2)$ with probability $\omega_0$, and $\mathcal{N}(\hat{S}-\hat{B}, \sigma_F^2+\sigma_b^2)$ with probability $\omega_1 = 1-\omega_0$ where $\hat{S}$ is the realized true surprise and $\hat{B}$ is the realized aggregate bias; the parameters to be estimated are $\omega_0$, $\mu_1 = \hat{S}$, $\sigma_1^2 = \sigma_F^2$, $\mu_2 = \hat{S}-\hat{B}$, and $\sigma_2^2 = \sigma_F^2+\sigma_b^2$. Similarly, for three-component normal, the additional parameters are the mean, variance and mixture weight of the third component. Note that the components in the mixture are interchangeable, i.e., we do not need to know which set of $\mu$'s and $\sigma$'s corresponds to the unbiased or biased component to fit the model. That means, fitting the mixture model does not represent a feasible way of coming up with a de-biased estimate for earnings surprise.

In order to estimate all the parameters associated with each model, we require at least 30 observations per firm-year pair. In addition, some firm-year pairs are dropped because the optimization procedure does not converge (for example, if there are too few unique values even though the number of observations exceeds 30). We end up with 929 firm-year pairs for this analysis. Figure IA.2 shows the differences in BIC values across firm-year pairs.[1] In both plots a positive difference means that the two-component mixture normal is better (has a lower BIC value) than the alternative. While there is no model that is always the best across all firm-year pairs, two-component mixture is better than one-component for 63.5% of the cases, while two-component is better than three-component for 79.4% of the cases. Overall the results suggest that two-component mixture is a reasonable modeling choice.

# II. Numerical Calculations and Extensions

In this section, we provide more color on how bias affects the relative performance of *CE*, *Rank(CE)* and *FOM* and why *FOM* is a robust measure of surprises $S$.

## A. *Biased Forecasts:* $\omega_1 > 0$

While *CE* can be large simply due to the existence of one very negative $F_i$, *FOM* is much less affected because each observation only contributes as 1 or $-1$ regardless of its magnitude. One consequence is that *CE* and *FOM* are no longer highly correlated. While we observe a rather low correlation in earnings data, which is also due to outliers, here we use simulations to reveal part of the dynamic caused by biased forecasts. We simulate data according to the model and calculate the correlations using $50,000$ samples, where the key parameters $\omega_1$ and $r_B = \frac{\sigma_B}{\sigma_A}$ vary over their range, and the others fixed at $N = 20$, $r_F = 1/2$ and $r_b = r_B/5$. Appendix Figure IA.3 shows how the correlation decreases with $r_B$, the relative uncertainty level of the bias component $B$. In terms of $\omega_1$, recall it is the proportion of biased forecasts, so the correlation first decreases with the introduction of biased forecasts as soon as $\omega_1$ becomes nonzero, and then picks up when both measures get equally bad.

Along with the lower correlation between these two measures, the discrepancy between their performance measuring market surprise also widens, mainly due to their different resistance to bias. We have shown earlier that *FOM* will eventually outperform *CE* as bias becomes more significant, because *FOM*'s correlation with $S$ has a positive lower bound whereas $\text{Cor}(CE, S)$ can be reduced to zero quickly. Indeed this is what we observe in simulation studies. As an illustration, again let the key parameters $\omega_1$ and $r_B$ vary over their range, with the others fixed at $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$. We directly compute the correlation between *CE* and $S$ from our formula and simulate $100,000$ samples of $X$ and $Y$ to compute the correlation between *FOM* and $S$. Appendix Figure IA.4 shows a representative pattern of their relative performance as a function of $\omega_1$ and $r_B$, where the difference between $\text{Cor}[CE, S]$ and $\text{Cor}[FOM, S]$ becomes negative (i.e., *FOM* outperforms) as the relative dispersion of bias $r_B = \sigma_B/\sigma_A$ increases.

### A.1. *CE and Rank(CE)*

In practice, people use *Rank(CE)*, i.e., sort *CE* into 10 deciles in order to be robust to outliers. However, this global adjustment may not work in the presence of bias. For example, one single large biased forecast can still move *CE* from decile 10 down to decile 1 and distort the ordering. Appendix Figure IA.5 shows a representative pattern of the difference in the performance of *Rank(CE)* and *FOM* (i.e., $\mathrm{Cor}[\mathit{Rank(CE)}, S] - \mathrm{Cor}[\mathit{FOM}, S]$) as a function of $\omega_1$ and $r_B$ with the same set of parameters as above, where each $\mathrm{Cor}[\mathit{Rank(CE)}, S]$ is computed using 50,000 simulated samples. Comparing with Appendix Figure IA.4, there is some improvement when $r_B$ is not too large. However, the essence of the analysis on *CE* carries over to *Rank(CE)* because when *CE* is greatly contaminated, the coding of *Rank(CE)* does not help much: the damage is already done. In this sense, *FOM* measure does the robustness adjustment on a local level, so the impact from bias is alleviated when aggregating $N$ forecasts, instead of afterwards. Therefore, *FOM* improves over *Rank(CE)* for the same reason as it does over *CE*, the reason being their sensitivity to large bias. That being said, *Rank(CE)* does have better property when treating the few outliers that overthrow *CE*.

## B. *Winsorized Mean and Median*

In order to be robust to the noisy forecasts, one may also Winsorize the forecasts. For example, a 5% Winsorization would set all forecasts below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile. The average of the resulting data is the Winsorized mean of forecasts. Similarly, we can define the Winsorized consensus error as

$$CE_{\lambda}^{win} = A - \bar{F}_{\lambda}^{win},$$

where $\lambda$ is the percentage of data on each tail being replaced. Note that when $\lambda = 50\%$, the Winsorized mean becomes median:

$$CE_{50\%}^{win} = CE_{med} = A - \mathrm{median}(F_i).$$

However, such measures do not show much, if any, improvement in our regression results of earnings announcement event study. This is not surprising because although Winsorization is designed to remove the two tails in a set of forecasts, it is by no means equivalent to removing the biased ones. Since the realization of bias is unknown in each draw, it is impossible for Winsorization to correctly pick up all the bad forecasts without sacrificing the good ones. In the same spirit as the analysis of consensus errors, the Winsorized measures by definition still strongly depend on the magnitude of forecasts, which inevitably leads to their vulnerability to bias. The more volatile $B$ is, the harder it is for Winsorization to achieve consistent performance. Appendix Figure IA.6 illustrates how the performance drops with increasing $r_B$ through 5000 simulations, where the other parameters in the model are set as $\omega_1 = 0.3$, $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$.

Furthermore, the performance also depends on the fraction of biased forecasts and the choice of $\lambda$ for Winsorization. Unfortunately, the fraction of biased forecasts $\omega_1$ is usually unknown in practice and may even be varying, so it is hard if not impossible to set $\lambda$, the single important parameter for Winsorization, and an inappproriate choice might result in undesirable performance. This is illustrated in Appendix Figure IA.7, where the relative performance of different Winsorized measures changes with the fraction of biased forecasts $\omega_1$, and the other parameters in the model are set as $r_B = 10$, $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$.

## C. *Remark on the Model*

A key assumption in our model is that for each stock a fraction of analysts are biased. Recall that under our modelling, the forecasts come from a mixture composed of two normal distributions, one centered around the unknown market expectation $e$ and the other biased by a magnitude of the realized $B$. While the aggregated bias magnitude $B$ can be huge or moderate, $\omega_1$ the weight of the biased distribution in the mixture is with respect to $N$ so the number of biased analysts scales with the total number and makes the law of large numbers fail. In this normal mixture framework, the bias component is essential and we have shown how it drives the behaviour of different measures that is consistent with our observations. If we remove the bias part of the modelling and instead introduce bad forecasts by having large variance in one of the distributions, it will fail to represent some important features in the real data. More specifically, suppose the forecasts are given by

$$F_i = e + \epsilon_i, \tag{1}$$

where $\epsilon_i$'s follow a mixture of two normal distributions: $\mathcal{N}(0, \sigma_0^2)$ with probability $\omega_0$ and $\mathcal{N}(0, \sigma_1^2)$ with probability $\omega_1 = 1 - \omega_0$, and $\sigma_1^2 > \sigma_0^2$. Notice that this is actually a limiting case of our baseline specification by setting $\sigma_B = 0$, which means $B$ is always 0 so that its impact disappears. Under this alternative modelling, even though individual forecasts can be very volatile, the variance of the average forecast error is given by:

$$\text{Var}[\frac{1}{N}\sum_{i=1}^{N}\epsilon_i] = \frac{1}{N}(\omega_0\sigma_0^2 + \omega_1\sigma_1^2), \tag{2}$$

so *CE* still converges to $S$ by the law of large numbers. That is, although $\sigma_1^2$ can be large, the distortion from fat-tails is greatly discounted and the variance decreases linearly in $N$, unlike in the original model the variance of the average noise never vanishes no matter how big $N$ is. This implies that *CE* or *Rank(CE)* should be better for larger $N$ under the alternative model, which does not quite match what we see in the real data.

Furthermore, in the absence of random bias all the forecasts are centered around the real market expectation $e$, so it is much easier for Winsorization to filter the bad forecasts. As a comparative example to Appendix Figure IA.6, Appendix Figure IA.8 illustrates the much stronger performance of Winsorized mean and median through 5000 simulations, which is again different from what we see in the empirical study and undermines the validity of this alternative modelling.

## D. *Extended Model*

Our model above assumes that the market's expectation conditions on information outside the set of analyst forecasts. But we can model the market's expectation as dependent just on the set of analysts' forecasts and obtain the same results.

Suppose now that $A \sim \mathcal{N}(0, \sigma_A^2)$ for simplicity. There are $i = 1, ..., N$ forecasts. We then assume that individual forecasts $i$ is given by

$$F_i = \begin{cases} A + \epsilon_i & \text{with prob. } \omega_0 \\ A + b_i + \epsilon_i & \text{with prob. } \omega_1 = 1 - \omega_0 \end{cases} \tag{3}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_F^2)$ and is uncorrelated with the randomness in $A$. Each forecast is unbiased with probability $\omega_0$, and is contaminated by an individual bias term $b_i$ with probability $\omega_1 = 1 - \omega_0$. We model the bias in the same manner as before. For each set of $N$ forecasts an aggregated bias level $B \sim \mathcal{N}(0, \sigma_B^2)$ is drawn first, and conditional on this realized $B$ individual bias $b_i$ follows $\mathcal{N}(B, \sigma_b^2)$.

We assume that investors are able to de-bias whereas the econometrician cannot. Hence, the market's posterior of $A$ is given by

$$\hat{A} = \frac{1}{N}\sum_{i=1}^{N} F_i^*, \tag{4}$$

where $F_i^* = A + \epsilon_i$ is the debiased forecasts. This follows from the usual Kalman Filtering results in linear-normal models where each forecast can be interpreted as a linear signal of the actual $A$. Since each signal has equal precision, there is then equal weighting of the signals in forming the posterior $\hat{A}$. The market surprise then is given by

$$S = A - \hat{A} \tag{5}$$

Notice that *CE* is now given by

$$CE = A - \frac{1}{N}\sum_{i=1}^{N} F_i \tag{6}$$

and *FOM* is now given by

$$FOM = \frac{1}{N}\sum_{i=1}^{N}(I_{F_i<A} - I_{F_i>A}) \tag{7}$$

We want to compare again the correlation of $CE$ and $FOM$ with the market surprise $S$, respectively, We can calculate that

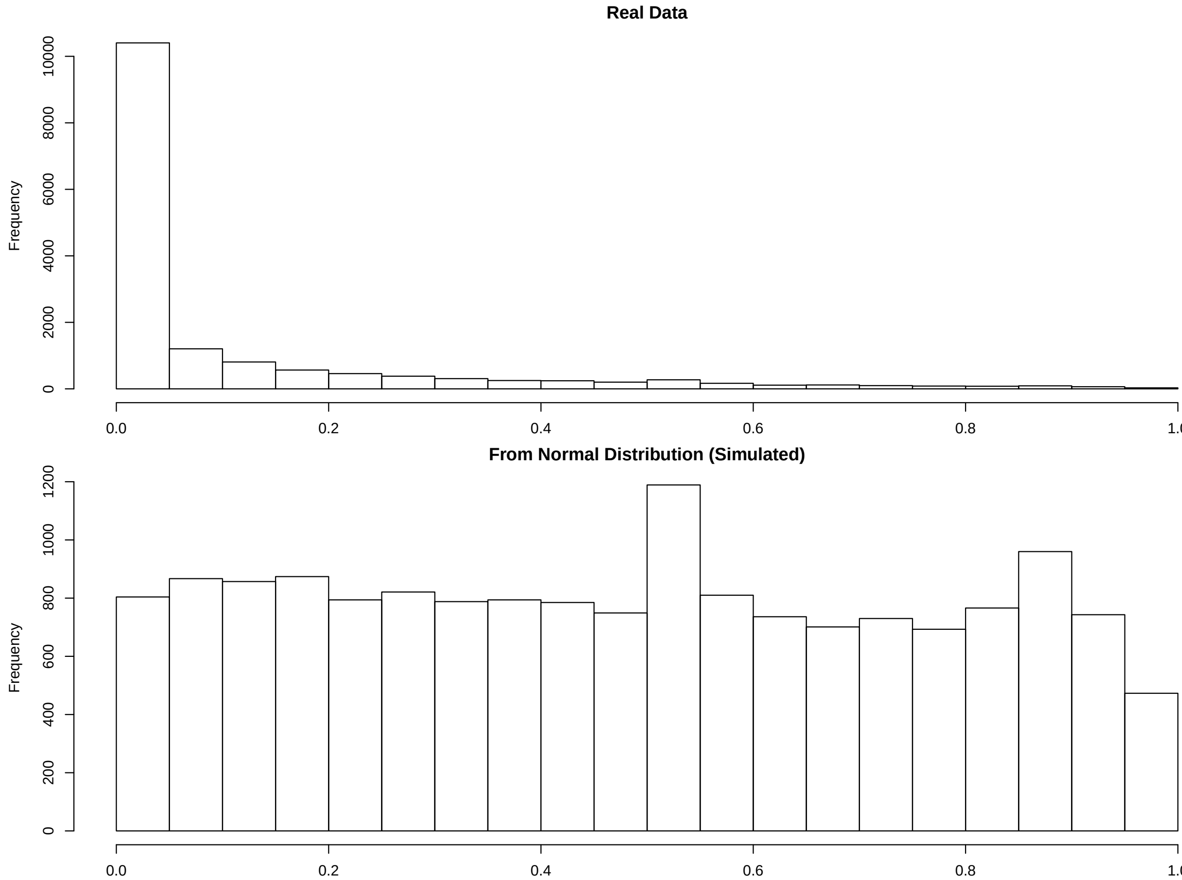
$$\mathrm{Cor}(CE, S) = \frac{1}{\sqrt{1 + \omega_0\omega_1 r_B^2 + \omega_1 r_b^2 + \omega_1^2 r_B^2 N}} \tag{8}$$

where $r_B = \sigma_B/\sigma_F$ and $r_b = \sigma_b/\sigma_F$. We can also show that

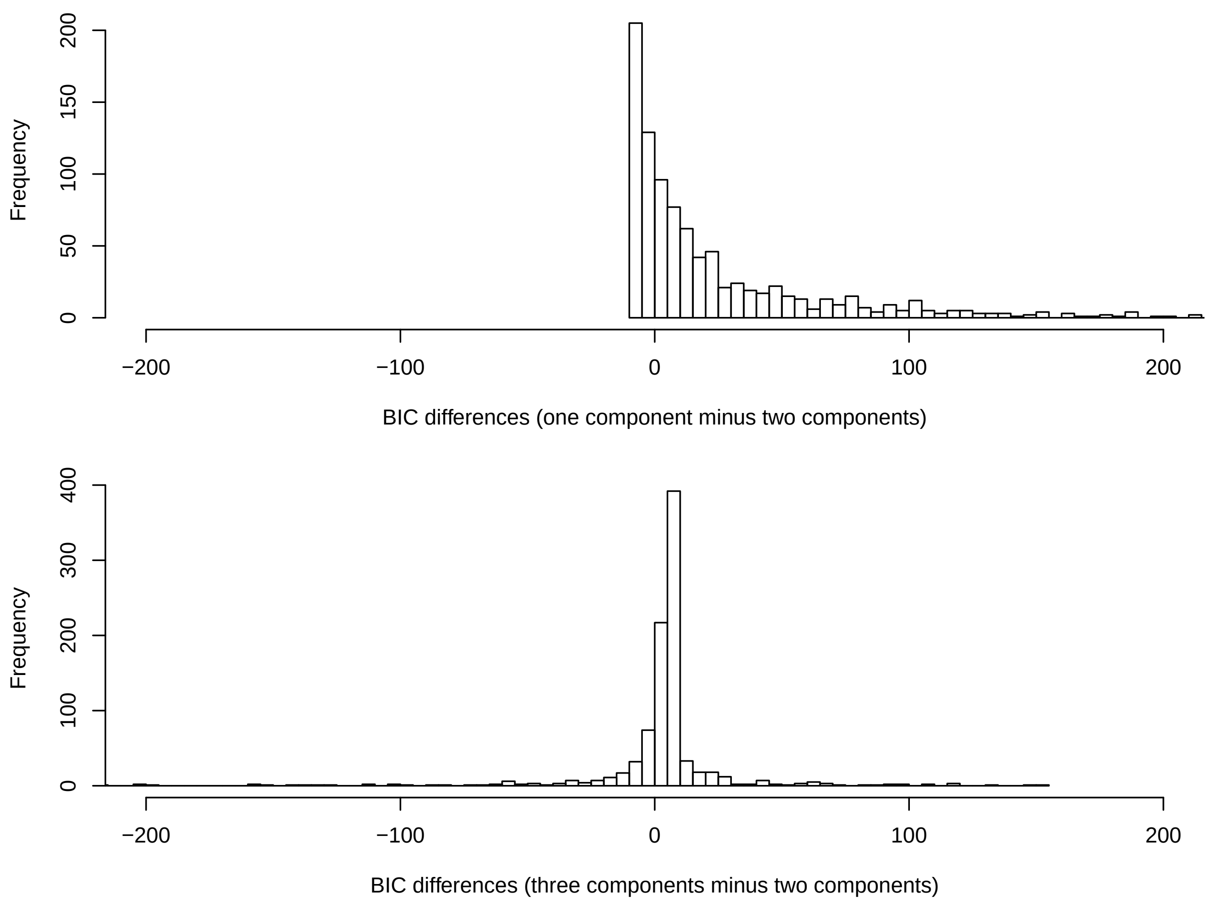
$$\mathrm{Cor}(FOM, S) = \frac{\omega_0 \frac{1}{\sqrt{2\pi}} + \omega_1 \mathrm{E}[X\Phi(\tilde{X} - Y)]}{\sqrt{\frac{\omega_0}{2}(1 - \frac{\omega_0}{2}) + \omega_1^2 \mathrm{E}[\Phi(\tilde{X} - Y)(1 - \Phi(\tilde{X} - Y))] + N\omega_1^2 \mathrm{Var}[\Phi(\tilde{X} - Y)]}} \tag{9}$$

where $X \sim \mathcal{N}(0, 1)$ and $\tilde{X} = X/r_b$ which is orthogonal to $Y \sim \mathcal{N}(0, \frac{r_B^2}{r_b^2})$.

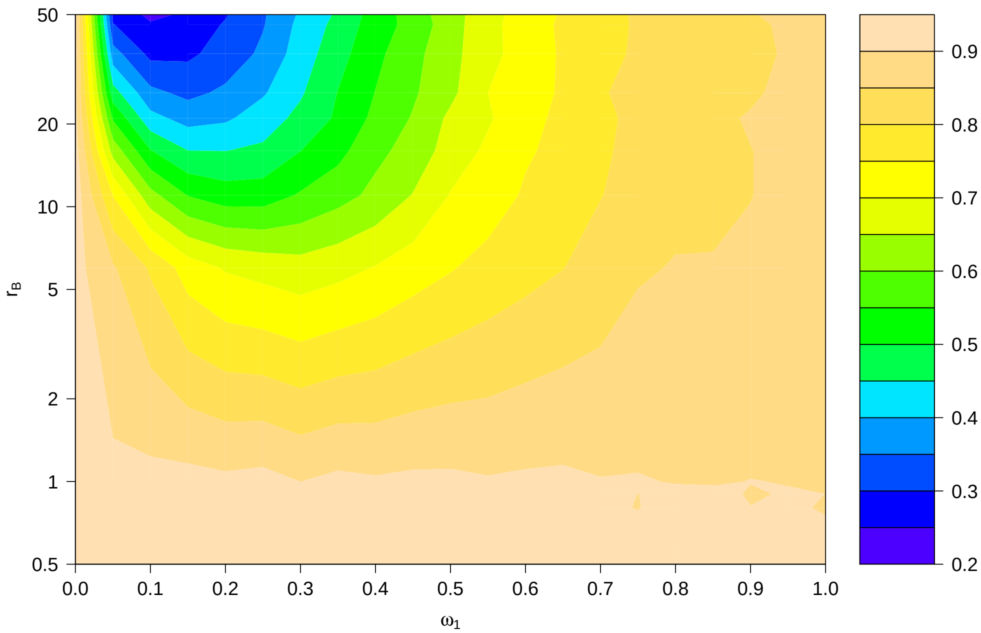
Since $\mathrm{Cor}(FOM, S) \geq \frac{w_0\sqrt{2/\pi}}{\sqrt{1+\omega_1^2 N}}$, it follows then that if $r_B$ gets large, then $\mathrm{Cor}(CE, S)$ drops below $\mathrm{Cor}(FOM, S)$. This then confirms our results in our baseline model.



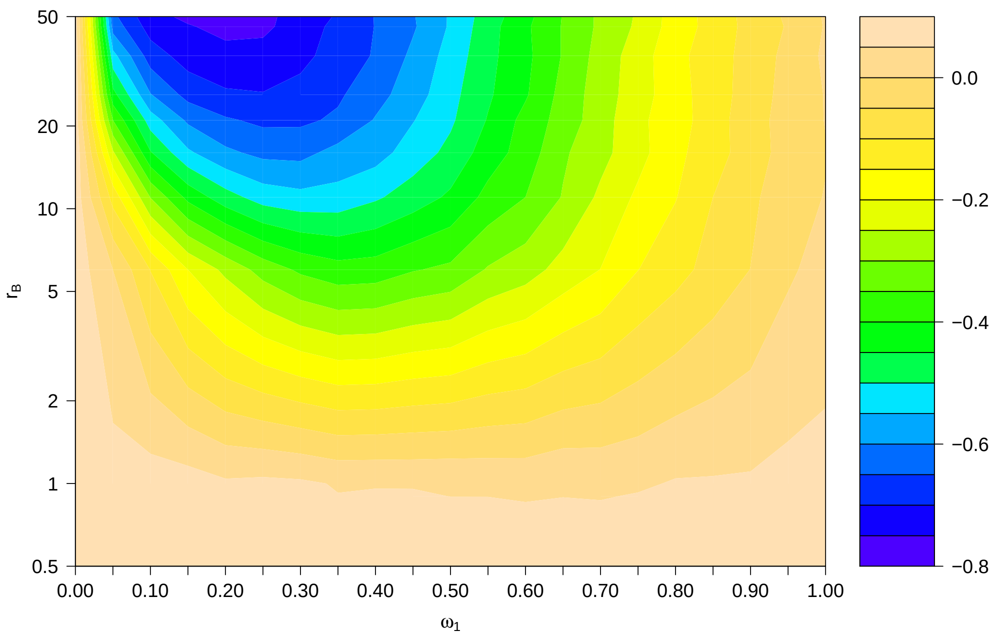

Appendix Figure IA.1: $p$ values of the Shapiro-Wilk tests



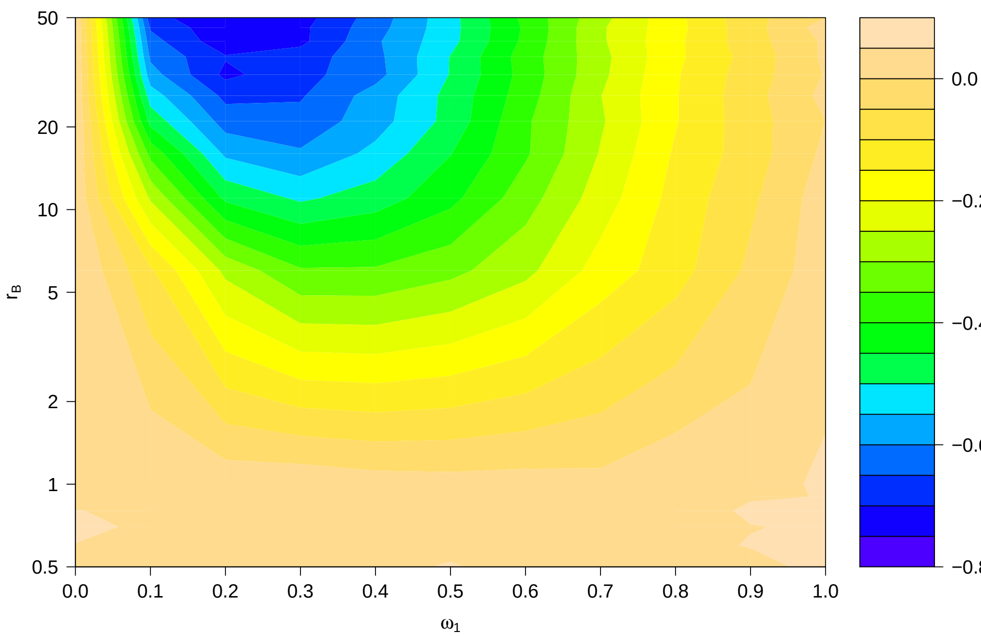

Appendix Figure IA.2: BIC differences comparing a one-component normal model to a 2-component mixture model in the top panel, and these difference comparing a 2-component mixture model to a 3-component mixture model in the bottom panel.



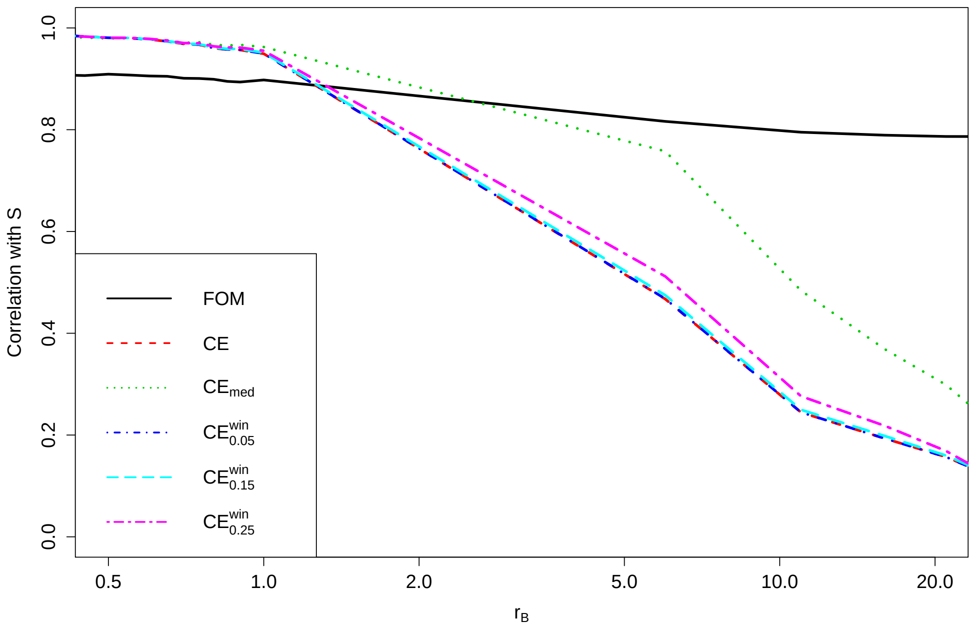

Appendix Figure IA.3: The contour plot of Cor[$CE, FOM$] as a function of the key parameters $\omega_1$ and $r_B$ in biased forecasts case. The contour value is the correlation between consensus errors $CE$ and fraction of misses $FOM$, the y-axis is $\omega_1$ the proportion of biased forecasts, and the x-axis is $r_B = \sigma_B/\sigma_A$ the ratio between the standard deviation of aggregated bias and the actual (shown in log-scale). The other parameters in the model are set as $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$.



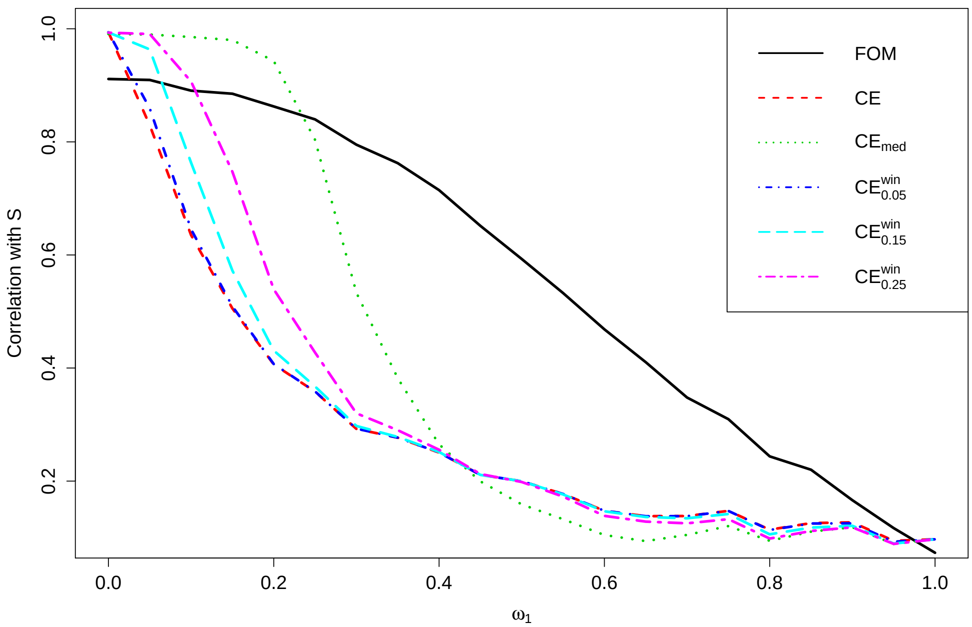

Appendix Figure IA.4: The contour plot of $\text{Cor}[CE, S] - \text{Cor}[FOM, S]$ as a function of the key parameters $\omega_1$ and $r_B$ in biased forecasts case. The contour value is the difference between the correlations of consensus errors $CE$ and fraction of misses $FOM$ to $S$ the market surprise, the y-axis is $\omega_1$ the proportion of biased forecasts, and the x-axis is $r_B = \sigma_B/\sigma_A$ the ratio between the standard deviation of aggregated bias and the actual (shown in log-scale). The other parameters in the model are set as $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$.



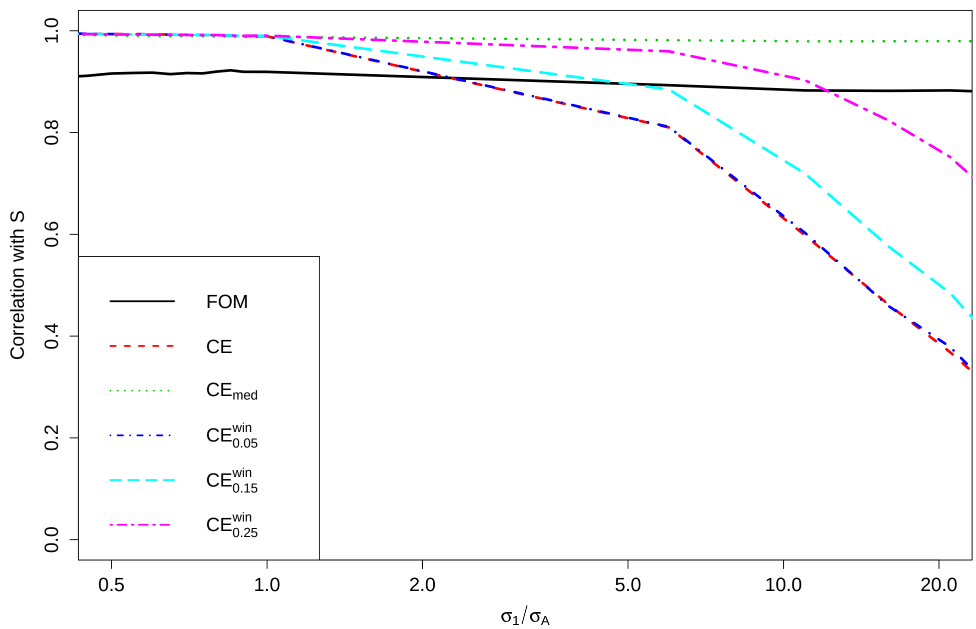

Appendix Figure IA.5: The contour plot of $\text{Cor}[Rank(CE), S] - \text{Cor}[FOM, S]$ as a function of the key parameters $\omega_1$ and $r_B$ in biased forecasts case. The contour value is the difference between the correlations of the rank score of consensus errors $Rank(CE)$ and fraction of misses $FOM$ to $S$ the market surprise, the y-axis is $\omega_1$ the proportion of biased forecasts, and the x-axis is $r_B = \sigma_B/\sigma_A$ the ratio between the standard deviation of aggregated bias and the actual (shown in log-scale). The other parameters in the model are set as $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$.




Appendix Figure IA.6: The comparison between the correlations of fraction of misses $FOM$ and different Winsorized measures $CE_{\lambda}^{\text{win}}$ to $S$ the market surprise as a function of $r_B$ in biased forecasts case, where $r_B = \sigma_B/\sigma_A$ is the ratio between the standard deviation of aggregated bias and the actual (shown in log-scale). The other parameters in the model are set as $\omega_1 = 0.3$, $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$.




Appendix Figure IA.7: The comparison between the correlations of fraction of misses *FOM* and different winsorized measures $CE^{\text{win}}_{\lambda}$ to $S$ the market surprise as a function of $\omega_1$ in biased forecasts case, where $\omega_1$ is the proportion of biased forecasts. The other parameters in the model are set as $r_B = 10$, $r_F = 1/2$, $r_b = r_B/5$ and $N = 20$.




Appendix Figure IA.8: The comparison between the correlations of fraction of misses *FOM* and different winsorized measures $CE_\lambda^{\text{win}}$ to $S$ the market surprise as a function of $\sigma_1/\sigma_A$ (shown in log-scale) under the alternative modelling without introducing bias, where $\sigma_1$ is the variance of bad forecasts. The other parameters in the model are set as $\omega_1 = 0.3$, $\sigma_0/\sigma_A = 1/2$ and $N = 20$.

# References

Laster, D., P. Bennett, and I. S. Geoum, 1999, "Rational bias in macroeconomic forecasts," *The Quarterly Journal of Economics*, 114(1), 293–318.

Lim, T., 2001, "Rationality and Analysts' Forecast Bias," *The Journal of Finance*, 56(1), 369–385.

Scharfstein, D. S., and J. C. Stein, 1990, "Herd behavior and investment," *The American Economic Review*, pp. 465–479.

Shapiro, S. S., and M. B. Wilk, 1965, "An analysis of variance test for normality (complete samples)," *Biometrika*, 52(3/4), 591–611.