Forecasting with Unknown Unknowns: Censoring and Fat Tails on the Bank of England's Monetary Policy Committee*

James Mitchell† and Martin Weale‡

July 2019

Abstract

This paper considers the production and evaluation of density forecasts paying attention to if and how the probabilities of outlying observations are quantified and communicated. Particular focus is given to the 'censored' nature of the Bank of England's fan charts, given that - which is commonly ignored - they describe only the inner 90% (best critical region) of the forecast distribution. A new estimator is proposed that fits a potentially skewed and fat tailed density to the inner observations, acknowledging that the outlying observations may be drawn from a different but unknown distribution. In forecasting applications, motivation for this could reflect the view that outlying forecast errors reflect (realised) unknown unknowns or events not expected to recur that should be censored before quantifying known unknowns.

Keywords: Forecast uncertainty; Fan charts; Skewed densities; Best critical region; Density forecasting; Censoring; Forecast evaluation
JEL Classification: C24; C46; C53; E58

*We are grateful to conference and seminar participants at the Bundesbank, the National Bank of Poland, Henley Business School (Reading), Strathclyde, WBS, LSE, SNDE (FRB Dallas) and FRB New York for helpful comments. Particular thanks for their comments to Jamie Bell, Andrew Harvey, Malte Knüppel, Gary Koop, David Latto and John Maheu.

†Warwick Business School, University of Warwick; Economic Statistics Centre of Excellence (james.mitchell@wbs.ac.uk)
‡King's College London; Economic Statistics Centre of Excellence (martin.weale@outlook.com)

1 Introduction

Following more than twenty years of practice at the Bank of England, density or fan chart forecasts are now widely published and used in many contexts as a means of communicating forecast uncertainties. Like the Bank of England, many forecasters and central banks' assessments of these future uncertainties are informed, at least in part, by monitoring past forecast errors.[1] As Reifschneider & Tulip (2019) review, this is the general approach to gauging unconditional forecast uncertainty at the US Federal Reserve, the European Central Bank, the Reserve Bank of Australia, the Bank of Canada and the Swedish Riksbank.

The Bank of England's Monetary Policy Committee (MPC) represents its forecast densities as fan charts, with shades of red (for inflation) and green (for GDP) representing regions with specified probabilities of outturns. Most analysis of these fan charts has drawn on the Bank's specification of the underlying probability distribution, defined by a two-piece normal distribution, and devised performance tests for these density forecasts making the assumption that the density function is described fully. The two-piece normal creates a skewed density by combining the two halves of two normal densities; and confers computational advantages over other skewed distributions as it can be analysed using analytical formula. While this density has been known under different names, and, as explained by Wallis (2014), has been rediscovered several times, it has a long history dating back to Fechner (1897).[2]

In this paper we explore two issues. Throughout our focus is on the Bank of England's density forecasts, but we note the international relevance of our analysis given similarities with practice at other central banks and institutions that produce density forecasts.

First, given increased attention (internationally) to forecast errors in the aftermath of the global financial crisis (e.g. see Alessi, Ghysels, Onorante, Peach & Potter (2014)), we examine whether it might be sensible for the Bank to depart from the assumption of two-piece normality. As Haldane (2012) has noted, theory and recent evidence suggest that macroeconomic data exhibit fat tails as well as skewness; and Adrian, Boyarchenko & Giannone (2019) emphasise non-Gaussian features when measuring the "vulnerability" of GDP growth to downside risks. Accordingly, we consider how more general density functions, that nest the two-piece normal, but allow for fat tails might be used to produce density forecasts.

Secondly, we draw attention to the fact that the Bank does not in fact specify the density function in full, but describes only the ninety per cent "best critical region" (BCR) which characterises the interval of shortest length with a target (nominal) coverage rate of 90%. In effect, the Bank publishes, what we call, "censored" density forecasts that do not take any view on the distribution of the outer tails beyond

[1] Central banks also use model-based approaches and their subjective judgement to gauge future economic uncertainties. But as emphasised by Ericsson (2002) and Knüppel (2014, 2018), quoting Wallis (1989), pp.55-56: "the model-based approach is of little help to the practitioner. It neglects the contribution of the forecaster's subjective adjustments".

[2] As Appendix A.1 explains, two-piece densities should be distinguished from the conceptually similar, but mathematically distinct, skew-normal classes of densities developed by Azzalini (1985). Outside of macroeconomics and central banking applications, the use of non-normal, skewed distributions is also growing in finance; e.g. see de Roon & Karehnke (2017).

saying that it does not overlap with the inner ninety per cent of the distribution. This appears to reflect, at some level, a Knightian distinction between known and unknown probability distributions (between *risk* and *uncertainty*). We examine the consequences of censoring both for estimation of the parameters of the density function of past forecast errors and for *ex post* assessment of forecast performance.

More generally, we propose and explore the properties of a new estimator that fits a potentially skewed and fat tailed density to the inner $100(1-\alpha)\%$ observations acknowledging that the outtermost $100\alpha\%$ observations may be drawn from a different distribution. As the censor points are defined by the bounds of the $100(1-\alpha)\%$ BCR - and therefore not determined exogenously - a fixed point estimator is proposed. The MPC fix 100α=10% and thereby do not quantify forecast uncertainties in the tails, beyond saying that there is, in sum, a 100α=10% chance of observing these more 'extreme' events. We emphasise that under asymmetry this need not imply that $100\alpha/2 = 5\%$ of the probability mass falls in each tail: the BCR need not amount to the region between the 5% and 95% quantiles. While there is no reason to set $\alpha = 10\%$ in other applications, the proposed estimator has relevance when interest resides with the central region of a density - free from or robust to outliers. In forecasting applications, motivation for this may reflect the view that outlying forecast errors reflect (realised) unknown unknowns and/or events not expected to recur, so they should be censored before quantifying expected known unknowns (risk). An *ad hoc* but commonly used alternative is to censor all ('old') observations that fall outside some historical window of the data.

The next section of the paper provides an account of density forecasting by the Bank of England's MPC. Section 3 describes the forecast data from the Bank of England that we use in our application. Section 4 sets out the parametric family of skewed distributions that we consider, and fits these to the MPC forecast errors. Section 5 describes the implications for estimation when the density function is not fully described. A new fixed-point estimator is proposed that fits the density acknowledging endogenous censoring of the outlying $100\alpha\%$ observations. Section 6 then explores the properties of this estimator via three sets of Monte Carlo experiment. Section 7 fits censored densities to the MPC's forecast errors, finding that especially for GDP there is much less evidence for skew when the outlying 10% of forecast errors are censored. Section 8 considers the consequences of censoring for evaluating (out-of-sample) forecast performance: it first sets out tests for the evaluation of censored density forecasts and then uses these to assess the absolute and relative forecast performance of the MPC's time-stamped fan charts. Despite many previous evaluations, this is the first time the MPC's forecast performance has been evaluated explicitly acknowledging the censoring. Section 9 concludes. An online Appendix contains supplementary details, results and robustness checks.

2 Density Forecasting by the Monetary Policy Committee

In 1993, following Britain's departure from the Exchange Rate Mechanism in September 1992 and the adoption of an inflation target, the Bank of England began to publish its *Inflation Report*. This included a forecast for inflation over the next two years. Initially uncertainty surrounding the forecast was represented graphically by a "trumpet" round the inflation projection, the width of which also represented the mean absolute forecast error, with the errors calculated not only from the *Inflation Report* forecasts but from earlier unpublished forecasts. In February 1996 the *Inflation Report* showed for the first time the now familiar fan chart. This presented deciles of subjective estimates of the probability distribution of the Bank's forecast.

The MPC was set up in June 1997 and it, rather than the Bank, has been responsible for the forecast of GDP and inflation since then. The MPC continued to present its forecasts in the format adopted by the Bank in 1996. In these fan charts the deciles of the density function represent nested "best" critical regions (BCRs). That is, they represent the shortest range of possible outcomes which have the required probability.[3] The outermost decile cannot, of course, be represented in this way unless the MPC takes the view that the maximum forecast errors are bounded. The decile bands fit naturally round a central forecast which is defined as the mode of the distribution; and the MPC's discussions appropriately focus on the most likely outcome and describe their forecast as representing the mode. Except when the forecast distribution is unimodal and symmetric, the BCR does not correspond to the central interval. In a decision theory framework, Wallis (1999) and Askanazi et al. (2018) show that the shortest interval is the *best* prediction interval when the loss function takes an all-or-nothing form. This is such that the loss (or cost) of an outturn falling outside the BCR in question is the same irrespective of how far away from the BCR the outturn falls.

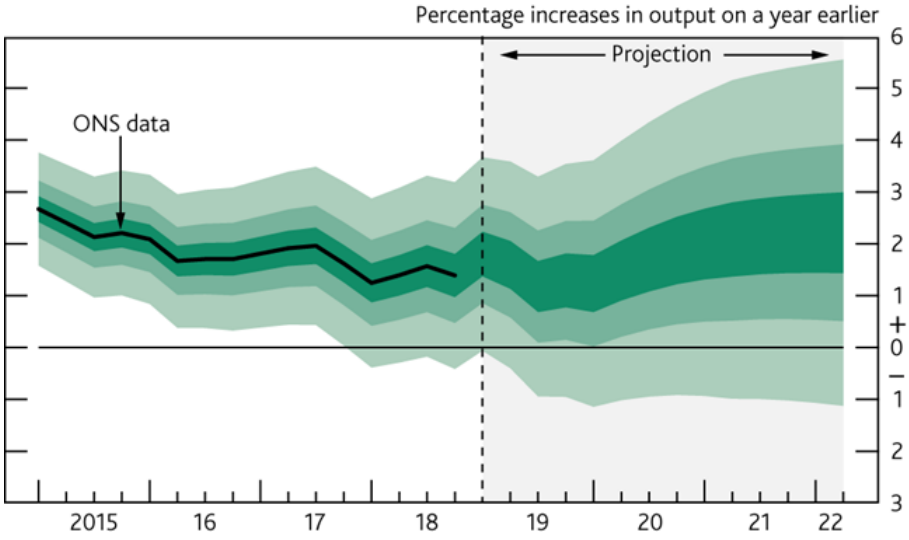
The MPC came to think that presenting the deciles of the density function gave a misleading indication of precision and, following a suggestion of Ken Wallis, reinforced by Stockton (2013), modified the charts in May 2013, to show only three BCRs, associated with thirty per cent, sixty per cent and ninety per cent of the probability mass. Charts in the original format, showing BCR deciles continue, however, to be made available on the Bank's website. In August 2013 the MPC added an additional fan chart showing its forecast for unemployment constructed on the same principles. Figure 1 shows the density forecast for GDP growth produced by the MPC in May 2019 with three best critical regions distinguished.

[3] See Britton, Fisher & Whitley (1998), Wallis (1999) and Askanazi, Diebold, Schorfheide & Shin (2018). To a Bayesian, the $100(1-\alpha)\%$ best critical region/interval for y might be re-interpreted as the highest posterior density (HPD) interval:

$$R_\alpha = \{y : f(y) \geq \pi_\alpha\}, \text{ where}$$
$$\pi_\alpha \text{ is the largest value for which } P(y \in R_\alpha) \geq 1-\alpha$$

A HPD interval has two main properties: (1) the density for every point inside the interval is greater than that for every point outside the interval and (2) for a given probability the interval is of shortest length; e.g. see Chan & Shao (1999) for methods to estimate HPD intervals.

Figure 1: The MPC's fan chart for GDP Growth (*Inflation Report*, May 2019)



Notes: The Monetary Policy Committee's fan chart for GDP growth (from the May 2019 *Inflation Report*). In their notes to this chart the Committee writes: "The fan chart depicts the probability of various outcomes for GDP growth... To the left of the vertical dashed line, the distribution reflects the likelihood of revisions to the data over the past; to the right, it reflects uncertainty over the evolution of GDP growth in the future... If economic circumstances identical to today's were to prevail on 100 occasions, the MPC's best collective judgment is that ... the mature estimate of GDP growth would lie within the darkest central band on only ... 30 of those occasions .... The fan chart is constructed so that outturns are also expected to lie within each pair of the lighter coloured areas on... 30... occasions. In any particular quarter of the forecast period, GDP growth, CPI inflation or the unemployment rate are therefore expected to lie somewhere within the fan on 90 out of 100 occasions. And on the remaining 10 out of 100 occasions they can fall anywhere outside the coloured area of the fan chart."

The fan charts have, since the MPC became responsible for them, always been described as representing the best collective judgment of the committee and as such the underlying probability distributions could take any form that the MPC judged to be appropriate.[4] In fact, however, the probability distributions have been almost invariably drawn using a two-piece normal distribution (Fechner 1897). Since the location parameter of this distribution is the mode, that in turn makes the MPC's focus on the mode coherent with the way in which it sees the forecast distribution.

The MPC has always claimed that its judgment of uncertainty is subjective. Given this there is, of course, no particular reason why it need use the two-piece normal distribution to represent that judgment. On the other hand, for any committee to complete its deliberations in a reasonable time, it is helpful to restrict the number of variables up for discussion. The limitations implied by the two-piece normal distribution have invariably proven to be helpful in this respect and allowed the MPC to devote much less time to discussing the density function than it does to the modal forecast. More generally, the MPC has stated that the parameters of the distribution of past forecast errors are helpful as a means of informing its choices. But if the MPC wishes its choices to be informed by past forecast errors, how best should we estimate the parameters of the distribution of past forecast errors? This is the issue addressed in this

[4] See http://www.bankofengland.co.uk/archive/Documents/historicpubs/qb/1998/qb980101.pdf

paper. As we discuss, the answer very much depends on the interpretation that the MPC puts on the fan charts.

Britton et al. (1998)[5] give no reason to doubt that the MPC, having adopted the two-piece normal distribution, had in mind any other distribution for the outermost ten per cent not covered by the fan chart. In May 2010, however, the *Inflation Report* stated, for the first time, that "on the remaining 10 out of 100 occasions GDP growth/inflation can fall anywhere outside the green/red area of the fan chart". That might be seen as no more than a statement of the obvious fact that the two-piece normal distribution is unbounded, although a normal distribution would not usually be described in such terms. In fact the MPC was not expressing any view on the density of points in the outlying ten per cent of the distribution.[6] This was made clear by the discussion of the Greek crisis in the first half of 2015. In February the MPC could see the risk that the problems faced by Greece might develop into a major financial crisis with obvious downside risks for growth in the UK. It took the view that the chance of this happening was less than ten per cent and thus felt perfectly comfortable in producing a symmetric fan chart. During the Spring, however, the crisis intensified and the MPC judged that the risk of a disorderly outcome was now higher than ten per cent, with the implication that it needed to show a degree of downside risk in the fan charts published in the May forecast. In the *Inflation Report* of May 2015 (p.48) the MPC observed, in its description of the fan chart for GDP growth, "that the risks around the central projection are skewed slightly to the downside for much of the forecast period, reflecting the possibility of a disorderly outcome to the current Greek negotiations, rather than balanced as in February." The downside risk was represented through the medium of the two-piece normal distribution. This can be seen by looking at Figure 2.

The top panel of Figure 2 shows the lower and upper 10% censoring points, y_L and y_U, together with the modal forecast μ and (subsequent) outturns for the MPC's two-year ahead GDP growth forecasts. Figure 3 presents estimates for their two-year ahead inflation forecasts. The Bank publishes the parameters of the two-piece normal, as shown graphically in the *Inflation Report* fan chart, each quarter on its website; but it does not publish y_U and y_L and so they have to be backed out by calculating the BCRs of the two-piece normal. For both GDP growth and inflation, Figures 2 and 3 show that the width of these *censoring* bands, y_U - y_L, has increased over time, implying the MPC perceived "risk" to be increasing. In particular, there is a marked and fairly rapid increase in the width of the bands in 2008 and 2009, with the apparent end of the "Great Moderation" as the Global Financial Crisis and inflation (oil price) shocks hit the UK economy. Note that the dates in the x-axes to Figures 2 and 3 indicate when the forecasts were made.

[5] See also http://www.bankofengland.co.uk/publications/Documents/inflationreport/ir02mayfanbox.pdf

[6] There was also, between 2010 and 2016, some discussion on the MPC about whether, even within the central ninety per cent of the distribution, it was expressing a precise view on the density or simply giving BCR decile ranges. But this was not material to its published density forecast and no conclusion was reached.

It is a matter of interpretation, formalised in the distinction between L_A^C and L_B^C in Section 5 below, but if the MPC took (or is assumed to have taken) a view on whether the unknown tail uncertainties (summing to 10%) were in the left or right hand tail, we can also compute the probabilities that $Y > y_U$ and $Y < y_L$. Recall for asymmetric densities, given y_U and y_L are defined as BCRs, these probabilities need not equal 5%.[7]

The middle panel of Figures 2 and 3 shows a measure of skew. The MPC represents the skew as the difference between the mean and the mode of the distribution, But it acknowledges that, in order to calculate the mean, it is necessary to make some assumption about the distribution of the censored part of the distribution; it assumes that this follows the same two-part normal distribution as the uncensored part, despite having no reason to believe this to be the case. Similarly, traditional measures of skewness rely on full knowledge of the relevant distribution. We therefore prefer to adapt instead the measure proposed by Arnold & Groeneveld (1995). They suggest as a measure of skewness:

$$1 - 2F(mode) \tag{1}$$

where F is the cumulative density function. With this defined only over the 90% best critical region we adapt their measure to provide a measure of skew in the presence of censoring:

$$Robust\ skew = 1 - 2F^*(mode)/0.9 \tag{2}$$

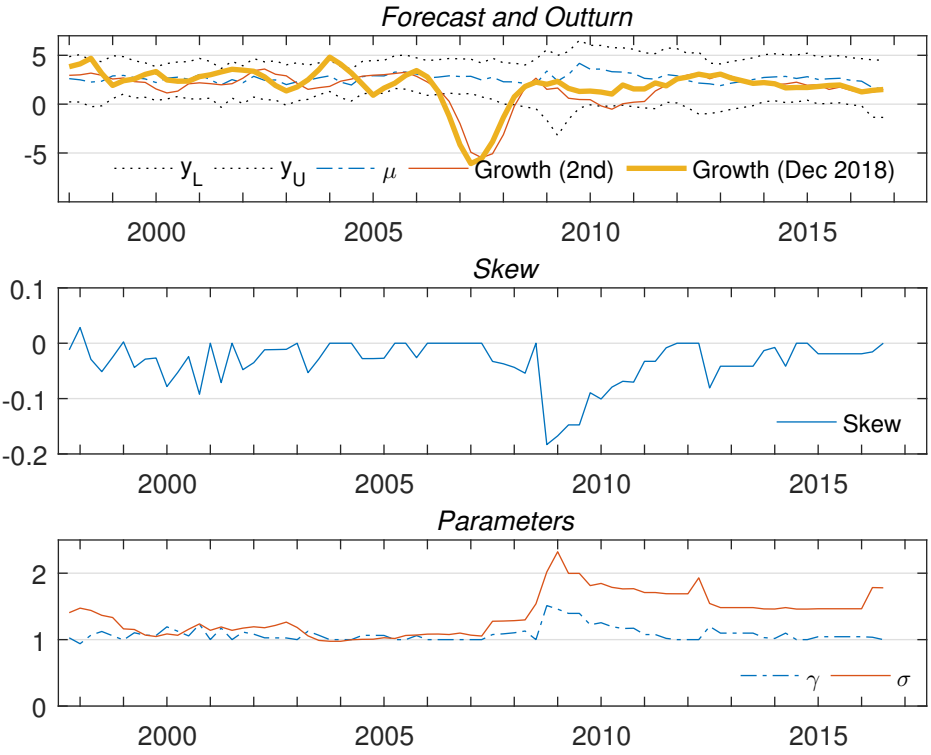
where F^* is the probability mass lying between the lower censor point and the mode. Looking at Figure 2, we can see that it was rare for the MPC to see an upside skew for GDP growth. It was most concerned about downside skew during the recession of 2008/9 and its aftermath. While, as noted above, a downside skew was introduced during the Greek crisis of 2015, Figure 2 shows that this was in fact quite modest.

In the third panel of each figure we show the scale and (traditional) skew parameters of the two-part (normal) distribution. While the MPC specifies the two scale parameters for the parts of the distribution on each side of the mode, we prefer to adopt a single scale parameter σ which would be equal to the standard deviation in the absence of skew, and a skew parameter γ. The scale parameter takes a value of $\sigma\gamma$ to the left of the mode and σ/γ to the right of the mode. This notation follows that adopted by authors such as Fernandez & Steel (1998), whose specification we use subsequently.

As noted in the introduction, we subsequently consider the use of more general skewed densities as an alternative to the two-piece normal distribution; and we explore the issues created by trying to fit only the central ninety per cent of a density function. But, before turning to that, we consider the MPC forecast error data in more detail.

[7] Figures A8 and A9 in the online Appendix shows that these probabilities do indeed vary over time and often differ from 5%. The strongest departures from equal (5%) probabilities occur for GDP growth in the aftermath of the financial crisis. In 2009, for example, the MPC gave close to a 7% probability that GDP, two years ahead, would fall below y_L.

Figure 2: Properties of the MPC's GDP growth forecasts: 10% BCR censoring thresholds



Note: μ=modal forecast; y_L=lower censor point; y_U=upper censor limit. Because of GDP data revisions, two estimates of the GDP outturn are considered: *Growth (2nd)* = the ONS's second GDP Growth Estimate; *Growth (Dec 2018)* = the ONS's estimate of GDP Growth using December 2018 vintage data; Skew (middle panel) = robust measure of skew as in (2); σ=scale parameter; γ=skew parameter (based on MPC estimate); dates relate to when the forecast was made.

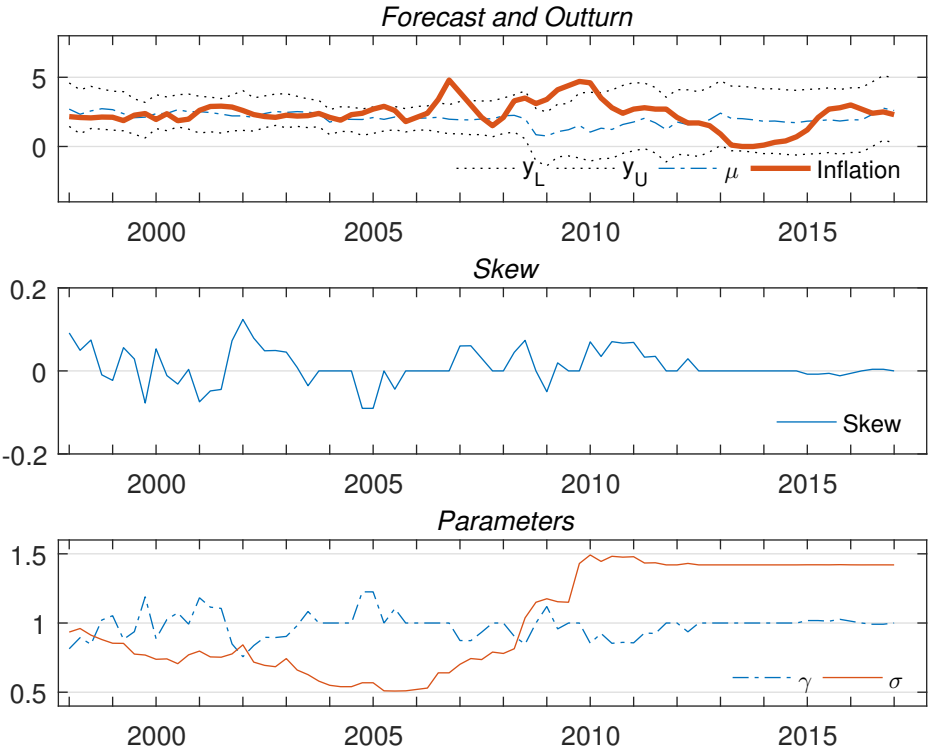
3 Forecast Error Data

As noted above, the MPC publishes density forecasts for economic growth, inflation and unemployment. The unemployment forecast has been published only since August 2013, not giving enough data to come to any informed view about forecast errors. We therefore focus our attention on the forecasts for inflation and output. The MPC publishes its forecasts quarterly up to three years ahead; with rates of change defined over one year. We focus our attention on forecasts at the two-year horizon because - not least for an inflation targeting central bank - these are probably of greatest interest, although it would be perfectly possible to apply our analysis to forecast errors at any other horizon.[8] We focus on forecasting errors relative to the modal forecast provided by the MPC.

The MPC's first inflation forecast was published in August 1997. We use, however, only the forecasts from February 1998 onwards, because the earlier forecasts were conditional on constant rather than

[8] An interesting extension would be to exploit cross-horizon dependencies between the forecast errors when estimating the forecast error densities. Knüppel (2014, 2018) proposes a pooled (across forecast horizons) estimator of the sample mean of the squared forecast errors.

Figure 3: Properties of the MPC's inflation forecasts: 10% BCR censoring thresholds



Note: μ=modal forecast; y_L=lower censor point; y_U=upper censor limit; Skew (middle panel) = robust measure of skew as in (2); σ=scale parameter; γ=skew parameter (based on MPC estimate); dates relate to when the forecast was made.

market interest rates. From February 1998 to November 2003 these are forecasts of the Retail Price Index excluding mortgage payments (RPIX) inflation, the target variable (set at $2\frac{1}{2}$ per cent per annum) for the MPC at the time. Thereafter, from February 2004, with the change in December 2003 to the targeted measure of inflation to Consumer Price Index (with the target set at 2 per cent), these are forecasts of CPI inflation. The (eight quarter ahead) forecasts are matched against the subsequent outturns for annual RPIX inflation (ONS code: CHMK) from 1999q4-2005q3 and CPI inflation (ONS code: D7G7) from 2005q4-2018q4 to make a time-series of 77 observations.[9]

When calculating forecast errors for annual GDP growth, given that GDP data are revised, we need to take a view on what vintage of GDP data to use. Since the MPC set out to forecast the "mature" values of GDP, as noted in footnotes to their fan charts, we use the latest available data vintage (from the quarterly national accounts published in December 2018) to define the outturn. This then delivers a series of GDP errors from 1999q4-2018q3, a time-series of 76 observations. However, in the out-of-sample analysis below, we do consider GDP errors defined against the third data release (which the ONS call

[9]While the wedge between RPIX and CPI has tended to increase, there is nothing in the data to suggest that it is wrong to treat the errors to the inflation forecasts as a single series.

the Second Quarterly National Account estimate, given that for most of our sample period the first estimate is called the "preliminary" estimate) and find that results are sensitive to the vintage of GDP data chosen.[10] In the analysis below, we define the forecast error as the outturn minus the forecast, so that negative errors are outturns below forecast. When referring below (in the figures) to the date of the forecast error, we continue to refer to the time of the forecast; given our focus on two-year ahead forecasts, this means the outturn itself was observed two years after the date indicated. The use of the second quarterly national account estimate as a benchmark does have the benefit that these results are not affected by forecast "errors" arising from changes to the definition of GDP.

4 The Distribution of MPC Forecast Errors

4.1 A Parametric Framework

In order to explore the suitability of the two-piece normal distribution, and model skewness and kurtosis in a flexible but practical way, we consider the general family of skew distributions defined in Arellano-Valle, Gomez & Quintana (2005). Like the two-piece normal these involve joining two distributions, with different scale (and perhaps shape) parameters. We defer the issue that the MPC has no views on the outer ten per cent of the distribution until Section 5.

A leading specific density within this family, that we focus on, is the two-piece t distribution described by Fernandez & Steel (1998). This depends on, in addition to the location, scale and skew parameters, the number of degrees of freedom of the t- distribution. In Appendix A.1 (for robustness) we explore more general and alternative skewed specifications. In general, we find that (in-sample) the two-piece t fits our data competitively relative to these alternatives. We therefore confine our attention to it (and its limiting case, the two part normal distribution) here. However, we first note that albeit at the expense of introducing extra parameters which may complicate estimation, especially for our relatively small sample, the ensuing discussion on the estimation of censored densities is general. It extends to skewed specifications beyond the two part t and normal. Secondly, in the policy making environment of the MPC, we suspect that in practice the introduction of not just a fourth (relative to the two-piece normal) but a fifth extra parameter would impede discussions of the economic interpretation of the parameters of the densities. Thirdly, from an estimation perspective, with a limited number of observations it is often helpful to limit the number of parameters to be estimated.

The density function of the two-piece t distribution is given as follows:

$$f(y_t) = \frac{2}{\sigma\left(\gamma + 1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\left(\pi\nu\right)^{1/2}} \left[1 + \frac{(y_t - \mu)^2}{\gamma^2 \nu \sigma^2}\right]^{-(\nu+1)/2} \quad \text{if } y_t < \mu \tag{3}$$

[10] The ONS changed its publication model and release calendars in the summer of 2018. We continue to use the quarterly national account estimate, even though, with the preliminary estimate being discontinued, this is now the second rather than third estimate.

$$f(y_t) = \frac{2}{\sigma\left(\gamma + 1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \left[1 + \frac{\gamma^2(y_t-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \quad \text{if } y_t \geq \mu.$$

where γ is the scalar skew parameter, $\gamma \in (0, \infty)$, ν is the degrees of freedom of the standard Student t distribution with location, μ, and scale, σ, and $\Gamma(.)$ is the gamma function.

The mode of the distribution is μ but this is the same as the mean only if $\gamma = 1$. The probability mass to the left of the mode is $\gamma^2/(\gamma^2+1)$ while that to the right of the mode is $1/(\gamma^2+1)$. So with $\gamma < 1$ the distribution is skewed to the right and with $\gamma > 1$ it is skewed to the left. A large number of degrees of freedom, ν, implies, of course, that the distribution is very close to normal; while a small number of degrees of freedom indicates that extreme values are appreciably more common than would be implied by a normal distribution with the same scale parameter.[11]

Given a scoring rule or loss function the parameters of this distribution can be estimated. Following Gneiting & Raftery (2007), optimum score estimators or M-estimators involve maximising the value of the (proper) scoring rule over the sample. We focus on the logarithmic scoring rule corresponding to maximum likelihood (ML) estimation.[12]

The log-likelihood function of a sequence of observations y_t, $t = 1, ..., T$, is, with $\mathbf{I}(y)$ an indicator function, $\mathbf{I}(y) = 1$ if $y \geq 0$ and $\mathbf{I}(y) = 0$ if $y < 0$, given as:

$$\begin{aligned} \log L = T \ln\left(\frac{2}{\sigma\left(\gamma + 1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}}\right) + \sum_{t=1}^{T} \mathbf{I}\left(y_t - \mu\right) \ln\left[1 + \frac{\gamma^2(y_t-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \\ + \sum_{t=1}^{T} \mathbf{I}\left(\mu - y_t\right) \ln\left[1 + \frac{(y_t-\mu)^2}{\gamma^2\nu\sigma^2}\right]^{-(\nu+1)/2}. \end{aligned} \tag{4}$$

For any given sample, the four parameters, μ, σ, γ and ν can be estimated by ML. Our sample suffers from the drawback that the forecast errors relate to overlapping periods. Nevertheless, parameter estimation by ML delivers consistent estimates of the four parameter values (White 1980).

4.2 Application to MPC Forecast Errors

We show in Figure 4 the histogram of the errors associated with the forecasts of inflation and, in Figure 5, the corresponding histogram for the forecast errors of GDP; recall that negative numbers indicate an outturn below forecast. For GDP growth we focus on forecast errors defined against the latest outturns - given the MPC's stated target of forecasting "mature" values of GDP. But for completeness Appendix A.2.1 contains an analogous figure using ONS "second" release GDP data to define the outturns.

Figures 4 and 5 are full-sample histograms and involve using all of the available forecast error data, discussed in Section 3. There are always questions about the appropriate sample period, or window

[11] In this specification the scale parameter to the left of the mode is $\sigma\gamma$ while to the right of the mode it is σ/γ. In the specification used by the MPC it is $\left(\frac{\sigma^2}{1-\phi}\right)^{1/2}$ to the left of the mode and $\left(\frac{\sigma^2}{1+\phi}\right)^{1/2}$ to the right of the mode. So it is easy to express γ in terms of ϕ and *vice versa*.

[12] The logarithmic score is known to be more sensitive to outliers than alternatives such as the Cumulative Ranked Probability Score (CRPS). Future work might consider estimators that minimise CRPS loss along the lines of Gebetsberger, Messner, Mayr & Zeileis (2018).

of data, over which to estimate forecasting models and evaluate forecast accuracy. In the presence of parameter instability, due to structural breaks, there is a trade-off between bias and forecast error variance when selecting the 'optimal' window of data to use to estimate the parameters of the model or density function (e.g. see Pesaran & Timmermann (2007)).

Here, given our relatively small samples of forecast error data, we elect to use as much data as possible when estimating the unconditional densities of the forecast errors. Importantly the point of departure for our censored estimator, introduced below, is that it lets the whole sample (of length T) determine which specific observations within the sample to censor - for a specified coverage rate, set at 90 percent by the MPC. This contrasts our understanding of practice at the Bank of England. Elder, Kapetanios, Taylor & Yates (2005) state that the MPC use, as a starting point, a rolling ten-year window to inform their judgement of uncertainty; with in more recent years (post financial crisis) shorter windows used, in effect, to censor forecast error observations not believed to be representative of current uncertainties. As well as ignoring all 'old' data (certainly more than 40 quarters old) irrespective of their properties, this practice also ignores the censoring that is later imposed when the MPC publish the fan chart only for the central 90% of observations.

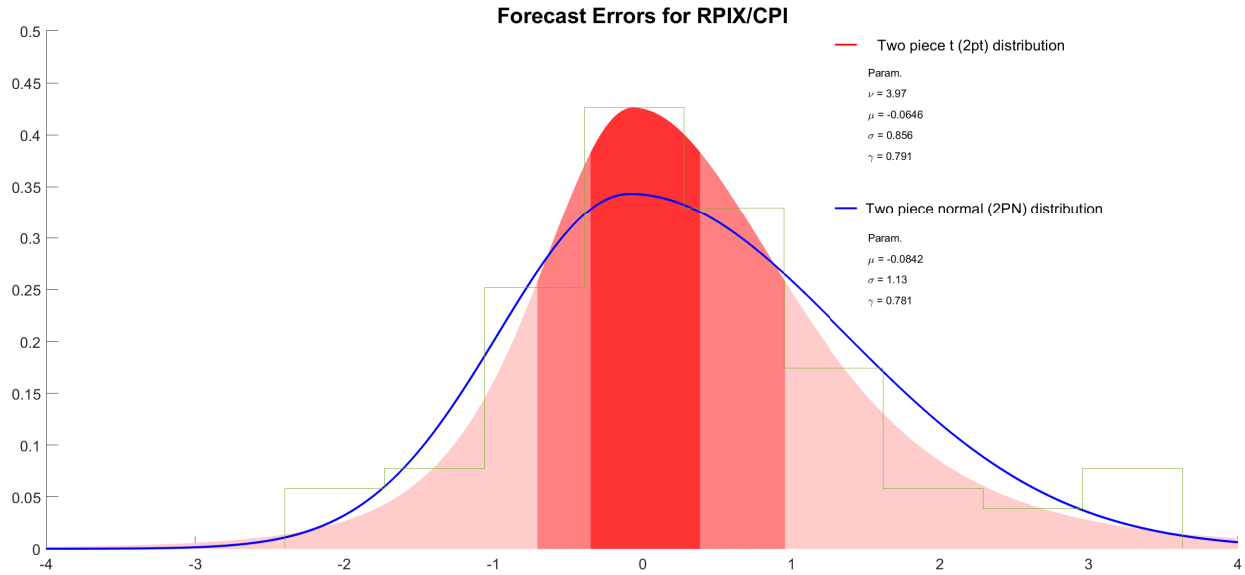
For both inflation and GDP growth in Figures 4 and 5 we also show the estimated ML parameters of the two-piece t and two-piece normal distributions fitted to the underlying forecast errors.[13] The darkest shaded region on each figure indicates the thirty per cent best critical region for the distribution. The band round this extends the best critical region to sixty per cent, and the palest band takes it to ninety per cent. Thus these bands correspond to what the MPC would display if it used the two-piece t distribution to represent forecast uncertainty, and based its judgments (entirely) on past forecast errors. We also show on the charts the density functions estimated by fitting two-piece normal distributions to the same observations.

While the inflation forecast errors are skewed to the right, those for GDP are even more clearly skewed but to the left. This skew is especially pronounced when the two-piece normal is fitted to the GDP forecast errors. Figure 5 shows forecast errors of up to (minus) eight per cent for GDP growth; these arose from a failure to forecast the recession of 2008/9. If one thinks that the underlying frequency of recessions is less than one in seventeen years (the length of our sample of data, in years), then it is possible that this figure overstates the true skew to the distribution of forecast errors. On the other hand, there have been six major recessions in the UK since the end of the First World War,[14] suggesting that an average frequency of about once in seventeen years (6%) is reasonable, and in that sense our data do not overstate the risk of forecasts being disrupted by recession.

[13] Estimation was performed in Matlab. Results were also cross-checked and verified with those from R using the sn and twopiece packages.

[14] There was, in addition, a two-quarter recession at the time of the General Strike in 1926, and a fall in output from 1944 to 1946 as the war effort was wound down. But these episodes are not generally treated as business cycle recessions.

Figure 4: Inflation: Forecast Error Histogram and Two-Piece Normal and t Densities

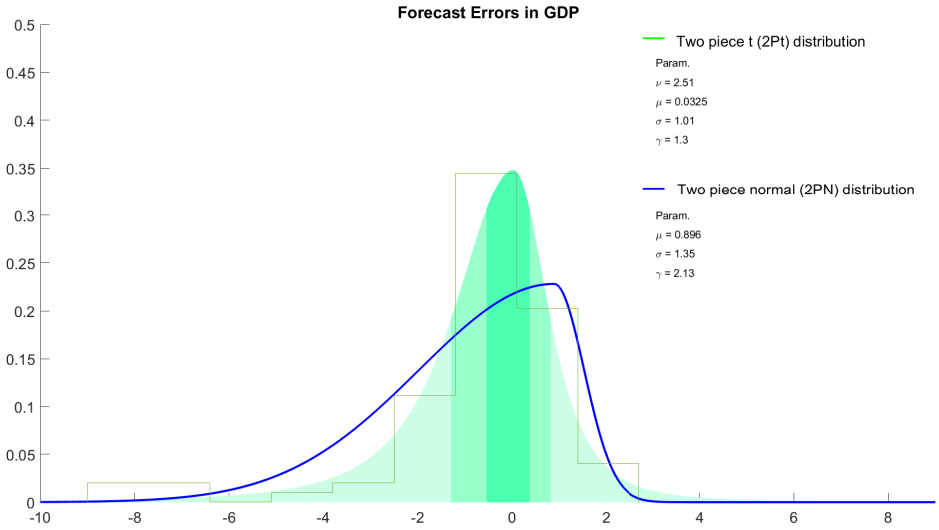


Note: RPIX forecasts (until Nov. 2003); CPI forecasts thereafter. 77 observations used. The p-values from tests for the uniformity of the probability integral transforms (for the 77 forecast error observations, as seen in the histogram, evaluated against the CDF of the fitted density), using an Anderson-Darling test are: 0.52 (2Pt), 0.02 (2PN), 0.17 (for a one-piece t) and 0.00 (for a one-piece normal).

In both cases the number of degrees of freedom fitted to the distributions is low, at 3.97 for inflation and 2.51 for GDP. It can be seen in both figures (including from inspection of the statistical tests referred to in the note to each figure), and particularly for GDP, that the two-piece t distribution does a better job of fitting the histogram density in the centre of the distribution than the two-piece normal distribution. This is confirmed by supplementary statistical evidence in Appendix A.1. While the t distribution is often described as having fat tails, the counterpart of this is a concentration of probability mass in the centre of the distribution. The problem with the two-piece normal distribution, once fitted to the sort of data we have here, is not so much that it means the probability of extreme events is understated. Rather it is that it understates the concentration of mass in the centre of the the distribution. It is also interesting that the two-piece t suggests less forecast error bias (a lower value for μ) than the two-piece normal density. For GDP this fall in bias is quite marked, with μ dropping from 0.90% to 0.03% when using the two-piece t rather than the two-piece normal density.

Despite their greater flexibility, the t distributions appear to have trouble in accommodating the extremes of the histograms. If, instead of being fitted to the whole distribution, they were fitted only to the central part, one might expect to see less skew, and perhaps a higher number of degrees of freedom. Thus the MPC might nevertheless be justified in assuming normality because it takes no view on the distribution of forecast errors outside the ninety per cent BCR. We explore this next by fitting censored

Figure 5: GDP Growth: Forecast Error Histogram and Two-Piece Normal and t Densities



Note: Latest release GDP estimates used to define the 'outturn'. 76 observations used. The p-values from tests for the uniformity of the probability integral transforms (for the 76 forecast error observations, as seen in the histogram, evaluated against the CDF of the fitted density), using an Anderson-Darling test are: 0.35 (2Pt), 0.00 (2PN), 0.29 (for a one-piece t) and 0.00 (for a one-piece normal).

two-piece t and normal distributions to the inflation and GDP forecast errors.

5 Fitting Censored Distributions

The principles of fitting censored distributions when the censor points are given exogenously are understood well. Typically it is clear whether observations are censored or not, but not where they lie in the censored region. In that situation, Diks, Panchenko & van Dijk (2011) have shown that in computing the likelihood function the censored observations are given a likelihood equal to the chance of being in the censored region, conditional of course on the parameters of the distribution. This yields ML estimates of the parameters, with standard properties.

The situation we face is different in two respects. First of all, while we have observations outside the 90% BCR we do not wish their position to have any influence on the estimated parameters of the distribution. This can be achieved if they are treated as though they are censored with a likelihood defined by the probability of being in the censored region. Thus, conditional on known censor points, this difference is not material for estimation.

The second difference is, however, very material. In the situation we face the censor points are defined by the bounds of the 90% BCR and thus by the parameter estimates. In such a situation the "regularity conditions" needed to prove, in particular, asymptotic normality of ML estimators are well known to be violated because the support of the density depends on its parameters, as in (5); e.g., see Woodroofe

(1972) and Smith (1985). We show that, in our case, the estimator degenerates in finite samples, and develop an alternative fixed point estimator whose properties we examine by means of Monte Carlo simulations.

5.1 Motivating a fixed point estimator

If the lower cut point, beyond which data are censored, is y_L and the upper cut point, above which data are censored, is y_U, then the conventional way of setting out the censored log likelihood is:

$$\log L_A^C = \left\{ \begin{array}{ll} \log(F(y_L)) & \text{if } (y < y_L) \\ \log L & \text{if } (y_L \leq y \leq y_U) \\ \log(1 - F(y_U)) & \text{if } (y > y_U) \end{array} \right\}, \tag{5}$$

where $F(y)$ defines the CDF of the density function, $F(y) = \int_{-\infty}^{y} f(y)dy$, and the BCR, set in our application to define a 100α=10% censored region, satisfies:

$$f(y_U, \beta) - f(y_L, \beta) = 0 \tag{6}$$

$$F(y_U, \beta) - F(y_L, \beta) = 1 - \alpha. \tag{7}$$

This likelihood function, however, still assumes that the MPC has a view on whether points are likely to be in the upper or the lower tail of the distribution, notwithstanding that the density function within those tails is not specified. It is hard to say how far that represents the MPC's views; certainly the issue was not discussed between 2010 and 2016.

An alternative likelihood function which is completely agnostic as to whether observations are going to be above the upper cut point or below the lower cut point can be defined, with conditions (6) and (7) again imposed, as:

$$\log L_B^C = \left\{ \begin{array}{ll} \log(F(y_L) + 1 - F(y_U)) & \text{if } (y < y_L) \\ \log L & \text{if } (y_L \leq y \leq y_U) \\ \log(F(y_L) + 1 - F(y_U)) & \text{if } (y > y_U) \end{array} \right\}. \tag{8}$$

While it might be desirable to compare the outcomes of using L_B^C rather than L_A^C, the reality of estimation with small samples is that we need to take advantage of the greater structure provided by L_A^C in order to be able to estimate the parameters of the distribution recursively. We subsequently show the importance of the distinction between L_A^C and L_B^C; it proves to be very material when fitting distributions to GDP forecast errors.

In either case, estimation of $\beta = [\mu, \sigma, \gamma, \nu]$, subject to (6) and (7), becomes more difficult, indeed potentially degenerate. This is because the censor points are treated as endogenous (or proportionate), rather than fixed (assumed known) due to discontinuities (boundary problems) as movements in the BCR cut points place observations either in the censored or the uncensored region.

Intuitively, for fixed (finite) T, we explain the degeneracy of full ML estimation of β, y_L and y_U as

follows. We illustrate for L_A^C although the same point is pertinent for L_B^C. Consider:

$$\max_{\beta, y_L, y_U} \sum \log L_A^C(y_t, \beta) + \lambda_1 \left(F(y_U, \beta) - F(y_L, \beta) - (1-\alpha)\right) \\ + \lambda_2 \left(f(y_U, \beta) - f(y_L, \beta)\right). \tag{9}$$

Suppose that we have a value of σ sufficiently small such that only one observation from a sample, say y_A, is in the uncensored region and that this is the value given to the mode of the distribution, μ: $\mu = y_A$. All other observations are then in the censored region - with, say, T_1 observations below y_L and T_2 observations above y_U. Then, when the constraints are met:

$$\log L^C = T_1 \log F(y_L, \beta) + \log f(y_A, \beta) + T_2 \log(1 - F(y_U, \beta)) \tag{10}$$

But as σ shrinks, for fixed T, $\log f(y_A, \beta)$ will increase without limit:

$$\log L^C \to \infty \text{ as } \sigma \to 0. \tag{11}$$

In the absence of censoring (or as $T \to \infty$), this would be offset by the likelihood associated with the other observations falling. But with the censored likelihood, for fixed T, that is not the case. In other words, the censor points y_L and y_U change as σ shrinks, but the probability of being in the censored tails, and thus $F(y_U, \beta)$ and $F(y_L, \beta)$, will not change. For fixed T the overall log likelihood, $\log L^C$, is therefore unbounded as σ shrinks to zero; there is no interior solution.

Accordingly, we suggest the following fixed-point estimator in finite-samples. It is motivated by the observation that, in large samples, estimates (for β) produced by maximising $\log L^C$, with *fixed censor points*, are independent of the censor points, provided all the uncensored observations are genuinely drawn from the specified distribution.

The proposed fixed point estimator is calculated by means of the following two steps:

$$\text{Step 1: } \beta_{r+1} = \max_{\beta} \sum \log L_j^C(y_t, \beta, y_{L,r}, y_{U,r}) \tag{12}$$

$$\text{Step 2: compute BCR of } f(y_t \mid \beta_{r+1}) \Rightarrow y_{L,r+1}, y_{U,r+1} \tag{13}$$

where we search over values of $y_{L,r}$ and $y_{U,r}$ $(r = 1, ..., R^*)$ to minimise $P_{r+1} = (y_{L,r+1} - y_{L,r})^2 + (y_{U,r+1} - y_{U,r})^2$. If $P_{r+1} = (y_{L,r+1} - y_{L,r})^2 + (y_{U,r+1} - y_{U,r})^2$ converges to zero as R^* increases, this provides a solution at which the ML estimates of the parameters of the censored distribution deliver censor points which, when used in estimation, deliver the same parameter estimates.

The contribution of each observation to the log likelihood depends on whether it is in the censored region or not. The log likelihood will not be continuous in the parameters because, for some parameter sets, observations may be uncensored while for others they will be censored. In large samples this effect is likely to be small; the contribution of each observation to the total log likelihood is low. But in small samples the discontinuities will be relatively greater and it may not be possible to find a solution for

which the quadratic term converges to zero. If the minimum $P_r = (y_{L,r+1} - y_{L,r})^2 + (y_{U,r+1} - y_{U,r})^2$ is larger than zero, only an approximation will have been found. It has to be a matter of judgement as to how good or bad that approximation is.

In practice in our experiments we found that, especially in moderate samples, maximisation of L or L_j^C (for fixed censor points) could prove problematic for some samples: the ML estimates of γ can diverge. Similar findings are reported by Sartori (2006) and Azzalini & Arellano-Valle (2013) for their skew normal and t densities (considered in more detail in Appendix A.1). This is because the likelihood can be monotone and the Fisher information matrix singular at the discontinuity point when skewness disappears, $\gamma = 1$. Accordingly, in the out-of-sample application below where sample sizes are smaller, in the spirit of Sartori (2006) and Azzalini & Arellano-Valle (2013), to avoid boundary estimates we maximise a penalised log-likelihood function, $PL_j^C(y_t, \beta)$, rather than L_j^C, where

$$PL_j^C(y_t, \beta) = \sum \log L_j^C(y_t, \beta) - \frac{1}{2} P_\lambda(|(\gamma - 1)|) \tag{14}$$

and $P_\lambda(|(\gamma - 1)|)$ is a nonnegative penalty function. We use the Lasso penalty, $P_\lambda(|(\gamma - 1)|) = \lambda\,|(\gamma - 1)|$ where λ is a tuning parameter. When $\lambda = 0$ estimation reduces to $L_j^C(y_t, \beta)$; and the higher the value of λ the more deviations from symmetry are penalised. We select λ by optimising the in-sample censored fit. We note that there is a connection between use of $PL_j^C(y_t, \beta)$ and Bayesian *a posteriori* estimates with a Laplace prior on $(\gamma - 1)$.

6 Monte Carlo Experiments

We carry out three sets of Monte Carlo experiment to assess the performance of the proposed fixed point estimator, (12)-(13). We also make comparison with the penalised estimator, (14). We focus on censoring at 100α=10%.

6.1 Experiment 1: Performance for different sample sizes

The first set of simulations test the performance of the censored estimator, under both L_A^C and L_B^C, in samples of different sizes as the degree of skew varies. Comparison is made with the uncensored ML estimator, L. T observations are drawn from a two piece t distribution, where $(\nu, \mu, \sigma, \gamma) = (5, 0, 1, 1.5)$ and $(5, 0, 1, 2.5)$. $\gamma = 1.5$ and $\gamma = 2.5$ correspond to moderate and high (positive) skew. We consider $T = 40, 100, 500, 1000$, noting that the Bank of England's use of just 40 observations to estimate their forecast error densities. We report results based on $R = 1000$ replications. For computational reasons we work with $1/v = 0.2$ rather than ν itself. For $\gamma = 2.5$, we also report results for $PL_j^C(y_t, \beta)$. We do not report (in part for space reasons) results for $PL_j^C(y_t, \beta)$ when $\gamma = 1.5$ since, as will be seen, the utility of the penalised estimators, relative to the unpenalised ones, is found to be greater in populations with high skew.

Table 1: Monte Carlo Experiment 1: Mean, median and standard deviation (across replications) of the parameter estimates from the censored estimators L_A^C, L_B^C, PL_A^C and PL_B^C and the uncensored ML estimator, L; and Proportion of observations placed in the censored region

		T=40					**T=100**					**T=500**					**T=1000**				
		$1/v$	μ	σ	γ		$1/v$	μ	σ	γ		$1/v$	μ	σ	γ		$1/v$	μ	σ	γ	
true		**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**
L_A^C	**mean**	0.21	0.01	0.90	2E+4	0.10	0.19	0.02	0.98	1.58	0.10	0.20	0.00	1.00	1.51	0.10	0.20	0.00	1.00	1.50	0.10
	med	0.15	0.03	0.91	1.55	0.10	0.18	0.02	0.98	1.52	0.10	0.20	0.00	1.00	1.51	0.10	0.20	0.00	1.00	1.50	0.10
	sd	0.23	0.52	0.31	5E+5	0.02	0.16	0.29	0.16	0.35	0.01	0.08	0.11	0.07	0.11	0.00	0.06	0.08	0.05	0.08	0.00
L_B^C	**mean**	0.19	0.01	0.63	7E+4	0.10	0.17	0.08	0.84	63.39	0.10	0.20	0.01	0.98	1.56	0.10	0.20	0.00	0.99	1.52	0.10
	med	0.00	0.12	0.70	1.81	0.10	0.12	0.07	0.91	1.62	0.10	0.20	0.01	0.98	1.53	0.10	0.20	0.00	0.99	1.50	0.10
	sd	0.26	0.79	0.72	2E+6	0.06	0.18	0.51	0.34	834	0.05	0.08	0.18	0.10	0.27	0.00	0.06	0.12	0.07	0.17	0.00
L	**mean**	0.16	0.03	0.95	3.09		0.17	0.02	1.00	1.58		0.20	0.01	1.00	1.51		0.20	0.00	1.00	1.50	
	med	0.13	0.07	0.97	1.54		0.17	0.03	0.99	1.52		0.20	0.01	1.00	1.51		0.20	0.00	1.00	1.50	
	sd	0.16	0.45	0.27	9.90		0.11	0.27	0.15	0.35		0.05	0.11	0.06	0.11		0.03	0.08	0.04	0.07	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
L_A^C	**mean**	0.20	0.02	0.71	1E+5	0.09	0.19	0.05	0.88	126.3	0.10	0.19	0.01	0.99	2.84	0.10	0.20	0.00	1.00	2.52	0.10
	med	0.09	0.10	0.78	2.83	0.10	0.17	0.06	0.92	2.65	0.10	0.19	0.00	0.99	2.51	0.10	0.20	0.00	1.00	2.50	0.10
	sd	0.24	0.53	0.52	2E+6	0.05	0.17	0.32	0.34	858.9	0.02	0.08	0.14	0.12	6.98	0.01	0.06	0.09	0.07	0.19	0.00
L_B^C	**mean**	0.16	0.05	0.41	2E+5	0.12	0.16	0.12	0.58	3E+3	0.11	0.19	0.03	0.91	33.9	0.10	0.19	0.01	0.98	5.08	0.10
	med	0.00	0.18	0.04	61.04	0.10	0.00	0.21	0.57	4.75	0.10	0.19	0.02	0.98	2.56	0.10	0.20	0.01	0.99	2.51	0.10
	sd	0.25	1.27	0.53	4E+6	0.17	0.19	0.46	0.52	6E+4	0.07	0.10	0.23	0.31	225	0.01	0.07	0.14	0.16	38.1	0.00
L	**mean**	0.15	0.06	0.79	16.31	0.00	0.17	0.04	0.92	5.24		0.19	0.00	0.99	2.55		0.20	0.00	1.00	2.52	
	med	0.11	0.10	0.88	2.81	0.00	0.17	0.06	0.96	2.68		0.19	0.01	0.99	2.53		0.20	0.00	1.00	2.50	
	sd	0.15	0.41	0.45	28.36	0.00	0.10	0.27	0.27	12.50		0.04	0.11	0.09	0.25		0.03	0.08	0.06	0.17	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
PL_A^C	**mean**	0.18	0.01	0.76	1385	0.10	0.17	-0.02	0.97	193.4	0.10	0.18	-0.05	1.03	2.44	0.10	0.19	-0.04	1.03	2.42	0.10
	med	0.08	0.07	0.82	2.79	0.10	0.15	0.00	0.99	2.50	0.10	0.18	-0.04	1.03	2.41	0.10	0.19	-0.04	1.03	2.41	0.10
	sd	0.23	0.51	0.47	2E+4	0.03	0.16	0.30	0.28	5E+3	0.02	0.08	0.13	0.11	0.26	0.00	0.06	0.10	0.07	0.19	0.00
PL_B^C	**mean**	0.13	-0.10	0.77	7E+5	0.11	0.11	-0.10	0.98	167.3	0.11	0.15	-0.12	1.10	2.70	0.11	0.16	-0.12	1.11	2.29	0.00
	med	0.00	0.10	0.74	3.29	0.10	0.00	-0.02	1.02	2.43	0.11	0.17	-0.10	1.10	2.22	0.10	0.19	-0.10	1.09	2.23	0.00
	sd	0.22	0.71	0.49	2E+7	0.03	0.17	0.46	0.42	3978	0.02	0.12	0.24	0.22	7.25	0.01	0.10	0.18	0.18	0.46	0.00

The mean and median values and the standard deviations (across the 1000 replications) of the estimates of the four parameters are shown in Table 1. We also report the proportion (averaged across the R replications) of the T observations that, for the censored estimators, are classified as falling in the censored region.

Not surprisingly, for large samples ($T = 1000$) Table 1 shows higher standard errors for the parameters fitted to the censored data using the fixed point method, than to the uncensored data by ML. At the same, time, however, the results confirm that in large samples the censored estimators work well, especially when skew is moderate ($\gamma = 1.5$) rather than extreme ($\gamma = 2.5$): under both L_A^C and L_B^C when $\gamma = 1.5$ the mean and median values equal (to two decimal places) those in the data-generating-process.[15] When $\gamma = 2.5$, L_A^C continues to have this property but there is extra noise in the estimates for L_B^C, which imposes less structure than L_A^C. This is seen by L_B^C showing deviations from the true parameter estimates for γ. These deviations reflect a few outlying estimates, for some iterations, with the median values closer to the true parameter values than the mean ones. Both L_A^C and L_B^C correctly place, on average across R, 10% of observations in the censored region with little variation even for small T.

Once the sample size drops to 100, the problems with estimation of γ really start to appear. This is so for L_A^C but especially L_B^C. An increasing number of the R draws return inaccurate (high, divergent) estimates of γ: the median parameter estimates remain closer to the true values than the mean ones.[16] With L_A^C there is slight evidence of bias (looking at the mean across replications) when the true $\gamma = 1.5$; but when instead L_B^C is used, we can see that the mean parameter estimate for γ is contaminated by some very high values (for these draws this is accompanied by extremely low values for σ). Essentially the divergence problems reported by Sartori (2006) and Azzalini & Arellano-Valle (2013) emerge. The problem is worse when $\gamma = 2.5$ than when $\gamma = 1.5$; and there is evidence of it even when the distribution is not censored and the likelihood function L is used. These problems become more acute when the sample size drops to 40 observations. At this stage, the mean estimates for γ from L_A^C, L_B^C and L all give contaminated results due to the increased risk that for some replications the estimates for γ diverge. Use of the penalised estimator does offer some help in these smaller samples when $\gamma = 2.5$ especially for L_B^C. While it does not prevent the mean estimate for γ (across replications) from rising above the true value, the median estimates are closer to the true values than when a penalty is not imposed. We therefore conclude that in very small samples it may prove helpful, in effect, to have a prior that the data are symmetric. As even if they are not, imposing this view via the penalised estimator improves the accuracy of the median estimates even when the data are in fact highly skewed. The penalised censored estimators continue to place 10% of the 40 observations in the censored region. In larger samples (e.g. $T = 1000$),

[15] Convergence was also satisfied, with P_r converging to zero. An alternative and simpler method which would also work in large samples would be to set fixed censor points to exclude the upper and lower ten per cent of the observations. This would allow the parameters to be estimated straightforwardly.

[16] We note that if were to assume negative rather than positive skew in the data-generating process, i.e. $\gamma = 1/\gamma$, then the estimated γ are at risk of diverging to zero rather than infinity.

imposing a penalty does cause γ to be underestimated slightly, and in turn σ to be overestimated. But this bias is relatively modest, about 1% for L_A^C (for the median estimates) and about double this for L_B^C.

6.2 Experiment 2: Performance for mixed distributions

The second set of experiments explore the performance of the censored estimators when not all of the underlying data are drawn from the same (skewed) distribution: in particular when (what will be) the censored observations come from a different distribution. Following Experiment 1, T observations are first drawn from a two piece t distribution, where $(\nu, \mu, \sigma, \gamma) = (5, 0, 1, 1.5)$ and $(5, 0, 1, 2.5)$. But then each of the 10% of these T observations that falls outside the 90% best critical regions, y_L and y_U, as estimated for each replication, is dropped and replaced, depending on whether it falls below y_L or above y_U, with a random draw from a uniform density between -10 and y_L or y_U and $+10$.

As in Experiment 1, Table 2 reports results for $T = 40, 100, 500, 1000$, where $R = 1000$. Let us consider the larger sample results first. When $T = 1000$, we find that the censored estimators, again especially L_A^C, do a good job at estimating the true parameter values, despite the censoring. They also correctly place 10% of observations in the censored region. But, as expected, the uncensored estimator - that assumes all T observations come from a single density - is not able to return as accurate estimates. It tends to over-estimate $1/v$, in an attempt to capture the 10% of tail observations drawn from the uniform densities.

Table 2: Monte Carlo Experiment 2: Mean, median and standard deviation (across replications) of the parameter estimates from the censored estimators L_A^C, L_B^C, PL_A^C and PL_B^C and the uncensored ML estimator, L; and Proportion of observations placed in the censored region

		T=40					**T=100**					**T=500**					**T=1000**				
		$1/v$	μ	σ	γ		$1/v$	μ	σ	γ		$1/v$	μ	σ	γ		$1/v$	μ	σ	γ	
true		**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**
L_A^C	**mean**	0.37	-0.02	0.83	369	0.10	0.33	-0.01	0.91	2.01	0.10	0.26	0.00	0.97	1.51	0.10	0.24	0.00	0.98	1.50	0.10
	med	0.32	0.01	0.84	1.52	0.10	0.31	-0.02	0.90	1.49	0.10	0.24	0.00	0.97	1.51	0.10	0.23	0.00	0.97	1.50	0.10
	sd	0.35	0.49	0.27	1E+4	0.02	0.26	0.30	0.17	10.0	0.01	0.14	0.12	0.08	0.12	0.00	0.11	0.08	0.06	0.09	0.00
L_B^C	**mean**	0.27	0.01	0.60	1E+5	0.12	0.24	0.09	0.83	43.7	0.11	0.21	0.07	0.96	1.66	0.11	0.21	0.05	0.97	1.61	0.10
	med	0.00	0.13	0.64	1.85	0.10	0.15	0.07	0.87	1.64	0.11	0.21	0.06	0.95	1.59	0.10	0.21	0.05	0.96	1.58	0.10
	sd	0.37	0.78	0.95	4E+6	0.10	0.28	0.53	0.30	655	0.06	0.14	0.19	0.11	0.36	0.01	0.10	0.12	0.08	0.19	0.01
L	**mean**	0.49	-0.05	0.78	3.53		0.51	-0.06	0.83	1.52		0.52	-0.06	0.85	1.45		0.53	-0.07	0.85	1.44	
	med	0.50	-0.02	0.78	1.50		0.52	-0.07	0.83	1.44		0.52	-0.06	0.85	1.44		0.53	-0.07	0.85	1.44	
	sd	0.21	0.49	0.22	12.09		0.12	0.30	0.12	0.43		0.05	0.11	0.05	0.12		0.04	0.08	0.04	0.08	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
L_A^C	**mean**	0.30	-0.06	0.72	2E+5	0.10	0.26	-0.01	0.88	709	0.10	0.23	0.00	0.98	2.55	0.10	0.22	-0.01	0.99	2.51	0.10
	med	0.24	0.01	0.80	2.65	0.10	0.24	0.00	0.92	2.53	0.10	0.22	0.00	0.97	2.50	0.10	0.22	0.00	0.98	2.50	0.10
	sd	0.29	0.56	0.51	4E+6	0.05	0.21	0.35	0.32	2E+4	0.02	0.10	0.13	0.11	0.47	0.00	0.07	0.08	0.07	0.20	0.00
L_B^C	**mean**	1E+3	1E+3	0.54	5E+6	0.14	0.18	0.14	0.58	692	0.11	0.19	0.08	0.89	9.84	0.10	0.20	0.05	0.94	3.62	0.10
	med	0.00	0.19	0.03	51.57	0.10	0.00	0.19	0.67	3.98	0.10	0.20	0.06	0.92	2.74	0.10	0.21	0.04	0.94	2.66	0.10
	sd	4E+4	4E+4	4.32	2E+8	0.18	0.23	0.42	0.48	5E+3	0.07	0.12	0.20	0.23	58.6	0.01	0.08	0.13	0.12	26.60	0.00
L	**mean**	0.30	-0.13	0.77	18.27		0.33	-0.15	0.92	5.50		0.36	-0.17	0.99	2.22		0.37	-0.17	1.00	2.19	
	med	0.27	0.00	0.85	2.67		0.35	-0.13	0.95	2.25		0.36	-0.16	0.99	2.20		0.37	-0.17	1.00	2.17	
	sd	0.21	0.59	0.48	30.89		0.14	0.40	0.29	13.7		0.05	0.15	0.09	0.28		0.03	0.10	0.06	0.18	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
PL_A^C	**mean**	0.25	-0.11	0.81	7E+4	0.11	0.24	-0.06	0.94	2E+3	0.11	0.23	-0.05	1.01	2.43	0.10	0.22	-0.04	1.01	2.43	0.10
	med	0.18	-0.01	0.89	2.55	0.10	0.23	-0.05	0.95	2.43	0.11	0.23	-0.05	1.01	2.41	0.10	0.22	-0.04	1.01	2.41	0.10
	sd	0.27	0.55	0.45	2E+6	0.03	0.20	0.33	0.28	5E+4	0.02	0.11	0.13	0.10	0.28	0.00	0.08	0.09	0.08	0.20	0.00
PL_B^C	**mean**	0.15	-0.14	0.77	4E+4	0.11	0.15	-0.09	0.94	158	0.11	0.15	-0.03	1.02	3.41	0.11	0.17	-0.04	1.02	2.56	0.11
	med	0.00	0.08	0.78	3.18	0.10	0.00	-0.01	0.97	2.58	0.11	0.16	-0.01	1.00	2.52	0.11	0.19	-0.02	1.00	2.50	0.10
	sd	0.25	0.74	0.48	9E+5	0.04	0.21	0.45	0.41	3E+3	0.03	0.14	0.21	12.7	12.7	0.01	0.11	0.16	0.16	0.87	0.01

As T decreases and γ increases, we again observe a higher chance that the censored estimates for γ diverge for some replications: as the mean estimates for γ again become too large with the standard deviation estimates for γ elevated. Table 2 shows that in smaller samples this afflicts L_B^C more than L_A^C. The median estimates for L_A^C are closer to the true parameter values than the mean ones, especially so for smaller T. For $T = 40$ and $\gamma = 2.5$, focusing on the median estimates for γ, L_A^C is considerably more accurate than L_B^C, with L_B^C again tending, for an increasing number of replications, to overestimate γ (and underestimate σ). Use of the penalised estimator mitigates this small-sample concern further. The median estimates for the penalised estimator, under L_A^C, are within 10% of the true parameter estimates when $T = 40$. But it does not eliminate the risk of the skewness estimates diverging, as the mean estimates from PL_A^C still diverge suggesting that in any specific application with small-samples care should be exercised, and parameter estimates closely inspected, if boundary values are to be avoided. The uncensored estimator continues to over-estimate $1/v$ in small samples.

6.3 Experiment 3: Small sample confidence intervals for the MPC's forecast errors

Finally, we are interested in testing whether any parameter estimates produced when fitting the censored two-piece t to the time series of forecast errors could have been, in reality, generated by an underlying symmetric normal distribution. Since our time series of forecast errors has 76 or 77 observations, we carry out our Monte Carlo test for samples of the same length, as well as considering the smaller sample of $T = 40$ and larger samples, $T = 500, 1000$. An issue we have to address is that the forecast errors relate to GDP growth or inflation over four quarters. If (unobserved) quarterly forecast errors are independently distributed, then errors over four quarters will follow a moving average process. If the underlying distributions is symmetric normal, then so too will be the four-quarter errors. Thus, in order to generate the data used in this experiment we draw $T + 3$ values, each denoted by u_k from a normal distribution with unit variance. We then construct T observations

$$\varepsilon_k = \left(u_k + u_{k+1} + u_{k+2} + u_{k+3}\right)/2; \; k = 1, ..., T \tag{15}$$

so that ε_k has the same variance as u_k but also follows the moving average process which arises from analysis of four-quarter forecast errors. To each set of T observations we fit the skewed t distributions, both uncensored and on the assumption that the distribution is fitted only to the central 90% of the observations. The true parameters values are $(1/\nu, \mu, \sigma, \gamma) = (0, 0, 1, 1)$.

Table 3: Monte Carlo Experiment 3: Mean, median and standard deviation (across replications) of the parameter estimates from the censored estimators L_A^C, L_B^C, PL_A^C and PL_B^C and the uncensored ML estimator, L; plus 90% confidence intervals (CI) for the parameter estimates under normality

		T=40					**T=77**					**T=500**					**T=1000**				
		$1/v$	μ	σ	γ	Prop	$1/v$	μ	σ	γ	Prop	$1/v$	μ	σ	γ	Prop	$1/v$	μ	σ	γ	Prop
	true	**0.00**	**0.00**	**1.00**	**1.00**		**0.00**	**0.00**	**1.00**	**1.00**		**0.00**	**0.00**	**1.00**	**1.00**		**0.00**	**0.00**	**1.00**	**1.00**	
L_A^C	mean	0.09	-0.01	0.75	3.12	0.09	0.06	0.01	0.87	1.11	0.09	0.03	0.01	0.97	1.01	0.10	0.03	0.00	0.98	1.00	0.10
	med	0.00	0.00	0.76	0.99	0.08	0.00	-0.01	0.87	1.01	0.09	0.00	-0.01	0.97	1.00	0.10	0.00	0.00	0.98	1.00	0.10
	sd	0.17	0.71	0.28	17.77	0.03	0.12	0.53	0.17	1.61	0.02	0.05	0.19	0.07	0.11	0.00	0.04	0.13	0.05	0.07	0.00
	low CI	**0.00**	**-1.20**	**0.04**	**0.32**	**0.03**	**0.00**	**-0.87**	**0.60**	**0.58**	**0.07**	**0.00**	**-0.29**	**0.85**	**0.85**	**0.09**	**0.00**	**-0.20**	**0.89**	**0.89**	**0.09**
	up CI	**0.40**	**0.88**	**1.09**	**1.97**	**0.14**	**0.37**	**0.68**	**1.10**	**1.48**	**0.11**	**0.11**	**0.38**	**1.06**	**1.24**	**0.10**	**0.10**	**0.22**	**1.05**	**1.12**	**0.10**
L_B^C	mean	0.08	0.00	0.50	42.77	0.08	0.06	0.00	0.69	9.98	0.09	0.03	0.02	0.95	1.04	0.10	0.03	0.00	0.97	1.01	0.10
	med	0.00	0.01	0.54	0.97	0.08	0.00	0.02	0.78	1.02	0.09	0.00	0.00	0.96	1.02	0.10	0.00	0.00	0.97	1.00	0.10
	sd	0.19	0.86	0.45	226.3	0.04	0.14	0.71	0.34	53.84	0.02	0.05	0.25	0.07	0.22	0.01	0.04	0.17	0.05	0.14	0.00
	low CI	**0.00**	**-1.41**	**0.00**	**0.00**	**0.03**	**0.00**	**-1.24**	**0.00**	**0.00**	**0.05**	**0.00**	**-0.39**	**0.83**	**0.73**	**0.09**	**0.00**	**-0.28**	**0.88**	**0.80**	**0.09**
	up CI	**0.37**	**1.10**	**1.01**	**101.2**	**0.14**	**0.44**	**1.12**	**1.03**	**5.09**	**0.12**	**0.13**	**0.37**	**1.07**	**1.32**	**0.10**	**0.11**	**0.26**	**1.04**	**1.19**	**0.10**
PL_A^C	mean	0.05	0.01	0.78	2.41	0.09	0.06	-0.04	0.88	1.02	0.09	0.04	-0.03	0.97	0.99	0.10	0.03	-0.02	0.98	0.99	0.10
	med	0.00	0.05	0.80	1.03	0.10	0.00	-0.04	0.88	0.99	0.09	0.00	-0.03	0.97	0.98	0.10	0.00	-0.02	0.98	0.98	0.10
	sd	0.12	0.73	0.26	12.88	0.03	0.11	0.48	0.16	0.32	0.02	0.05	0.17	0.07	0.10	0.01	0.04	0.14	0.05	0.08	0.00
	low CI	**0.00**	**-1.28**	**0.14**	**0.46**	**0.05**	**0.00**	**-0.78**	**0.62**	**0.59**	**0.06**	**0.00**	**-0.30**	**0.86**	**0.83**	**0.09**	**0.00**	**-0.25**	**0.89**	**0.86**	**0.09**
	up CI	**0.30**	**1.12**	**1.13**	**2.37**	**0.13**	**0.29**	**0.74**	**1.15**	**1.56**	**0.12**	**0.15**	**0.22**	**1.08**	**1.17**	**0.10**	**0.10**	**0.22**	**1.05**	**1.13**	**0.10**
PL_B^C	mean	0.06	0.03	0.55	3.56	0.09	0.04	-0.08	0.73	1.35	0.09	0.03	-0.06	0.95	0.97	0.10	0.02	-0.05	0.97	0.97	0.10
	med	0.00	0.09	0.65	1.03	0.10	0.00	-0.08	0.80	0.97	0.09	0.00	-0.07	0.95	0.94	0.10	0.00	-0.05	0.98	0.95	0.10
	sd	0.15	0.89	0.37	14.89	0.03	0.11	0.67	0.31	2.27	0.02	0.05	0.25	0.08	0.21	0.01	0.04	0.17	0.05	0.13	0.00
	low CI	**0.00**	**-1.47**	**-0.01**	**0.00**	**0.05**	**0.00**	**-1.24**	**0.00**	**0.00**	**0.06**	**0.00**	**-0.46**	**0.83**	**0.70**	**0.09**	**0.00**	**-0.31**	**0.87**	**0.77**	**0.09**
	up CI	**0.43**	**1.41**	**1.04**	**7.89**	**0.13**	**0.32**	**1.02**	**1.11**	**3.61**	**0.12**	**0.14**	**0.43**	**1.07**	**1.31**	**0.11**	**0.11**	**0.26**	**1.04**	**1.22**	**0.10**
L	mean	0.03	0.04	0.83	1.13		0.02	0.03	0.91	1.05		0.01	0.04	0.98	1.03		0.01	0.02	0.98	1.02	
	med	0.00	0.12	0.83	1.03		0.00	0.12	0.90	1.03		0.00	0.07	0.98	1.03		0.00	0.04	0.98	1.02	
	sd	0.07	0.51	0.23	1.07		0.05	0.39	0.15	0.26		0.02	0.16	0.06	0.09		0.02	0.12	0.04	0.07	
	low CI	**0.00**	**-0.95**	**0.49**	**0.52**		**0.00**	**-0.70**	**0.67**	**0.65**		**0.00**	**-0.26**	**0.88**	**0.87**		**0.00**	**-0.19**	**0.91**	**0.90**	
	up CI	**0.17**	**0.69**	**1.11**	**1.57**		**0.14**	**0.46**	**1.10**	**1.37**		**0.07**	**0.30**	**1.07**	**1.18**		**0.04**	**0.18**	**1.04**	**1.11**	

Table 3 shows the mean, median and standard deviation of each parameter taken from the $R = 1000$ draws, together with the upper and lower ninety per cent confidence limits.

A number of things stand out from Table 3, beyond the obvious point that the fit of L_A^C and L_B^C is much worse, with the confidence intervals wider, with small samples than with $T = 1000$ observations. $1/\nu$ cannot be expected to be symmetric around its true value of zero, so a bias inevitably exists in the small-sample estimates. A related bias appears in the estimate of σ. A low number of degrees of freedom and a high value of σ are both ways of accommodating observations distant from the mode, so bias in one implies a bias in the other. There is little evidence of bias in the mode, μ, since the confidence limits are reasonably symmetric. γ appears skewed to the right, especially so for small samples and for L_B^C rather than L_A^C; the confidence limits are asymmetric.

When the distribution is not censored the estimates for σ and γ in Table 3 are better determined than when censor points are estimated simultaneously. This is not very surprising. But for smaller samples a bias does appear in L's estimates for μ. It is also worth noting that, even when the distribution is not censored, when $T = 77$ an estimate of $1/0.14 = 7$ degrees of freedom has a 5% chance of arising from an underlying normal distribution.

When the data are censored so that the distribution is fitted to only the central 90% of observations, then the estimated value of the number of degrees of freedom has to be 2.7 (2.2) or lower under L_A^C (L_B^C) before one can reject, at a 90% level, the hypothesis that the underlying distribution is normal.

There is again evidence of a higher possibility of divergence in the censored estimates of γ (and in turn those for σ) for smaller samples as evidenced by a higher standard deviation for γ. But L_A^C is less contaminated than L_B^C by some high values for γ. Contamination for both estimators is worse when T drops from 77 to 40, which should be borne in mind in our out-of-sample application (section 8.2 below). We note that the median estimates for γ from L_A^C and L_B^C remain accurate, close to unity, even when $T = 40$. But, as seen from comparison with Table 1, this feature is specific to when there is no skew in the population data. Recall we found that when there is population data skew, the median estimates for γ from L_A^C and L_B^C differ from the true value - and the penalised estimators are likely to be preferred in such small-samples. Table 3 shows that with symmetric data the penalised estimators continue to lower the chance of divergent estimates for γ.

Overall, Table 3 shows that in small samples L_A^C continues to be preferred to L_B^C as its estimates for the four parameters are better determined - its mean and median estimates are closer to the true values with lower standard deviations and tighter and more symmetric confidence bands. But (even without the population data skew considered in Tables 1 and 2) there remains a chance in smaller samples that estimation using L_A^C delivers divergent values for γ. So in practice, including in the out-of-sample application in Section 8.2 below, we recommend looking closely at the parameter estimates for fear they

involve an (economically) unappealing boundary value for γ. If the estimated value of γ diverges, the resulting density in effect becomes a half or folded density; for an illustration of such a cliff-edged density see Figure A1 in the online Appendix. If estimates do diverge, based on the results in Tables 1 to 3, we suggest use of our penalised estimator as it is found to mitigate, albeit not eradicate, the possibility of boundary values in small samples. Moreover, in any applications when estimation does appear to reflect divergence - with the estimates of γ (or $1/\gamma$) rising above a threshold value of say 5 or 10 implying a half or folded density - the estimates might be rejected and model/density estimation reconsidered.

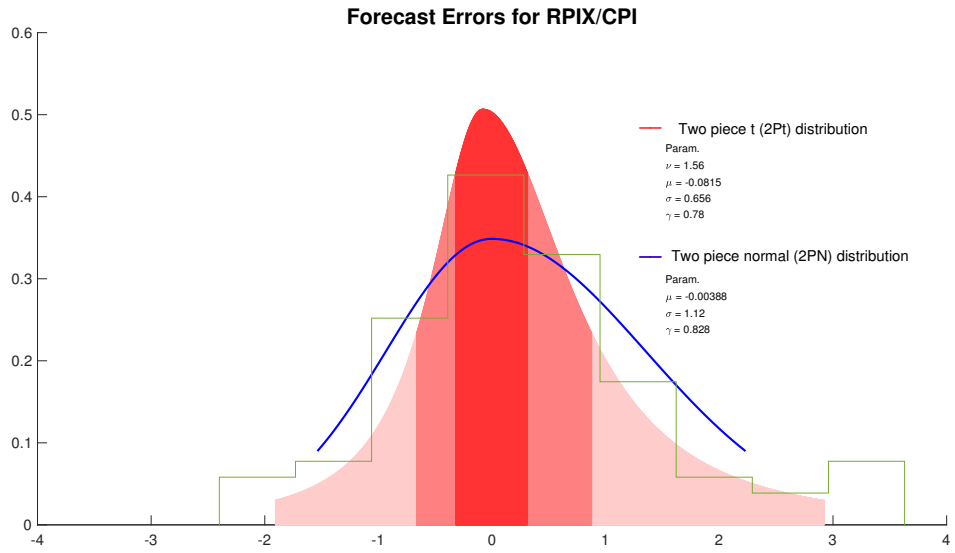
7 Censored densities fitted to the MPC forecast errors

We now fit the censored density functions to the MPC's forecast errors for inflation and GDP. We use L_A^C because, consistent with the Monte Carlo evidence and attempts to fit L_B^C to the forecasts errors (reported in section A.2.2 of the online Appendix), we found L_A^C better able to avoid boundary solutions for γ when fitting the two-piece normal distribution. In all cases $P_r = 0$ (for large r), confirming satisfactory estimation. The results for inflation are shown in Figure 6 and those for GDP in Figure 7.

Looking at GDP growth first, comparing with the uncensored distributions of Figure 5, we see that not allowing the outlying ten per cent of the observations to influence the shape of the distribution has a considerable effect on skewness. For the two-piece normal density the degree of skew present in Figure 7 drops considerably. Many of the negative forecast errors (observed over the period of the financial crisis) are now censored, placed in the left tail, rather than accommodated, as in Figure 5, via a higher skew estimate. It remains the case, however, that a t distribution with a low number of degrees of freedom, 3.9, is needed to capture the peak of the distribution of the GDP forecast errors. Although Table 3 shows that this falls within the 90% confidence interval for a normal density when there are only $T = 77$ observations. The two-piece normal distribution does a poorer job of reflecting the peak of the distribution in Figure 7 because of the extra spread needed to fit the outer parts of the uncensored region; right skew is also present.

By way of contrast, as shown in section A.2.3 of the online Appendix, if uncensored two-piece t and normal densities are fitted not to all the forecast errors as in Figure 5, but just to a sample of GDP growth errors before or after the financial crisis, we also find much less skew than in Figure 5. This is especially so for the two-piece normal density. This supports the view that analysing forecast errors over a rolling window, as is apparently practiced at the Bank of England, also amounts to a form of censoring. But it is *ad hoc* and inconsistent with the fact that the density is later, at a second step, censored. In fact, when using only forecast errors since the financial crisis, there is no skew to the two-piece normal and the preferred density is normal (symmetric) with a similar variance to the two-piece t in Figure 7. One implication of this type of *censoring* is that the probability of large forecast errors is much lower

Figure 6: Inflation: Forecast Error Histogram and Censored Two-Piece Normal and t Densities using L_A^C



Note: RPIX forecasts (until Nov. 2003); CPI forecasts thereafter. 77 observations used. The p-values from tests for the uniformity of the in-sample probability integral transforms (for the forecast error observations, seen in the histogram, that are not censored), using an Anderson-Darling test are: 0.94 (2Pt) and 0.12 (2PN). 9% of the observations fall outside 2Pt and 15% outside 2PN.

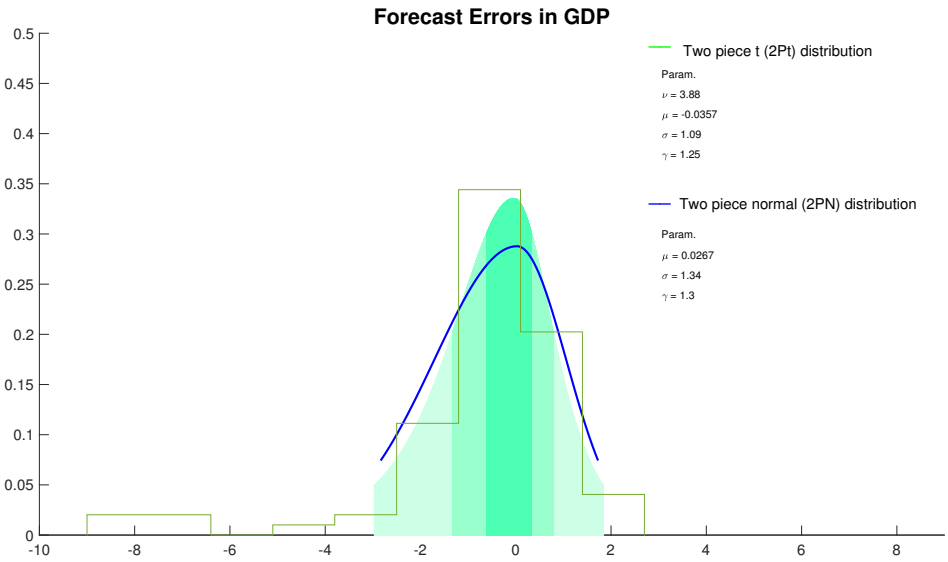
than in Figure 5 or Figure 7.

Turning to inflation, comparison of Figures 4 and 6 shows censoring to have a less material affect on the skew parameter of the estimated distributions than it did for GDP growth. There is evidence that extending the two-piece normal to the two-piece t helps when censoring - as the ML estimate for ν is lower in Figure 6 than Figure 4. The smaller value for ν, which is now outside the 90% small-sample confidence interval for a normal density (cf. Table 3), enables the distribution to accommodate the concentration of probability mass close to zero better; a lower variance is also required than in Figure 4. We observe in Figure 6 wider censoring bounds using the two-piece t rather than the two-piece normal; with close to 10% of observations placed outside these bounds. We also note that when testing the uniformity of the probability integral transforms (computed by evaluating the CDF for all non-censored observations) we see some evidence, as the p-values are higher, that the two-piece t fits the non-censored observations better.[17]

In summary, we conclude that inference on the estimated parameters is affected if the censored nature of the forecast density is (correctly) acknowledged when fitting distributions to past forecast errors. In particular, notably for GDP, one would not produce nearly so a skewed density having censored the outlying ten percent of observations. The shape of fan charts (estimated from past forecast errors) can be materially affected by whether the censoring is accommodated in estimation.

[17] Section A.2.3 of the online Appendix also plots uncensored densities fitted to rolling windows of inflation forecast errors. These densities do not resemble those in Figure 6.

Figure 7: GDP Growth: Forecast Error Histogram and Censored Two-Piece Normal and t Densities using L_A^C



Note: 76 observations used. Latest release estimates used to define the outturns. The p-values from tests for the uniformity of the in-sample probability integral transforms (for the forecast error observations, seen in the histogram, that are not censored), using an Anderson-Darling test are: 0.40 (2Pt) and 0.23 (2PN). 11% of the observations fall outside 2Pt and 13% outside 2PN.

8 Evaluation of censored density forecasts

The MPC's density forecasts attract considerable attention as the focal point of monetary policy debate and communication. This attention has involved testing the forecasts to evaluate their *ex post* accuracy. But none of these studies have, at least directly, acknowledged the censoring. While all previous evaluations of the MPC densities ignore censoring, those that reduce the density to fewer than 10 intervals and then evaluate these can be coincidentally robust to it since they do not, in effect, test the fit of the density in the censored tails.[18]

We first consider, in section 8.1, evaluation tests appropriate for censored density forecasts before considering, in section 8.2, their application to the MPC densities and our forecast error data.

8.1 Evaluation tests

Following the distinction made when evaluating (uncensored) density forecasts, we consider statistical tests, suitable for censored density forecasts, for both absolute and relative forecast accuracy. The former test forecast accuracy relative to the 'true' but unobserved density; and prominent tests (following Diebold, Gunther & Tay (1998)) involve application of goodness-of-fit tests to the probability integral

[18] Previous studies at the Bank of England include: Boneva, Fawcett, Massolo & Waldron (2018); Independent Evaluation Office (2015); Hackworth, Raidia & Roberts (2013); Elder et al. (2005). External studies include: Wallis (2003); Wallis (2004); Clements (2004); Mitchell (2005); Mitchell & Hall (2005); Hall & Mitchell (2007); Dowd (2007); Dowd (2008); Boero, Smith & Wallis (2011); Gneiting & Ranjan (2011) and Galbraith & van Norden (2012).

transforms (PITs). The latter involve comparison of two or model competing density forecasts via application of scoring rules.[19]

8.1.1 PITs based evaluation of censored density forecasts

Let f and F continue to denote the (time-varying) probability and cumulative density functions of the (two-sided) censored density forecast; and let y_t denote the subsequent outturn with $t = 1, ..., T$ now denoting the out-of-sample evaluation period. The forecast density is assumed to be censored at $100\alpha\%$ ($\alpha \in (0, 1)$), between the thresholds $y_{L,t}$ and $y_{U,t}$, where $y_{U,t} > y_{L,t}$, such that $\int_{y_{L,t}}^{y_{U,t}} f(y_t) = 1 - \alpha$.

The PITs are defined, in the usual way, as $z_t = F(y_t)$. But, given the censoring, we also define PIT thresholds $z_{L,t} = F(y_{L,t})$ and $z_{U,t} = F(y_{U,t})$; e.g., $z_{L,t} = 0.05$ and $z_{U,t} = 0.95$ for 10% censoring with symmetric thresholds (about the mean). The censored density forecast $f(y_t)$ is well-calibrated when $z_{c,t}$, defined as:

$$z_{c,t} = z_t \text{ if } z_{L,t} < z_t < z_{U,t}, \tag{16}$$

rather than z_t, is uniformly distributed. So calibration should involve testing $E(z_{c,t}) = 0.5(z_{L,t} + z_{U,t}) = 0.5$ and $Var(z_{c,t}) = (1/12)(z_{U,t} - z_{L,t})^2$. Outside of the uncensored range, $z_{L,t} < z_t < z_{U,t}$, calibration of $f(y_t)$ requires correct unconditional coverage:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{I}(z_t \leq z_{L,t}) + \mathbf{I}(z_t \geq z_{U,t}) = \alpha. \tag{17}$$

Practically, with statistical testing in mind, it is convenient to adjust the range of $z_{c,t}$ to accommodate the censoring, by defining and then subjecting $\left\{\frac{z_{c,t}-z_{L,t}}{z_{U,t}-z_{L,t}}\right\}_{t=1}^{T}$ to standard goodness-of-fit tests, such as the likelihood ratio (LR) test proposed by Berkowitz (2001). But such tests should be complemented with a separate test for correct unconditional coverage, (17), via, for example, a Christoffersen (1998) LR test. As Askanazi et al. (2018) explain, a consequence of using BCRs is that under correct unconditional coverage no other set of interval forecasts (extracted from the same density forecast) with shorter intervals can also satisfy this condition.

A joint test is offered by considering a two-sided variant of the censored LR test proposed by Berkowitz (2001). This ignores the degree of forecast failure in both the 10% (left and right-hand-side) tails but importantly does account for their frequency (as argued for by, e.g., Diks et al. (2011)). Specifically, following Berkowitz (2001), take an inverse normal CDF transformation, Φ^{-1}, of the PITs to define

[19] In both cases, we consider unweighted versions of the tests, although we note that with greater assumed knowledge of the MPC's loss function tests could be developed that attach more importance to predicting accurately the probabilities in certain (uncensored) ranges of (economic) interest. See Clements (2004) for an interesting application to the MPC's forecasts, albeit one that does not acknowledge the censoring. As emphasised by Diebold et al. (1998) and Granger & Pesaran (2000), a rationale for the statistical based PITs tests is that if one cannot reject correct calibration in an absolute sense then the density forecast can be considered 'optimal' irrespective of the form of the MPC's loss function.

$z_t^* = \Phi^{-1}(z_t)$; and also define $z_{L,t}^* = \Phi^{-1}(z_{L,t})$ and $z_{U,t}^* = \Phi^{-1}(z_{U,t})$ such that:

$$z_{c,t}^* = z_t^* \text{ if } z_{L,t}^* \leq z_t^* \leq z_{U,t}^* \tag{18}$$

$$z_{c,t}^* = z_{L,t}^* \text{ if } z_t^* < z_{L,t}^* \tag{19}$$

$$z_{c,t}^* = z_{U,t}^* \text{ if } z_t^* > z_{U,t}^* \tag{20}$$

so that the log likelihood function for estimation of m and s, which should be $(0, 1)$ under correct calibration, is given as:

$$\begin{aligned} L(m, s \mid z_{c,t}^*) = & \sum_{z_{L,t}^* < z_{c,t}^* < z_{U,t}^*} \log \frac{1}{s} \phi \left(\frac{z_{c,t}^* - m}{s} \right) \\ & + \sum_{z_{c,t}^* = z_{L,t}^*} \log \Phi \left(\frac{z_{L,t}^* - m}{s} \right) \\ & + \sum_{z_{c,t}^* = z_{U,t}^*} \log \left(1 - \Phi \left(\frac{z_{U,t}^* - m}{s} \right) \right). \end{aligned} \tag{21}$$

Therefore, a censored (or tail) LR test statistic can be constructed as:

$$LR_{tail} = -2(L(0, 1) - L(\widehat{m}, \widehat{s})) \tag{22}$$

that is distributed $\chi^2(2)$ under the null hypothesis that the censored density forecast is correctly calibrated (i.e., $m = 0$ and $s = 1$). This two degrees of freedom variant of Berkowitz's test (see Clements (2004)) does not test for independence in the PITs; we should not expect independence, under correct calibration, in applications like ours where the forecast horizon is greater than one-step ahead.

We can also evaluate the censored density forecast by reducing to it a series of (albeit there are an infinite number of) interval forecasts. Following Wallis (2003), a Pearson chi-squared test dividing the PITs into k intervals might be constructed. As Wallis (2003) explains, this is a generalisation of Christoffersen's test for $k > 2$ intervals. How we construct these k intervals depends on how we should interpret the underlying density forecast. For example, for the MPC's density forecasts these intervals should arguably be constructed not using quantiles but BCRs. This is an important illustration of how the way in which the MPC choose to communicate their forecast affects the appropriate evaluation test.

8.1.2 Scoring rules with censored density forecasts

Scoring rules evaluate the quality of probability forecasts by assigning a numerical score based on the forecast and the subsequent outturn. As Gneiting & Raftery (2007) review, various scoring rules have been proposed. We focus on a popular one from within the class of strictly "proper" scoring rules - the logarithmic score - and show how it can be applied in a manner that acknowledges the censoring.[20] While

[20] A scoring rule is said to be "proper" if it always ranks the true conditional density forecast above any incorrect density forecast; see Gneiting & Raftery (2007). As Diks et al. (2011) and Gneiting & Ranjan (2011) explain, care has to be exercised when evaluating densities over regions of interest to ensure the scoring rule is proper. An alternative scoring rule, the Continuous Ranked Probability Score (CRPS), is also popular. It too can be straightforwardly applied to censored density forecasts, in effect by integrating the quantile scores between $z_{L,t}$ and $z_{U,t}$.

the previous literature has not applied these tests specifically to the MPC's forecasts, their development reflects interest in other areas of applied statistics in assessing predictive accuracy in regions of perceived interest.

Diks et al. (2011) propose conditional and censored logarithmic scores; and discuss their interpretation and properties. They find that a censored scoring rule performs better in many cases, and so in our application we focus on it. The average censored logarithmic score, in our context, is defined as the out-of-sample analogue of $\log L_A^C$ seen in (5):

$$\overline{LS}_A^C = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \{\mathbf{I}(y_t - y_{L,t})(1 - \mathbf{I}(y_t - y_{U,t}))\}\log f(y_t) + \\ \mathbf{I}(y_t - y_{U,t})\log(1 - F(y_{U,t})) + (1 - \mathbf{I}(y_t - y_{L,t}))\log F(y_{L,t}) \end{bmatrix} \tag{23}$$

Similarly, an out-of-sample analogue, $\overline{LS}_B^C$, can be defined for $\log L_B^C$ as seen in (8).

$\overline{LS}_A^C$ considers the likelihood associated with the outturn y_t being in the uncensored region. But when an outturn falls in the censored region, like (21), it ignores the shape of $f(y_t)$. However, knowledge of the asymmetries in tail risks is acknowledged. $\log F(y_{L,t}) > \log(1 - F(y_{U,t}))$ when $(1 - F(y_{U,t})) > F(y_{L,t})$, i.e. when the forecaster believes there to be higher chance they are going to be surprised on the upside than downside so that the upside is censored more than the downside. Tests of equal forecast performance, across competing censored density forecasts, can be constructed based on their $\overline{LS}^C$ values as discussed in Diks et al. (2011) and Gneiting & Ranjan (2011). We use these in our application below.

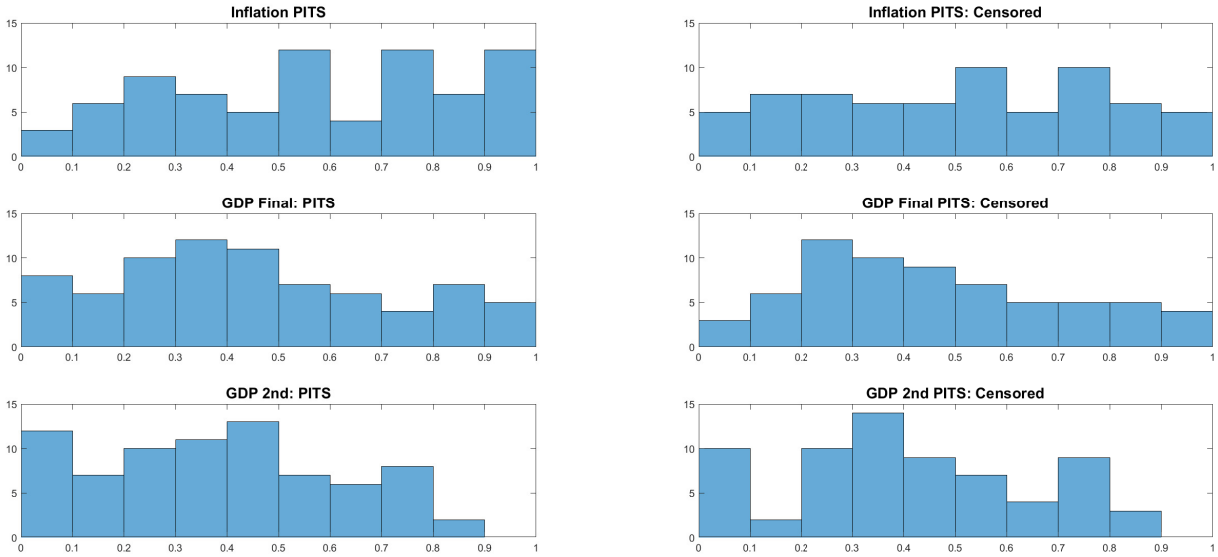
8.2 Evaluating the MPC's (censored) forecasts

We first evaluate the *ex post* accuracy of the MPC's density forecasts, but extend previous studies by acknowledging the censoring explicitly. We then compare the MPC forecasts with benchmark statistical density forecasts formed by recursively fitting censored normal, t, two-piece normal and two-piece t densities to the MPC's historical (modal) point forecast errors. This is done as if in real-time. This comparison helps us both assess the relative value of the MPC's judgement-informed forecasts; it also serves as an illustrative example of what data-based forecasts, acknowledging the censoring, could have been presented in real-time to the MPC ahead of each quarterly *Inflation Report*.

8.2.1 Evaluating the MPC's forecasts as censored density forecasts

Figure 8 provides a visual impression of the calibration of the PIT values implied by the MPC's GDP growth and inflation density forecasts. GDP 'outturns' are defined using both "final" and second estimates. PIT histograms are shown both ignoring (left plots) and then reflecting (right plots) the self-imposed censoring that the MPC applies. We note that application of the joint test, LR_{tail}, seen in (22), rejected calibration of both the inflation and GDP growth density forecasts (with p-values of 0.00 in all cases). To help understand this rejection we now look for uniformity of the PITs and the coverage rates of the censored density forecasts separately.

Figure 8: MPC Inflation and GDP Forecast PITs (for 2 year ahead forecasts) calculated using outturns from 1999q4 to 2018q3/q4



Note: The histogram figures on the left ignore the MPC's censoring and plot z_t; those on the right do not and plot $\left\{\frac{z_{c,t}-z_{L,t}}{z_{U,t}-z_{L,t}}\right\}$. The GDP histograms are shown using both the final or latest (December 2018) data release and the second data release to define the GDP outturns.

Specifically, the left panels of Figure 8 show histograms for the PITs, z_t, for the 76 or 77 forecasts, with ten bins of width 0.1. The plots on the right are for the rescaled PITs acknowledging the censoring: $\left\{\frac{z_{c,t}-z_{L,t}}{z_{U,t}-z_{L,t}}\right\}$. Note that the MPC know $z_{L,t}$ and $z_{L,t}$ (given their 10% BCR thresholds) at the time they make their forecasts. So subsequent PITs outside this range should not be tested for uniformity.[21]

If the MPC did not censor its density forecasts, the more any of these PITs histograms deviate from uniformity the weaker the evidence for correct forecast calibration. But, with censoring, only the plots on the right-hand-side of Figure 8 must be uniform. That is, non-uniformity of the PITs in the plots on the left could be a feature not of calibration failure, but of failing to account for the censoring in evaluation. Thinking about the left plots, we should not expect 10% of PITs to be less than 0.1 or 10% to be greater than 0.9 when the density forecast is asymmetric - given the MPC's use of BCRs.

The uncensored PITs plot in the top left of Figure 8 - for inflation - is reminiscent of figures in Independent Evaluation Office (2015) (see Box 4 on their pages 50-51). That is, relative to forecast, there were too many high outturns for inflation; in other words, the MPC understated the probability of high inflation outturns, especially from 2007. But this specific plot, as discussed, does not acknowledge the censoring. When we correct for this and look instead at the top right plot the story seems to change. The top right histogram plot appears more uniform. This suggests that the MPC's forecasts were not so

[21] In Section A.4 of the online Appendix we also show plots having reduced the density forecast to 10 interval forecasts. As anticipated in Section 8.1, we break the density forecast into interval forecasts based on both (central) percentiles and 10% BCRs. We find inference is sensitive to this choice.

bad after all, when we rightly acknowledge the censoring. We subject each of the histograms in Figure 8 to Pearson chi-squared tests as discussed in Wallis (2003) and find p-values of 0.14 (top left), 0.85 (top right). These p-values therefore confirm our visual impression that calibration is better in the top right than top left panel of Figure 8.

The second and third rows of Figure 8 provide the PITs plots for the MPC's GDP forecasts, defining errors against final vintage and second release GDP outturn data, respectively. Comparison of these two rows suggests that calibration is still imperfect, but stronger, when measured against final release rather than second release GDP data. This is because the plots against the second release data reveal a greater tendency for the MPC's GDP forecasts to be too low. This is reflected by the downward slope of the histograms in the bottom panels of Figure 8 and in particular an empty bin from $0.9 - 1$. These differences illustrate the importance of GDP data revisions, the tendency for these revisions to be positive (see Galvao & Mitchell (2019)) and the MPC's stated ambition of forecasting "mature" GDP data.

But looking more closely in Figure 8 at GDP forecast performance against the final vintage data, we see that inference about calibration is again sensitive to censoring. The PITs plot in the middle left panel shows that when we do not acknowledge the censoring we see many outturns fall towards the left of the MPC's density - this reflects, in part, the forecasting errors made over the crisis with outturns falling well below forecasts (this is no longer true). While this is still true of the plots on the middle right, which acknowledge the censoring, the pattern of density forecasting errors is a little different.

8.2.2 Comparison of the Estimated Parameters with those used by the MPC

We now move on from an assessment of the MPC's forecasts to a comparison between its density forecasts and those generated using the methods set out above. We compare the MPC's own (censored) judgement-based density forecasts with censored two-piece normal (2PN) and two-piece t densities (2Pt) recursively fitted to the MPC's historical (modal) point forecast errors. We also experiment with censored symmetric densities - the normal and t - aware from the Monte Carlo study that estimation of censored asymmetric densities may suffer from divergent skew estimates especially in smaller samples.

In what follows we assess the performance of the forecasts generated by the different parameter sets; thereby we identify if, when and how the MPC's judgement deviated from the data-based evidence.

These exercises involve recursively re-estimating the parameters of the 2Pt, 2PN, Normal and t densities using information that would have been available to the forecaster in real-time. Specifically, we consider that the forecaster had access to the MPC's historical forecast errors (for GDP defined against the latest available GDP vintage extracted from the Bank of England's "GDP real-time database"[22]) but lagged to reflect both publication lags (for GDP) and the fact that they have to wait two years to define the forecast error (for their 2-year ahead forecasts). The four censored densities are fitted uncondition-

[22] Available at http://www.bankofengland.co.uk/statistics/Pages/gdpdatabase/default.aspx

ally, albeit recursively adding an observation each quarter, to the historical forecast errors dating back to 1998q1. We estimate the error densities using PL_A^C, although we note that for the majority of the recursive estimations this amounts to use of L_A^C since the estimated $\lambda = 0$. But on occasion, consistent with the Monte Carlo findings in Tables 1 to 3, imposing a penalty prevented divergent estimates for γ.[23] We then set the mode of these densities to the latest modal forecast from the MPC and thereby compute censored forecast densities for GDP growth and inflation, noting that if $X \sim 2Pt(\mu, \sigma, \gamma, \nu)$ and, if $W = c + X$, then $W \sim 2Pt(c + \mu, \sigma, \gamma, \nu)$. We start making forecasts in 2003q2 for 2005q1 (two years ahead), and recursively update the sample so that the final forecast we make is in 2016q4 (2017q1) of 2018q3 (2018q4, for inflation). This means our out-of-sample window is from 2005q1 to 2018q3 for GDP and to 2018q4 for inflation.

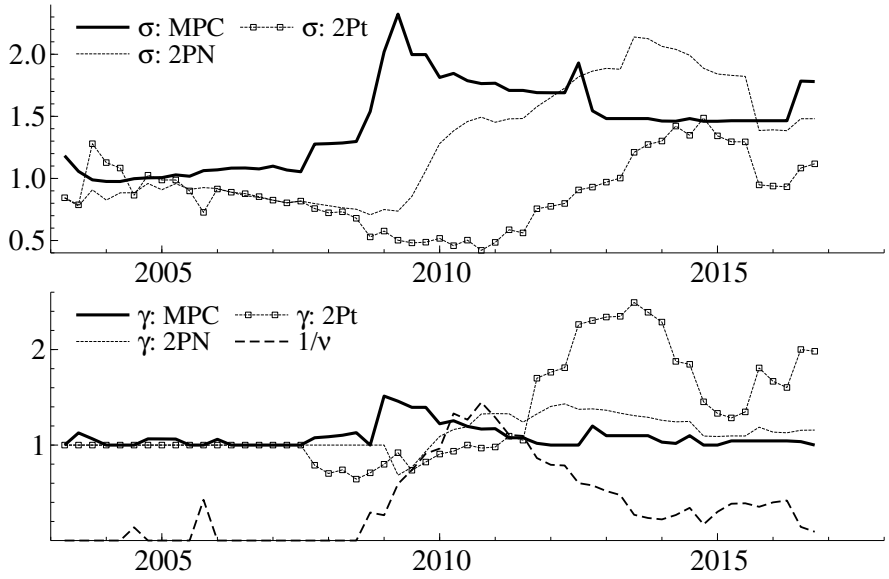
Figure 9 shows the time series of the scale and skew parameters, σ and γ, for the GDP distributions used by the MPC and also for the distributions fitted recursively to past GDP forecast errors when using the two-piece normal and two-piece t distributions. It confirms that divergent estimates of γ are avoided by the censored estimators despite the small samples, especially at the beginning of the out-of-sample period. We also show the reciprocal of the number of degrees of freedom for the two-piece t distribution, $1/\nu$. Looking first at the MPC's parameters, we can see that there was a sharp increase in σ at the time of the financial crisis. Specifically, in late 2008 and during 2009 the MPC increased σ: the MPC saw a break, while the data based methods adjust only gradually. This is borne out by the fact that a negative (four-quarter) GDP growth rate outturn was not observed (published) until late January 2009 (with the ONS estimate for 2008q4 growth). Data revisions subsequently identified 2008q3 as the first quarter of negative growth and also increased the depth of the recession.[24] The MPC has gradually retreated from the high values set at the height of the crisis, although the recent numbers show an increase. The skew parameter was set close to 1 until the crisis, when it was raised delivering a pronounced downside skew. Since then it has gradually declined, albeit with some reversals. The increase in skew associated with the 2015 Greek crisis is very modest, as our earlier graph showed.

Looking now in Figure 9 at the fitted parameters for the two-part normal distribution, we see that, as might be expected with recursive estimation, it did indeed take some time for σ to rise to accommodate the GDP forecasting errors of the financial crisis. The sharp fall in σ in late 2015 is explained by the fact that, with accrual of data, what had previously been large errors in the non-censored part of the distribution

[23] Even when imposing a penalty we found it hard to obtain sensible fitted densities using PL_B^C. L_A^C and PL_A^C impose more discipline in estimation, and this appears to help, especially out-of-sample with the smaller sample sizes. To estimate λ we searched over a grid, selecting that λ that maximised L_A^C.

[24] We undertook structural break tests for a break in the unconditional variance of the forecast errors, to see how quickly the forecast error data alone would have picked up a change. Under the assumption that the forecast errors follow a normal distribution with mean zero, then $\sqrt{\pi/2}\,|y_t - \mu|$ is an unbiased estimator of the standard deviation of $y_t - \mu$. We may then test for a structural break in the unconditional volatility of the forecast errors by testing for a break in the mean of their absolute values. We treat the break point as unknown and use the sup-Wald statistic with approximate asymptotic p-values as in Hansen (1997). Recursive application of these tests suggests that a break would not have detected statistically, with a p-value less than 0.05, until late 2011 although the p-values do decline from late 2009.

Figure 9: Data-based and MPC Parameters for the GDP Fan Chart (2 year ahead forecasts)



Note: Recursively updated latest vintage used to define GDP outturns. Dates refer to when the two-year ahead forecast was made.

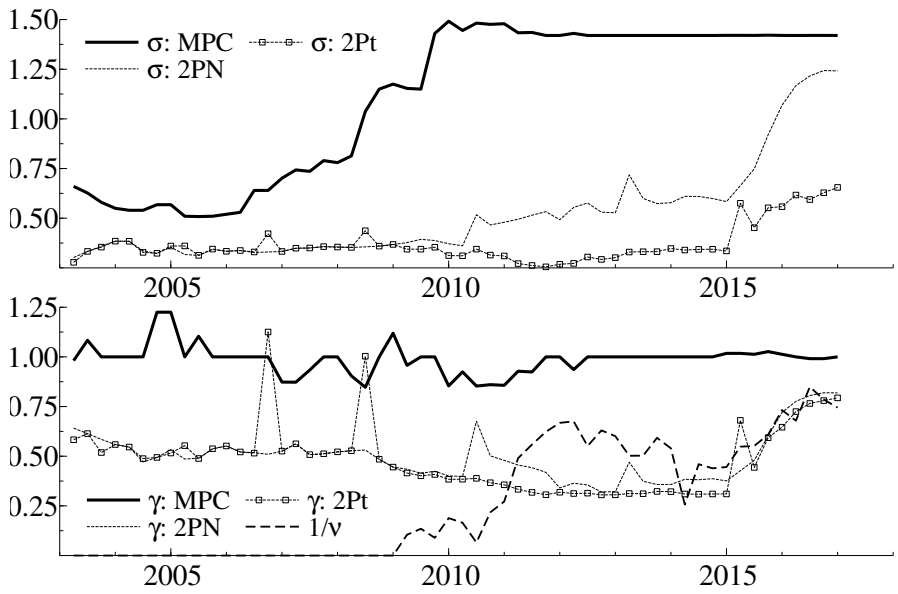
are moved to the censored part of the distribution, weakening their influence on the parameter. The skew parameter also rose in the aftermath of the crisis and has drifted back, as censoring is able to do more work.

Figure 9 also shows the parameters for the two-piece t distribution. Here the rise in σ is even slower and more modest than with the two-part normal distribution. This is because the increased frequency of large errors is initially accommodated by a falling number of degrees of freedom (rising value of $1/\nu$). There is also much more movement in the skew parameter. Only late in the period does recursive estimation suggest that the spread is better explained by the value of σ than a low value of ν.[25]

The most striking feature of the inflation parameters in Figure 10 is that throughout the MPC sets a value of σ which is large relative to that implied by the data. The MPC also saw very little skew. In contrast, the data-based parameters show a clear upside risk. A likely explanation of this is that the MPC bases its value of σ on past errors on the assumption that there is no skew present. The positive skew observed with, say, the fitted two-piece normal distribution is indicative that upside risks are much greater than downside risks for inflation. In turn this reflects the sticky downward nature of many prices and the rarity with which inflation rates below zero are observed. With the fitted parameters it is not until the very low inflation rates of 2015-2016 influence the calculations that we see the skew rises closer to 1 and the pattern of inflation errors is instead explained by a higher value of σ. But even here the difference between the two-part t and the two-part normal is clear. The two-part t prefers to rely on a

[25] As can be seen from Figure 9 the degrees of freedom parameter ν does temporarily, in the aftermath of the financial crisis, fall to unity and below. Given we define our censored densities between finite intervals, infinite moments, as observed for Cauchy distributions, are not a concern; for related discussion see Nadarajah & Kotz (2007).

Figure 10: Data-based and MPC Parameters for the Inflation Fan Chart (2 year ahead forecasts)



Note: Dates refer to when the two-year ahead forecast was made.

decreased number of degrees of freedom and puts less weight on the value of σ than does the two-piece normal distribution.

8.2.3 Out-of-sample comparison of the MPC with data-based censored density forecasts

Our comparison of forecast performance takes two forms. First we assess the number of occasions on which an outturn fell in the outer, censored part of the distribution. Secondly, we compare the forecast scores for the different distributions.

Table 4 reports the percentage of times that an outturn subsequently fell in the censored region of the MPC's or the data-based fan chart. We show results for: i) the full forecast period, ii) the period for which recursive (data-based) forecasts are possible and iii) the period since the financial crisis of 2008-09. Table 5 then reports the densities' average logarithmic scores, $\overline{LS}_A^C$ and $\overline{LS}_B^C$, over the latter two out-of-sample evaluation periods - when a comparison between the MPC and the data-based densities is possible. Recall the higher the score the better the relative performance of the density. The score for the best performing density is highlighted in bold in Table 5. Each data-based censored density forecast is also tested relative to the MPC, with an asterisk indicating rejection of the null of equal forecast performance at the 95% level using HAC standard errors.[26]

We draw our main conclusion from Table 4: the data-based censored forecasts are, on average over the longer out-of-sample window (from 2005q1-2018q3/q4), too narrow. This reflects overconfidence, a lack

[26]For completeness, Section A.4 of the online Appendix also provides some PITs plot for the data-based densities as well as the MPC. These are consistent with the analysis below using both scoring rules and looking at the proportion of outturns that fall inside the censored region, (17).

Table 4: Percentage of outturns falling in the 10% BCR censored region of the MPC and data-based two-year ahead inflation and GDP growth density forecast (dates refer to outturns, with the forecasts made two years previously)

	99q4-18q4	99q4-18q3		05q1-18q4	05q1-18q3		11q4-18q4	11q4-18q3	
	Inflation	GDP Growth		Inflation	GDP Growth		Inflation	GDP Growth	
		2nd	Latest		2nd	Latest		2nd	Latest
MPC	13	13	11	18	16	16	3	4	0
2Pt	–	–	–	41	22	22	38	4	0
2PN	–	–	–	43	29	23	38	17	0
N	–	–	–	29	23	23	41	21	11
t	–	–	–	41	31	29	14	7	0

of perceived risk - as more than 20%, and up to 40%, of outturns subsequently fell outside the data-based 90% BCRs. Calibration is better, i.e. closer to 10%, for the MPC forecasts but still too high (over this period from 2005, although it is better over the full-sample from 1999). The tendency for the BCRs, especially for the data-based densities, to be too narrow is much greater for inflation than GDP. It is also greater in the earlier part of the evaluation period than the latter period since the financial crisis. This is seen by, in general, the lower proportion of outturns falling in the censored region since 2011q4 especially for GDP growth (when measured using the latest vintage data). Indeed, except for the normal density, Table 4 shows that no GDP outturns fell in the censored region of any of the other densities from 2011q4. So, since the financial crisis the BCR bounds from both the data-based and MPC densities widened but for GDP growth by too much; too few (latest vintage) GDP outturns are subsequently censored except when using the (one-piece) normal. Figures A17 -A20 in AppendixA.4.3 illustrate these features visually for the 2Pt and 2PN density forecasts for inflation and GDP (latest vintage); Figures 2 and 3 above showed analogous plots for the MPC.

The fact that too few GDP outturns fell in the censored region since 2011q4 of course needs qualification. As discussed earlier, with (*big*) recessions occurring about every seventeen years (roughly the length of our full sample) and assuming that these are the main cause of large forecast errors, we need one recession in seventeen years of GDP data to get a good impression of censoring. In such circumstances, and also bearing in mind the Monte Carlo evidence that estimation can be problematic in small-samples especially of only about $T = 40$ observations, judgement is likely to work better. The implication of this is also that, as seen in Table 4, we should expect a very low proportion of outturns to fall in the MPC's censored region when looking at the sub-period 2011q4-2018q3.

Table 5 complements this discussion of coverage by reporting the censored logarithmic scores over the two more recent evaluation periods. The ranking of the different forecasts proves sensitive to whether we use $\overline{LS}_A^C$ or $\overline{LS}_B^C$. Recall $\overline{LS}_A^C$ penalises outlying observations, that fall in the censored regions, more heavily than $\overline{LS}_B^C$ that does not take a stance on the relative frequency of outlying observations in the

left and right tails. This explains why, for a given forecast, $\overline{LS}_A^C \geq \overline{LS}_B^C$. Looking at inflation first, using $\overline{LS}_A^C$ we see that, consistent with Table 4, none of the data-based forecasts match the performance of the MPC. But these gains for the MPC are not statistically significant. But for $\overline{LS}_B^C$ the data-based densities are more competitive and the normal density, in fact, delivers the highest score over both evaluation periods. This is at apparent odds with Table 4 where the normal density was seen to censor too many inflation outturns. We rationalise this by noting that by having narrower BCR intervals for inflation (than the MPC) the data-based densities do, in general, have poorer coverage rates. They censor too many observations. But when an inflation outturn does fall within its narrower BCR a higher score is awarded than for the MPC with its wider BCR intervals. The penalty for placing too many observations outside the BCR interval is weaker for $\overline{LS}_B^C$ than $\overline{LS}_A^C$ given that the cost is capped at $ln(0.1)$ for $\overline{LS}_B^C$. In large(r) samples, as proper scoring rules, both censored scoring rules would reward correct coverage.

For GDP growth when using the latest vintage estimates, Table 5 shows the MPC densities to be superior than the data-based densities under both $\overline{LS}_A^C$ and $\overline{LS}_B^C$. These gains are statistically significant. Although the second release is not the stated target for the MPC, the importance of data revisions is revealed by observing that when using the second release GDP data to define the outturns, the two-piece normal does outperform the MPC in three of the four cases considered in Table 5. These gains are not, however, statistically significant.

The averaged scores reported in Table 5 may, of course, mask temporal changes in absolute and relative forecast performance. Accordingly, to provide an indication of the relative and potentially time-varying performance of the five forecasts, we looked at their quarter-by-quarter censored log scores. For space reasons, results are shown in Section A.4 of the online Appendix. But summarising, we found that the data-based forecasts have a more volatile performance over time. They do especially poorly, relative to the MPC, when inflation or GDP growth peaks or troughs. But they often perform better during more stable periods.

This leads us to conclude not that the data-based censored density forecasts developed here are superior to judgement-based forecasts, like those from the MPC, - in fact the reverse holds in general - but that they offer a helpful benchmark. It is by comparison with these data-based alternatives that the judgements made by the MPC about the parameters of its forecast densities can be understood. We therefore encourage their routine production and analysis, as an input into subsequent discussions.

Table 5: Out-of-sample average censored log scores, $\overline{LS}_A^C$ and $\overline{LS}_B^C$

	Evaluation: 2005q1-2018q3/q4						Evaluation: 2011q4-2018q3/q4					
	$\overline{LS}_A^C$			$\overline{LS}_B^C$			$\overline{LS}_A^C$			$\overline{LS}_B^C$		
	Inflat	GDP Growth		Inflat	GDP Growth		Inflat	GDP Growth		Inflat	GDP Growth	
		2nd	Latest		2nd	Latest		2nd	Latest		2nd	Latest
MPC	**-1.60**	-1.70	**-1.68**	-1.47	-1.61	**-1.57**	**-1.60**	**-1.68**	**-1.44**	-1.58	-1.66	**-1.44**
2Pt	-1.88	-1.71	-1.92*	-1.47	-1.62	-1.76*	-2.29	-1.85*	-1.81*	-1.61	-1.84*	-1.81*
2PN	-1.75	**-1.63**	-1.82*	-1.48	**-1.52**	-1.68*	-2.03	-1.72	-1.73*	-1.59	**-1.65**	-1.64*
N	-1.70	-1.75	-1.87*	**-1.40**	-1.55	-1.68*	-1.81	-1.81*	-1.72*	**-1.53**	-1.67	-1.64*
t	-1.70	-1.76	-1.88*	-1.52	-1.61	-1.73*	-1.82	-1.86	-1.72*	-1.75	-1.81	-1.72*

Notes: The score for the best performing density is highlighted in bold. Each data-based censored density forecast is tested relative to the MPC, with an asterisk indicating rejection of the null of equal censored forecasting performance at the 95% level using HAC standard errors. Dates for the two evaluation periods refer to outturns, with the forecasts made two years previously

9 Conclusion

This paper considers a hitherto largely overlooked feature of the Bank of England's published fan charts. The MPC at the Bank of England, in effect, publish "censored" density forecasts that do not take any view on the outer ten percent of the distribution beyond saying that it does not overlap with the inner ninety per cent. The probabilities in the ten percent tails are unknown, they are not specified: they may well be drawn from a different (perhaps unknown or unspecified) distribution to the inner ninety percent. While there is no reason, in other applications, to fix the censoring at ten percent, what the Bank of England appear to be providing are density forecasts that communicate the known unknowns but also acknowledge the (possibility of) unknown unknowns. Indeed, an important question for future research, in anticipating applications beyond the MPC, is whether the degree of censoring should vary over time. This could reflect the judgement of the forecaster - at times when the forecaster is especially uncertain about their probability forecasts, they may choose to censor more than ten percent of their density forecast. By setting the censoring at ten percent the MPC are stating that there is a one in ten chance that the unexpected happens; although their use of BCRs and asymmetric forecast densities means that this ten percent need not be evenly split between the left and right tails of the forecast density. The MPC could communicate this directly - if they wished to alert the public to upside or downside *uncertainties*, as opposed to *risks*.

We examine the consequences of censoring both for estimation of the parameters of the density function of past forecast errors - which the MPC use as the basis for constructing their forecasts - and for *ex post* evaluation of the MPC's forecasts.

Accordingly, we propose and then evaluate, through Monte Carlo, a new fixed point estimator that fits a potentially skewed and fat tailed density to the inner observations but does not take a view on what distribution the outer observations come from. Our estimator is relevant to any researcher with a small sample of data who is concerned that the outtermost observations may be drawn from a distribution different from that defining the central observations. More specifically, we hope that our estimator will become increasingly relevant for the MPC and the Bank of England as sample sizes are now large enough, according to our Monte Carlo evidence, to estimate censored skewed and fat tailed densities more reliably than was possible in our out-of-sample experiments which had to split the available sample.

In re-evaluating the MPC's density forecasts, but for the first time acknowledging the censoring, we find that the MPC's two-year ahead inflation densities are in fact better calibrated than if the censoring is (incorrectly) ignored. Comparison with data-based censored forecasts also reveals the value of the MPC's judgement about the width of the censoring bounds - a measure of risk. The importance of acknowledging the censoring when fitting densities to past forecast errors is seen via comparison with uncensored densities. Censored density forecasts, especially for GDP growth, are less skewed: in effect,

the skew present in the uncensored density is a consequence of forcing all observations to be drawn from the same (potentially fat tailed and skewed) density. Given the importance of assessments of skew for statements about the balance of risks in the macroeconomy (e.g. see Adrian et al. (2019)), this paper demonstrates that the choice of statistical estimator used to produce the density forecast is more than a dry statistical issue.

References

Adrian, T., Boyarchenko, N. & Giannone, D. (2019), 'Vulnerable growth', *American Economic Review* **109**(4), 1263–89.

Alessi, L., Ghysels, E., Onorante, L., Peach, R. & Potter, S. (2014), 'Central bank macroeconomic forecasting during the global financial crisis: The european central bank and federal reserve bank of new york experiences', *Journal of Business and Economic Statistics* **32**(4), 483–500.

Arellano-Valle, R., Gomez, H. & Quintana, F. (2005), 'Statistical inference for general class of asymmetric distributions', *Journal of Statistical Planning and Inference* **128**, 427–443.

Arnold, B. & Groeneveld, R. (1995), 'Measuring skewness with respect to the mode', *American Statistician* **49**, 34–38.

Askanazi, R., Diebold, F. X., Schorfheide, F. & Shin, M. (2018), 'On the comparison of interval forecasts', *Journal of Time Series Analysis* **39**(6), 953–965.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**(2), 171–178.

Azzalini, A. & Arellano-Valle, R. (2013), 'Maximum penalized likelihood estimation for skew-normal and skew-t distributions.', *Journal of Statistical Planning and Inference* **143**, 419–433.

Berkowitz, J. (2001), 'Testing Density Forecasts with Applications to Risk Management', *Journal of Business and Economic Statistics* **19**, 465–474.

Boero, G., Smith, J. & Wallis, K. (2011), 'Scoring rules and survey density forecasts', *International Journal of Forecasting* **27**(2), 379 – 393.

Boneva, L., Fawcett, N., Massolo, R. & Waldron, N. (2018), 'Evaluating UK Point and Density Forecasts from an Estimated DSGE Model: the Role of Off-model Information over the Financial Crisis', *International Journal of Forecasting* **35**, 100–120.

Britton, E., Fisher, P. & Whitley, J. (1998), 'The Inflation Report Projections: understanding the Fan Chart', *Bank of England Quarterly Bulletin* **Q1**, 30–37.

Chan, M.-H. & Shao, Q.-M. (1999), 'Monte Carlo estimation of Bayesian credible and HPD intervals', *Journal of Computational and Graphical Statistics* **8**, 69–92.

Christoffersen, P. (1998), 'Evaluating interval forecasts', *International Economic Review* **39**, 841–862.

Clements, M. P. (2004), 'Evaluating the Bank of England density forecasts of inflation', *The Economic Journal* **114**(498), 844–866.

de Roon, F. & Karehnke, P. (2017), 'Addendum: A Simple Skewed Distribution with Asset Pricing Applications', *Review of Finance* **21**(6), 2401–2401.

Diebold, F., Gunther, T. & Tay, A. (1998), 'Evaluating Density Forecasts with Applications to Financial Risk Management', *International Economic Review* **39**, 863–883.

Diks, C., Panchenko, V. & van Dijk, D. (2011), 'Likelihood-based Scoring Rules for comparing Density Forecasts in Tails', *Journal of Econometrics* **163**, 215–230.

Dowd, K. (2007), 'Too good to be true? the (in)credibility of the uk inflation fan charts', *Journal of Macroeconomics* **29**, 91–102.

Dowd, K. (2008), 'The gdp fan charts: an empirical evaluation', *National Institute Economic Review* **203**(1), 59–67.

Elder, R., Kapetanios, G., Taylor, T. & Yates, T. (2005), 'Assessing the MPC's Fan Charts', *Bank of England Quarterly Bulletin* **45**, 326–348.

Ericsson, N. R. (2002), Predictable uncertainty in economic forecasting, *in* M. Clements & D. Hendry, eds, 'A Companion to Economic Forecasting', Blackwell, Oxford, chapter 2, pp. 19–44.

Fechner, G. (1897), *Kollektivemasslehre*, Engelmann, Leipzig.

Fernandez, C. & Steel, M. (1998), 'On Bayesian Modelling of Fat Tails and Skewness', *Journal of the American Statistical Association* **93**, 359–371.

Galbraith, J. & van Norden, S. (2012), 'Assessing gross domestic product and inflation probability forecasts derived from Bank of England fan charts', *Journal of the Royal Statistical Society Series A* **175**(3), 713–727.

Galvao, A. & Mitchell, J. (2019), 'Measuring Data Uncertainty: An Application using the Bank of England's "Fan Charts" for Historical GDP Growth', *ESCoE Discussion Paper 2019-08* .

Gebetsberger, M., Messner, J. W., Mayr, G. J. & Zeileis, A. (2018), 'Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood', *Monthly Weather Review* **146**(12), 4323–4338.

Gneiting, R. & Raftery, A. (2007), 'Strictly Proper Scoring Rules; Prediction and Estimation', *Journal of the American Statistical Association* **102**, 359–378.

Gneiting, R. & Ranjan, R. (2011), 'Comparing Density Forecasts using Threshold and Quantile-weighted Proper Scoring rules', *Journal of Business and Economic Statistics* **29**, 411–422.

Granger, C. W. J. & Pesaran, M. H. (2000), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting* **19**, 537–560.

Hackworth, C., Raidia, A. & Roberts, N. (2013), 'Understanding the MPCs Forecast Performance since Mid-2010', *Bank of England Quarterly Bulletin* .

Haldane, A. G. (2012), Tails of the unexpected. Speech given at "The Credit Crisis Five Years On: Unpacking the Crisis", conference held at the University of Edinburgh Business School, 8-9 June.

Hall, S. G. & Mitchell, J. (2007), 'Combining density forecasts', *International Journal of Forecasting* **23**, 1–13.

Hansen, B. E. (1997), 'Approximate asymptotic p values for structural-change tests', *Journal of Business and Economic Statistics* **15**(1), 60–67.

Independent Evaluation Office (2015), Evaluating Forecast Performance. Bank of England.

Knüppel, M. (2014), 'Efficient estimation of forecast uncertainty based on recent forecast errors', *International Journal of Forecasting* **30**(2), 257–267.

Knüppel, M. (2018), 'Forecast-error-based estimation of forecast uncertainty when the horizon is increased', *International Journal of Forecasting* **34**(1), 105–116.

Mitchell, J. (2005), 'The National Institute density forecasts of inflation', *National Institute Economic Review* **193**(1), 60–69.

Mitchell, J. & Hall, S. G. (2005), 'Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR "fan" charts of inflation', *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.

Nadarajah, S. & Kotz, S. (2007), 'A skewed truncated Cauchy distribution with applications in economics', *Applied Economics Letters* **14**(13), 957–961.

Pesaran, M. H. & Timmermann, A. (2007), 'Selection of estimation window in the presence of breaks', *Journal of Econometrics* **137**(1), 134–161.

Reifschneider, D. L. & Tulip, P. (2019), 'Gauging the Uncertainty of the Economic Outlook Using Historical Forecasting Errors: The Federal Reserve's Approach', *International Journal of Forecasting* . Forthcoming.

Sartori, N. (2006), 'Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions', *Journal of Statistical Planning and Inference* **136**(12), 4259 – 4275.

Smith, R. (1985), 'Maximum Likelihood Estimation in a Class of Non-regular Cases', *Biometrika* **72**, 67–90.

Stockton, D. (2013), Review of the Monetary Policy Committee's Forecasting Capability. Bank of England.

Wallis, K. F. (1989), 'Macroeconomic Forecasting: A Survey', *The Economic Journal* **99**(394), 28–61.

Wallis, K. F. (1999), 'Asymmetric Density Forecast of Inflation and the Bank of England Fan Chart', *National Institute Economic Review* **167**, 106–112.

Wallis, K. F. (2003), 'Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts', *International Journal of Forecasting* **19**, 165–175.

Wallis, K. F. (2004), 'An assessment of Bank of England and National Institute inflation forecast uncertainties', *National Institute Economic Review* **189**, 64–71.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**, 817–838.

Woodroofe, M. (1972), 'Maximum Likelihood Estimation of a Translation Parameter of a Truncated Distribution', *The Annals of Mathematical Statistics* **43**, 113–122.

A Online Appendix: Supplementary Results for:

Forecasting with Unknown Unknowns: Censoring and Fat Tails on the Bank of England's Monetary Policy Committee by Mitchell and Weale

A.1 Estimation of skew densities

In order to explore further the suitability of the two-piece t and normal distributions we consider the general family of skew distribution parameterisations defined in Arellano-Valle et al. (2005) and Rubio and Steel (2014).[27] Like the two-piece normal of Fechner (1897), this family of distributions involves joining two distributions, but not necessarily normal, with different scale parameters σ_1 and σ_2 on either side of the location parameter, μ. Specifically, Arellano-Valle et al. (2005) reparameterise these two scale parameters in terms of a common scale, σ, and a skewness parameter, α, and define the family of distributions as:

$$f(y_t|\mu,\sigma,\alpha) = \frac{2}{\sigma\left(a(\alpha)+b(\alpha)\right)} f\left(\frac{y_t-\mu}{\sigma b(\alpha)}\right) \text{ if } y_t < \mu \tag{A.1}$$

$$f(y_t|\mu,\sigma,\alpha) = \frac{2}{\sigma\left(a(\alpha)+b(\alpha)\right)} f\left(\frac{y_t-\mu}{\sigma a(\alpha)}\right) \text{ if } y_t \geq \mu \tag{A.2}$$

where f is a symmetric density and $a(\alpha)$ and $b(\alpha)$ are known and positive asymmetry functions. Asymmetries are introduced when $a(\alpha) \neq b(\alpha)$.

A leading specific density within this family (when $a(\gamma)=\gamma$, $b(\gamma)=1/\gamma$, for $\gamma>0$ and $f(.)$ is the t density), that we focus on in the main paper, is the two-piece t distribution described by Fernandez & Steel (1998)[28]:

$$f(y_t|\mu,\sigma,\gamma) = \frac{2}{\sigma\left(\gamma+1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \left[1+\frac{(y_t-\mu)^2}{\gamma^2\nu\sigma^2}\right]^{-(\nu+1)/2} \text{ if } y_t < \mu \tag{A.3}$$

$$f(y_t|\mu,\sigma,\gamma) = \frac{2}{\sigma\left(\gamma+1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \left[1+\frac{\gamma^2(y_t-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \text{ if } y_t \geq \mu. \tag{A.4}$$

This estimates, as well as the location and scale parameters, the number of degrees of freedom of the t-distribution.

Generalisations of (A.1)-(A.2) involve introducing additional (shape) parameters; see Rubio and Steel (2015). Rubio and Steel's (2015) five-parameter double two-piece distribution (DTP) uses different scale but also different shape parameters either side of the mode, μ. The DTP family contains the original

[27] Arellano-Valle et al. (2005) generalise Mudholkar and Hutson (2000) who introduced the so-called epsilon-skew-normal family of densities. This family reparameterises Fechner (1897) so that the two piece normal is re-expressed in terms of an explicit skewness parameter. When this parameter equals zero, the epsilson-skew normal density reduces to the normal density.

[28] This is an instance of the so-called Two-Piece Scale (TPSC) family of densities introduced by Rubio and Steel (2015) when $a(\alpha)=\sigma_1/\sigma$, $b(\alpha)=\sigma_2/\sigma$; σ_1 and σ_2 denote the scale of each of the two distributions being joined.

two-piece densities as a subclass, as well as a four-parameter distribution (DTSH) that varies only the shape on each side of the mode. Rubio and Steel (2015) define the DTP as:

$$f(y_t|\mu,\sigma_1,\sigma_2,\delta_1,\delta_2) = \frac{2\varepsilon}{\sigma_1} f\left(\frac{y_t-\mu}{\sigma_1};\delta_1\right) \text{ if } y_t < \mu \tag{A.5}$$

$$f(y_t|\mu,\sigma_1,\sigma_2,\delta_1,\delta_2) = \frac{2(1-\varepsilon)}{\sigma_2} f\left(\frac{y_t-\mu}{\sigma_2};\delta_2\right) \text{ if } y_t \geq \mu \tag{A.6}$$

where

$$\varepsilon = \frac{\sigma_1 f(0;\delta_2)}{\sigma_1 f(0;\delta_2) + \sigma_2 f(0;\delta_1)}; \tag{A.7}$$

or

$$f(y_t|\mu,\sigma,\gamma,\delta_1,\delta_2) = \frac{2}{\sigma c(\gamma,\delta_1,\delta_2)} f(0;\delta_2) f\left(\frac{y_t-\mu}{\sigma b(\gamma)};\delta_1\right) \text{ if } y_t < \mu \tag{A.8}$$

$$f(y_t|\mu,\sigma,\gamma,\delta_1,\delta_2) = f(0;\delta_1) f\left(\frac{y_t-\mu}{\sigma a(\gamma)};\delta_2\right) \text{ if } y_t \geq \mu \tag{A.9}$$

where

$$c(\gamma,\delta_1,\delta_2) = b(\gamma) f(0;\delta_2) + a(\gamma) f(0;\delta_1). \tag{A.10}$$

Special cases of DTP include the distribution considered by Zhu and Galbraith (2010) that allows the number of degrees of freedom in (A.3)-(A.4) to be different on each side of the mode. Note also how the DTP includes four-parameter two piece scale (TPSC) distributions, such as the two-piece t distribution seen in (A.3)-(A.4), by setting $\delta_1 = \delta_2 = \delta$, when $f(.)$ is a t density. Rubio and Steel (2015) also consider the subfamily of two-piece shape (TPSH) distributions obtained when $\sigma_1 = \sigma_2 = \sigma$ in (A.5)-(A.6). This produces distributions with different shape parameters in each direction; following Rubio and Steel (2015) let ζ explain the difference between the shapes on either side of the model, where $\delta_1/\delta_2 = b^*(\zeta)/a^*(\zeta)$ and $\{a^*(\zeta), b^*(\zeta)\}$ are positive differentiable functions.

We consider five-parameter DTP and four-parameter DPSC and TPSH distributions with $f(.)$ chosen to be the t density and the symmetric sinh-arcsinh (SAS) distribution of Jones and Pewsey (2009), denoted s_{JP} with asymmetry parameter ε. The SAS distribution allows for both heavier and lighter tails than the normal distribution which is a special case when $\delta_1 = \delta_2 = 1$ and $\gamma = 0$.

As a robustness check, we compare the in-sample fit of the two-piece t distribution with these other classes of distributions. As in Rubio and Steel (2015) we do so through classical information criteria (the AIC and BIC) based on the ML estimates. Estimation makes use of the `sn` package in R and R packages available at http://rpubs.com/FJRubio/DTP and http://rpubs.com/FJRubio/BTV. For comparison purposes, we also consider the skew normal distribution of Azzalini (1985) and the skew t distribution of Azzalini and Capitanio (2003).[29] The skew normal of Azzalini (1985) is defined by the density function

$$f(y_t|\mu,\sigma,\alpha) = \frac{2}{\sigma}\phi\left(\frac{y_t-\mu}{\sigma}\right)\Phi\left(\alpha\frac{y_t-\mu}{\sigma}\right) \tag{A.11}$$

[29] The skew t distribution of Azzalini and Capitanio (2003) has also found recent application in macroeconomics; e.g. see Adrian et al. (2019).

where ϕ and Φ denote the standard normal probability density function and distribution function, respectively, and α which regulates the skew or shape. The skew t of Azzalini and Capitanio (2003) is defined by the density function

$$f(y_t|\mu,\sigma,\alpha,\nu) = \frac{2}{\sigma} f(\frac{y_t-\mu}{\sigma}|\mu,\sigma,\nu) F(\alpha \frac{y_t-\mu}{\sigma} \sqrt{\frac{\nu+1}{\nu+\left(\frac{y_t-\mu}{\sigma}\right)^2}}|\mu,\sigma,\nu+1) \tag{A.12}$$

where f and F denote the Student t density function and distribution function, respectively with ν degrees of freedom. Again α regulates the shape; when $\alpha = 0$ the skew t reduces to the t and when $\alpha = 0$ and $\nu = \infty$ the density reduces to the Gaussian with mean μ and standard deviation σ. And we consider the Normal Laplace distribution of Ramierez-Cobo et al. (2010) which is the convulution of a normal distribution and a two-piece Laplace distribution with location 0 and two parameters α and β. The Normal Laplace density has heavier tails than the normal density.

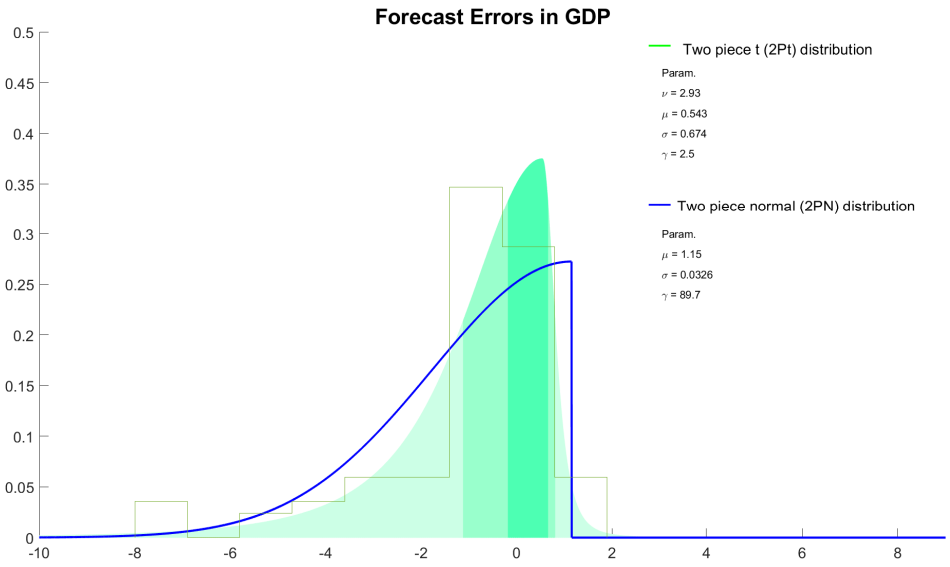
Table A1 shows the ML parameter estimates and the AIC and BIC values for these twelve density functions when fitted to the forecast error data considered in Section 3. Looking across all three forecast error series, we see that the two piece t fits the data competitively relative to the alternatives. While improvements in in-sample fit are achieved by the more flexible DTP and TPSC (with four or five parameters), the more parsimonious (three parameter) two piece t is always ranked in the top half of the twelve densities in terms of goodness of fit - according to both the AIC and BIC. The ML parameter estimates, across the different densities, also confirm the impression from Figures 2, 3 and Figure A1 (below) that asymmetries are most important for GDP growth. There is also, consistent with results in the main paper, evidence that allowing for fat tails improves fit. For both sets of GDP forecast errors, the two piece normal density (and the skew normal density of Azzalini (1985)) do not fit the data as well as the two piece t (and the skew t density of Azzalini and Capitanio (2003)). When using second release data, as in Figure A1, we also see that the skewed normal densities have divergent skew parameters, in contrast to the skewed t densities.

For inflation, Table A1 shows that the BIC, which favours parsimony, in fact selects a Gaussian density as the preferred density. But when using the AIC this Gaussian density ranks only eighth, with the asymmetric t and normal densities offering improvements in fit.

Table A1: GDP and inflation forecast error data (considered in Section 3): ML estimates of different skewed and fat tailed density functions and AIC and BIC values

	GDP final							
	AIC	BIC	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\gamma}$ or $\widehat{\alpha}$	$\widehat{\nu}$	$\widehat{\delta}$	$\widehat{\zeta}$
2Pt	294.57	303.90	0.03	1.01	1.30	2.51		
2PN	327.28	331.94	0.90	1.35	2.13			
DTP SAS	288.33	299.98	−1.13	15.45	−0.98		7.95	−0.96
DTP t	297.21	308.87	0.35	1.10	0.44	2.93	−0.08	
TPSC SAS	296.06	305.38	−0.16	0.52	0.22		0.52	
TPSH SAS	291.91	301.23	−0.21	0.62			0.60	−0.17
s_{JP}	293.91	303.23	−0.22	0.56	($\widehat{\varepsilon}$) −0.22		($\widehat{\beta}$) 0.55	
SN	304.15	311.14	1.46	3.00	−4.83			
St	293.12	302.44	0.50	1.33	−1.20	2.75		
Normal Laplace	294.58	303.90	0.65	4.76	0.62		($\widehat{\beta}$) 0.79	
N	327.28	331.94	−0.74	2.03				
t	327.28	331.94	−0.74	2.03		15.78		
	GDP second release							
	AIC	BIC	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\gamma}$ or $\widehat{\alpha}$	$\widehat{\nu}$	$\widehat{\delta}$	$\widehat{\zeta}$
2Pt	275.51	284.84	0.54	0.67	2.50	2.93		
2PN	281.08	288.08	1.14	0.29	10.00			
DTP SAS	271.82	283.47	−0.41	9.11	−0.95		9.13	−0.95
DTP t	276.20	287.86	0.46	0.98	0.63	26.06	−0.90	
TPSC SAS	275.46	284.78	0.56	0.60	0.75		0.60	
TPSH SAS	272.78	282.10	0.01	0.74			0.85	−0.39
s_{JP}	282.32	291.65	1.77	0.02	($\widehat{\varepsilon}$) − 6.64		($\widehat{\beta}$) 1.22	
SN	279.90	286.89	1.15	2.93	−311573.60			
St	275.07	284.39	0.85	1.78	−6.17	3.19		
Normal Laplace	275.08	284.41	0.48	5.24	0.58		($\widehat{\beta}$) 0.40	
N	319.97	324.63	−1.05	1.93				
t	319.97	324.63	−1.05	1.93		12.60		
	Inflation							
	AIC	BIC	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\gamma}$ or $\widehat{\alpha}$	$\widehat{\nu}$	$\widehat{\delta}$	$\widehat{\zeta}$
2Pt	247.60	256.97	−0.06	0.86	0.79	3.97		
2PN	247.73	254.76	−0.08	1.13	0.78			
DTP SAS	246.63	258.35	−0.40	0.54	−0.67		0.59	−0.23
DTP t	249.60	261.31	−0.05	0.88	−0.22	4.05	0.04	
TPSC SAS	244.66	254.04	−0.09	0.53	−0.24		0.64	
TPSH SAS	244.86	254.23	0.07	0.54			0.66	0.10
s_{JP}	262.09	271.46	−2.42	7.54	($\widehat{\varepsilon}$) 1.84		4.83	
SN	251.30	258.33	0.33	1.19	0.00			
St	248.12	257.50	−0.68	1.41	1.55	10.75		
Normal Laplace	248.40	257.77	0.13	0.96	1.06		($\widehat{\beta}$) 0.09	
N	249.30	253.99	0.33	1.19				
t	249.30	253.99	0.33	1.19		15.59		

Figure A1: GDP Growth (Second Release Outturns): Forecast Error Histogram and Uncensored Two Piece Normal and t Densities



Note: Second release GDP estimates used to define the 'outturn'. 76 observations used. The p-values from tests for the uniformity of the probability integral transforms, using an Anderson-Darling test, against the 76 observations are: 0.15 (2Pt), 0.00 (2PN), 0.00 (for a one-piece t) and 0.01 (for a one-piece normal).

A.2 Fitting uncensored and censored densities: robustness

Here we report supplementary results referred to in the main body of the paper.

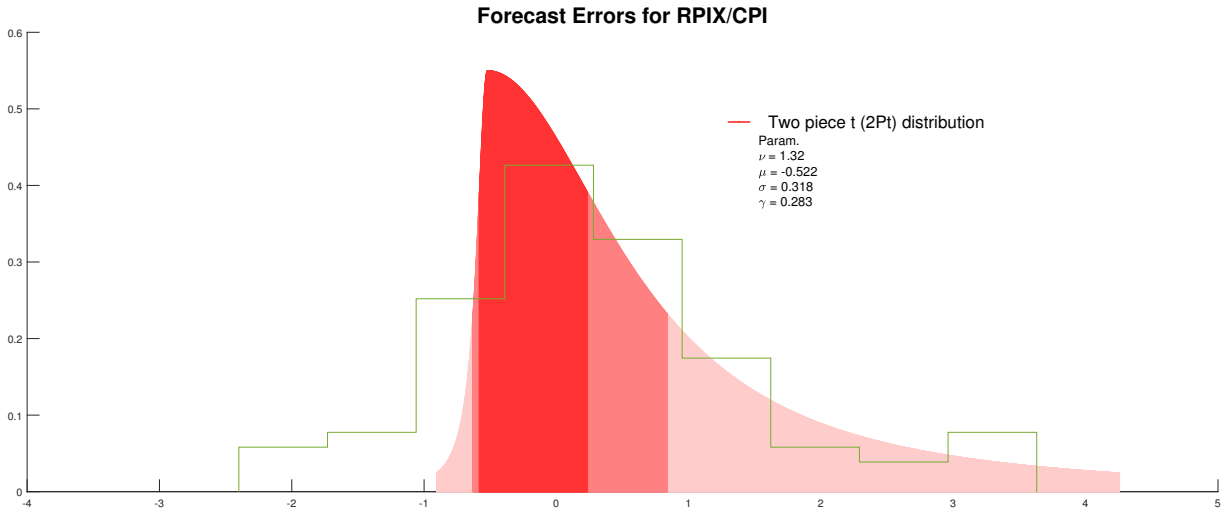
A.2.1 Use of second release estimates to define forecast errors

Figure A1 shows the uncensored two-piece t and normal densities fitted to forecast errors using the second release GDP growth data as the outturn. Comparison with Figure 5, that shows analogous densities but with outturns measured using "mature" estimates of GDP, reveals that data revisions matter. Figure A1 indicates more skew to the forecast errors when second release data are used as outturns. For the two-piece normal the skewness parameter diverges to 89.7. For the two-piece t, γ rises from 1.3 to 2.5.

A.2.2 Censored densities: use of L_B^C

Figures A2 and A3 show the consequences of fitting L_B^C rather than L_A^C to the inflation and GDP (final vintage) forecast errors. Here the quadratic criterion, P_r, also converges to a value of 0 but only for the two-piece t densities. For the two-piece normal densities, despite experimentation, it did not prove possible to obtain satisfactory estimation and $P_r > 0$ even as $r \to \infty$. The estimated densities failed to meet the requirements of a BCR (i.e. the probability density of being at either censoring point should be equal); we therefore do not report them. Recall L_A^C distinguishes the lower from the upper tail, with each having its own probability. This means that the process of fitting is likely to place some observations in

Figure A2: Inflation: Forecast Error Histogram and Censored Two Piece t Density using L_B^C



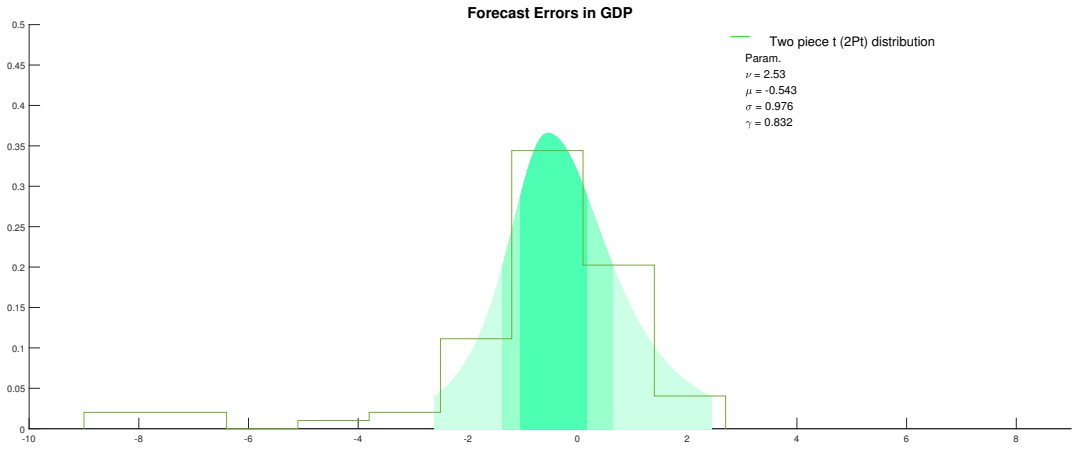
Note: RPIX forecasts (until Nov. 2003); CPI forecasts thereafter. 77 observations used. The p-values from tests for the uniformity of the probability integral transforms, using an Anderson-Darling test, are: 0.84 (2Pt). 10% of the observations fall outside 2Pt.

each tail rather than locating all the censored observations in only one of the tails as in L_B^C - our focus here. The expected number of observations in each tail depends on the skew parameter.

Figure A2 shows that for inflation the two-piece t fits the data with a very low number of degrees of freedom. We also observe, which is odd, that L_B^C places all of the censored observations in the left tail - and as a result, to fit the remaining errors, it requires a much more skewed density than L_A^C in Figure 2.

For GDP growth, Figure A3 indicates both less evidence for asymmetry relative to L_A^C (Figure 3) and that the nature of the observed asymmetry has switched from right to left skew. This is because the majority of the censored observations associated with the recession are now all in the lower tail. As a result, the distribution is more symmetric because no attempt is made to place any censored observations in the right-hand tail, as in Figure 3 using L_A^C. We also, find, however, that even when no effort is made to accommodate the recession (given that the recessionary data are censored), a low number of degrees of freedom is selected.

Figure A3: GDP (final vintage): Forecast Error Histogram and Censored Two Piece t Density using L_B^C



Note: Latest release GDP estimates of used. 76 observations used. The p-values from tests for the uniformity of the probability integral transforms, using an Anderson-Darling test, are: 0.91 (2Pt). 9% of the observations fall outside 2Pt.

A.2.3 Densities using specific windows of data

Figures A4-A7 follow in the spirit of practice at the Bank of England by plotting some illustrative (uncensored) densities fitted to specific (rolling) samples of forecast error data.

For inflation, Figures A4 and A5 consider, respectively, the first ten years of our forecast error data and the last ten years. For GDP growth, we select the sample period more carefully/subjectively, aware of the effects of the global financial crisis in 2008 on the GDP forecast errors. Accordingly, Figure A6 considers a sample of forecast error data before the crisis; while Figure A7 considers a sample after the crisis. Experimentation revealed that the choice of estimation window for these uncensored densities could have a large effect on the shape of the densities fitted to the GDP forecast errors. The dates in the figures refer to outturns, with the forecasts made two years previously.

Figure A4: Inflation: Forecast Error Histogram and Uncensored Two Piece Normal and t Densities fitted to Error data from 1999q4-2008q3

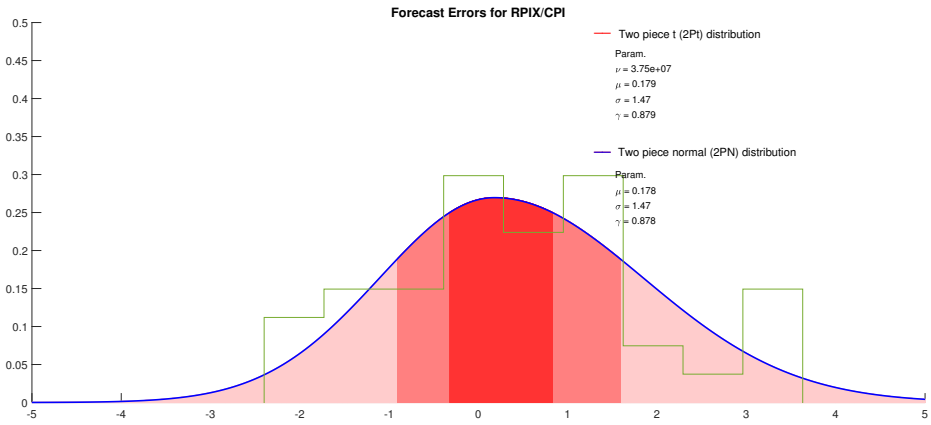


Figure A5: Inflation: Forecast Error Histogram and Uncensored Two Piece Normal and t Densities fitted to Error data from 2009q1-2018q4

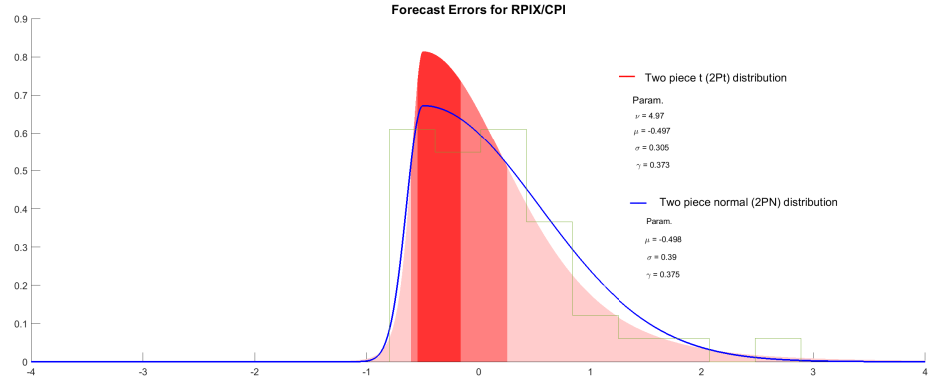


Figure A6: GDP growth: Forecast Error Histogram and Uncensored Two Piece Normal and t Densities fitted to Error data from 1999q4-2008q2

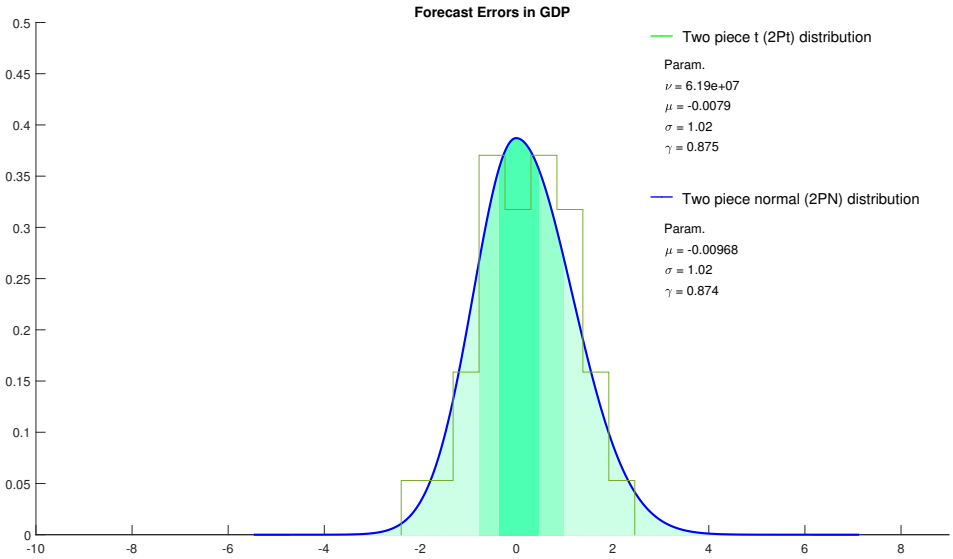
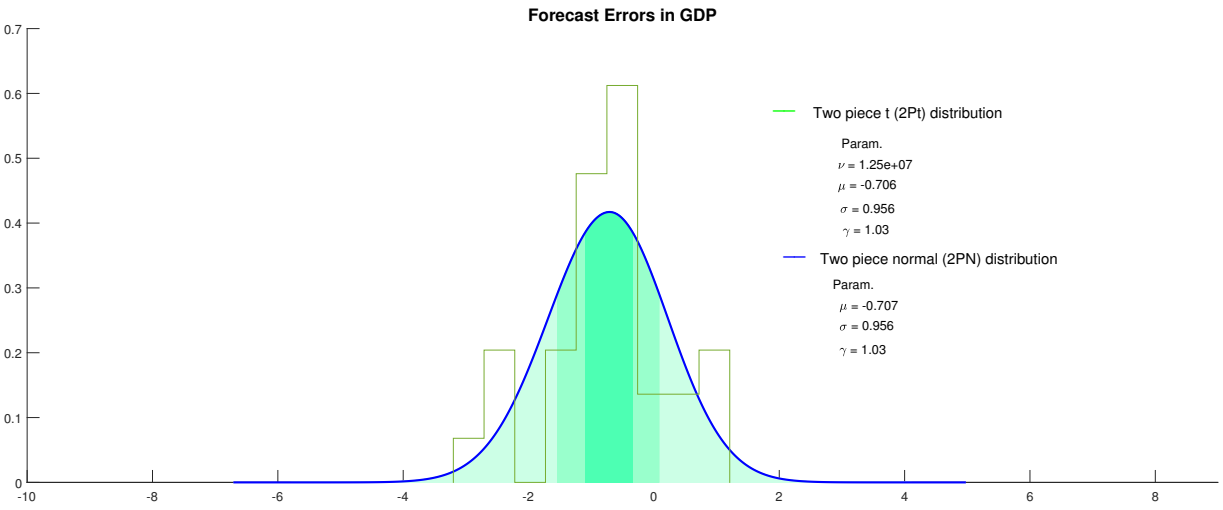


Figure A7: GDP growth: Forecast Error Histogram and Uncensored Two Piece Normal and t Densities fitted to Error data from 2011q2-2018q3



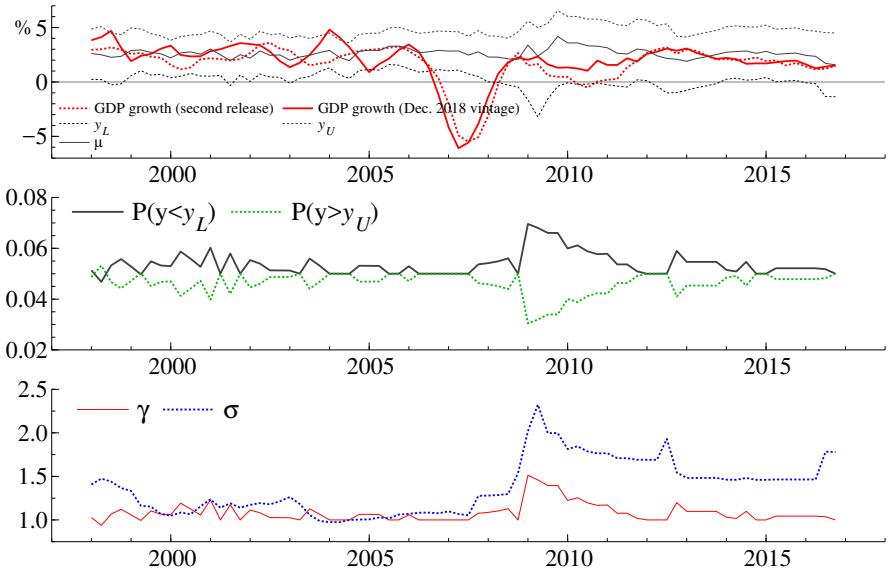
A.3 Further Properties of the MPC's Forecasts

A.3.1 Probabilities of Outturns Falling Above or Below the MPC's Censor Points

It is a matter of interpretation, formalised in the distinction between L_A^C and L_B^C in equations (5) and (8), but if the MPC took (or is assumed to have taken) a view on whether the unknown tail uncertainties (summing to 10%) were in the left or right hand tail, we can then sensibly compute the probabilities that $Y > y_U$ and $Y < y_L$. Recall for asymmetric densities, given y_U and y_L are defined as BCRs, these probabilities need not equal 5%.

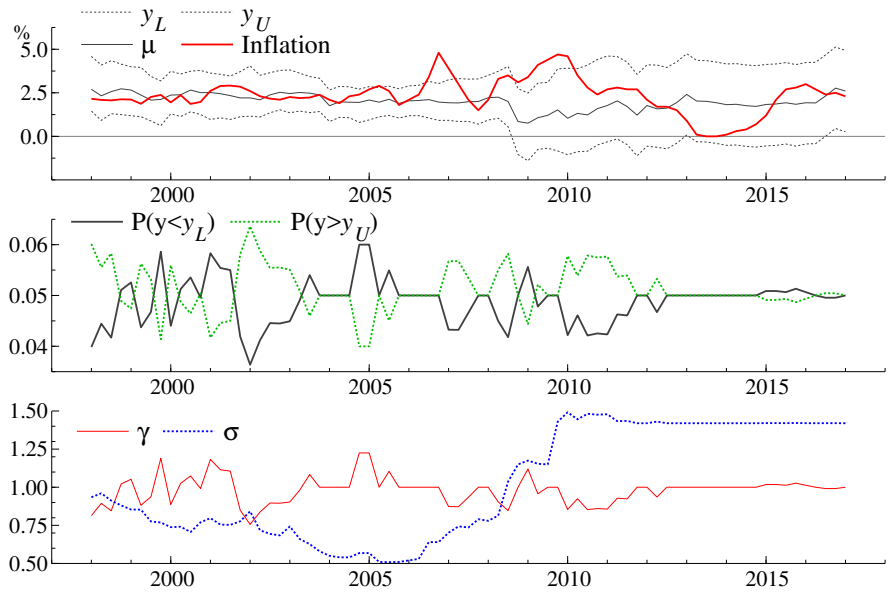
Figures A8 and A9 (focusing on the middle panels) show that these probabilities vary over time and often differ from 5%. The strongest departures from equal (5%) probabilities occur for GDP growth in the aftermath of the financial crisis. In 2009, for example, there was close to a 7% probability that GDP, two years ahead, fell below y_L.

Figure A8: Properties of the MPC's GDP growth forecasts: 10% BCR censoring thresholds



Note: μ=modal forecast; y_L=lower censor point; y_U=upper censor limit; *Growth (2nd)* = GDP Growth 2nd Estimate; *Growth (Dec 2018)* = GDP Growth December 2018 Vintage; σ=scale parameter; γ=skew parameter; dates relate to date of forecast

Figure A9: Properties of the MPC's inflation forecasts: 10% BCR censoring thresholds



Note: μ=modal forecast; y_L=lower censor point; y_U=upper censor limit; σ=scale parameter; γ=skew parameter; dates relate to date of forecast

A.3.2 Evaluation of interval forecasts defined using the central percentiles or the BCRs

Here we illustrate empirically that it matters whether interval forecasts are extracted from the MPC's density forecasts based on central percentiles or $k = 10$ BCRs. In Figures A10 to A12 we plot PIT histograms using both of these options for inflation and the two measures of GDP growth (final vintage and second release).

In each Figure, the top left plot (for comparison purposes) repeats the relevant panel from Figure 8 by showing histograms for the PITs, z_t. The top right plot is for the censored PITs: $\left\{\frac{z_{c,t}-z_{L,t}}{z_{U,t}-z_{L,t}}\right\}$. The bottom two panels plot the PITS having sorted them into $k = 10$ bins. In effect, this involves reducing the density forecast to ten interval forecasts. As anticipated in Section 8.1, we break the density forecast into interval forecasts based on both (central) percentiles and 10% BCRs: in each case, the PITs are arranged so that the far left bin indicates how many observations fell in the outermost 10% interval (both left and right tail, so for intervals based on percentiles this is PITs both between 0 and 0.05 and between 0.95 and 1); the bin to the immediate right of this plots the number of observations that fell between the 10% and 20% interval (both left and right tail, so for intervals based on percentiles this is PITs both between 0.05 and 0.1 and between 0.9 and 0.95), and so on, until the bin on the far right plots the number of outturns that fell in the innermost 10% interval (so for intervals based on percentiles this is PITs between 0.45 and 0.55).

If the MPC did not censor its density forecasts, the more any of these PITs histograms deviate from uniformity the weaker the evidence for correct forecast calibration. But, with censoring, only the plots on the right-hand-side of Figures A10 to A12 should be uniform. That is, non-uniformity of the PITs in the plots on the left could be a feature not of calibration failure, but of failing to account for the censoring in evaluation.

The uncensored PITs plot in the top left of Figure A10 - for inflation - looks familiar compared to figures in Independent Evaluation Office (2015) (see Box 4 on their pages 50-51). That is, relative to forecast there were too many high outturns for inflation; in other words, the MPC understated the probability of high inflation outturns, especially from 2007. But this specific plot, as discussed, does not acknowledge the censoring. When we correct for this and look instead at the top right plot the story seems to change. The top right histogram plot appears more uniform. This suggests that the MPC's forecasts were not so bad after all, when we rightly acknowledge the censoring. The bottom two plots also appear to indicate less bias, than the top left plot, although when we look at the bottom right plot there do appear to be too many observations falling in the centre of the density. We did subject each of the histograms in Figure A10 to Pearson chi-squared tests as discussed in Wallis (2003) and find p-values of 0.14 (top left), 0.85 (top right), 0.55 (bottom left) and 0.28 (bottom right). So these p-values do confirm our visual impression that calibration is better both in the top right than top left panel, but not

in the bottom left than bottom right (though the test in the bottom left, as in the top left, is incorrect, as it ignores the censoring).

Figures A11 and A12 plot the PITS for the MPC's GDP forecasts, defining errors against final vintage and second release GDP outturn data, respectively. Comparison of these two figures suggests that calibration remains imperfect, but slightly stronger, when measured against final release rather than second release GDP data. Comparing the bottom left and bottom right plots we also see that the shape of the histograms is sensitive to whether we define the ten intervals using percentiles or BCRs. There is tentative evidence that calibration is better when BCRs are used - as is the MPC's (albeit overlooked) intention. This is confirmed by p-values from the Pearson chi-squared test: 0.04 (bottom left) and 0.25 (bottom right) in Figure A11 and 0.00 (bottom left) and 0.23 (bottom right) in Figure A12.

Figure A10: MPC Inflation Forecast PITs (2 year ahead forecasts made from 1998q1-2017q1 of outturns in 1999q4-2018q4)

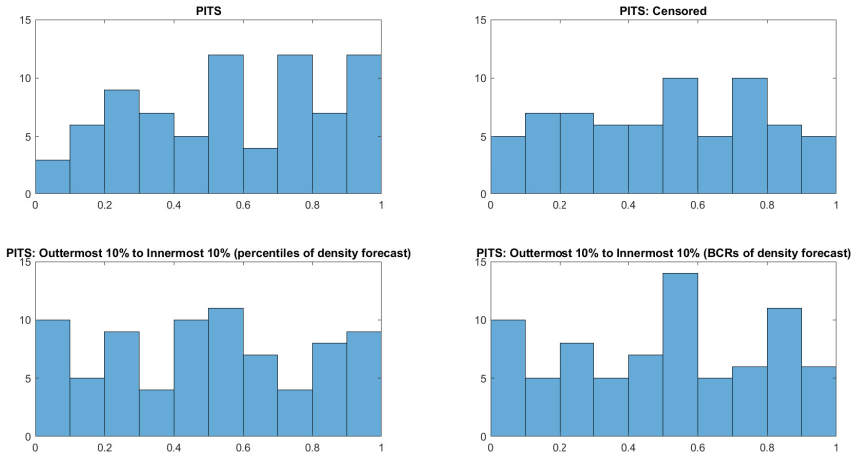


Figure A11: MPC GDP Growth Forecast PITs, final vintage (2 year ahead forecasts made from 1998q1-2016q4 of outturns in 1999q4-2018q3)

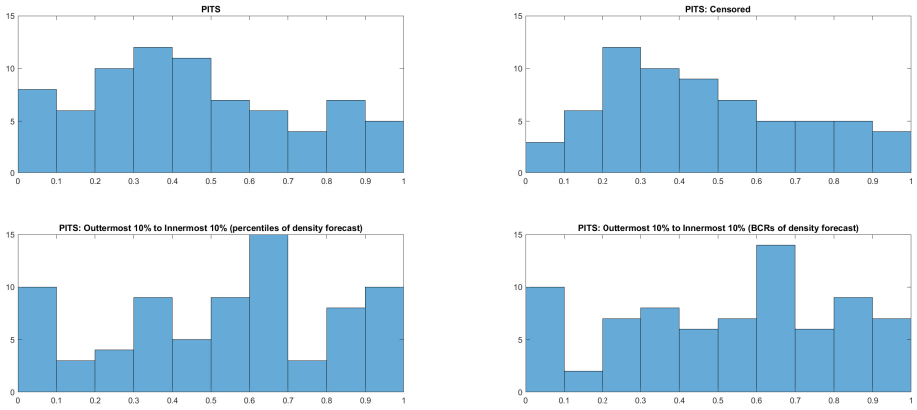


Figure A12: MPC GDP Growth Forecast PITs, second vintage (2 year ahead forecasts made from 1998q1-2016q4 of outturns in 1999q4-2018q3)

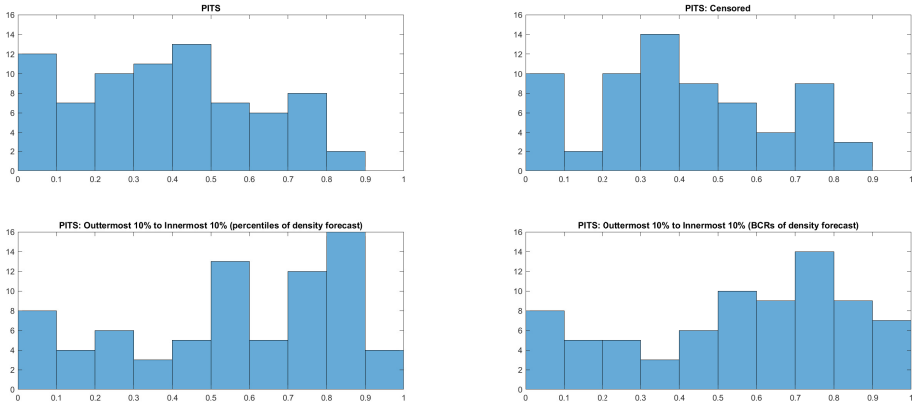
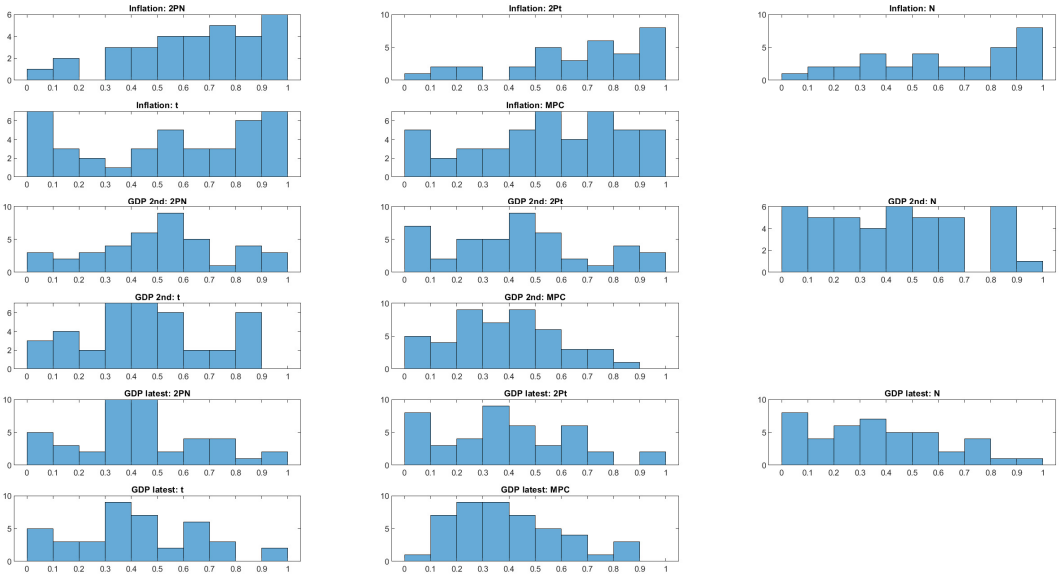


Figure A13: PIT histograms for the censored densities over the out-of-sample evaluation period, 2005q1-2018q3/q4



A.4 Additional out-of-sample results: 2005q1-2018q3/q4

A.4.1 PITs plots

Figure A13 plots histograms for the rescaled PITs, $\left\{\frac{z_{c,t}-z_{L,t}}{z_{U,t}-z_{L,t}}\right\}$, for the MPC and all four data-based censored forecasts over our (longer) out-of-sample period analysed in Tables 4 and 5. None of the histograms appear particularly uniform. But the MPC's inflation density forecasts do appear flatter. All of the data-based density forecasts, like the MPC, are unable to deliver uniform PITs for GDP growth. There remains a preponderance of outturns falling towards the left of the GDP density forecast, largely due to the failure to forecast (two years ahead) the global financial crisis and ensuing recession in the U.K..

A.4.2 Time-varying out-of-sample performance

To provide an indication of the relative and potentially time-varying performance of the five forecasts, Figures A14 to A16 plot their quarter-by-quarter censored logarithmic scores. To make it easier to observe when an outturn falls in the the censored region, we report results using $\overline{LS}_B^C$ - since this gives a (constant) score of $\log(0.1) = -2.3$ when the outturns falls in the censored region - irrespective of whether this is the upper or lower tail. We do this for inflation in Figure A14 and for GDP, against both measures of the outturn, in Figures A15 and A16.

The Figures show that the data-based forecasts are more volatile in terms of their performance over

time. They do especially poorly, relative to the MPC, when inflation or GDP growth peaks or troughs. But they often provide better performance during the more stable periods.

We observe that the MPC forecasts descend to the lower bound score, of -2.3, far less frequently than the data-based forecasts. In other words, out-of-sample the (subsequent) outturns fell in the censored region of the data-based density forecasts far more frequently than they did for the MPC density. This is consistent with the finding in Table 4 that outturns fall in the censored region much more than 10% of the time: the data-based densities censor too many observations out-of-sample. In contrast, the MPC densities, especially over the longer-sample, have a much better coverage rate, closer to 10%. This suggests that the data-based forecasts set too narrow a width for the 90% BCR. In contrast, no doubt reflecting their access to more up-to-date information than the data-based forecasts which use forecast error data at least two years out-of-date, the MPC appear better able to set the (time-varying, as seen in Figures 2 and 3) widths of the 90% BCR intervals.

Figure A14: Censored log scores ($LS^C_{B,t}$) of MPC and data-based density forecasts of inflation

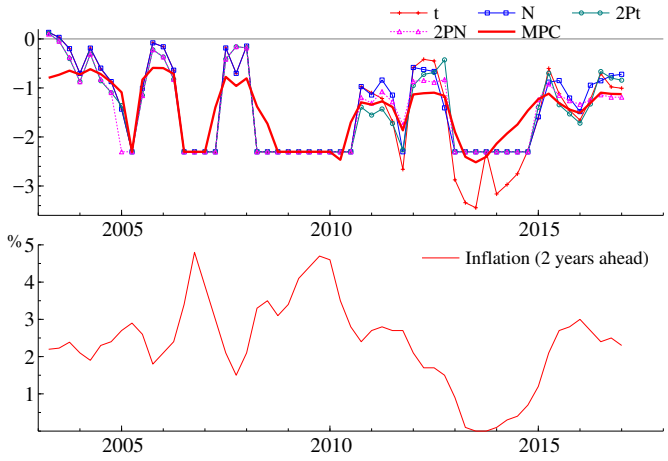


Figure A15: Censored log scores ($LS^C_{B,t}$) of MPC and data-based density forecasts of GDP growth (second estimate)

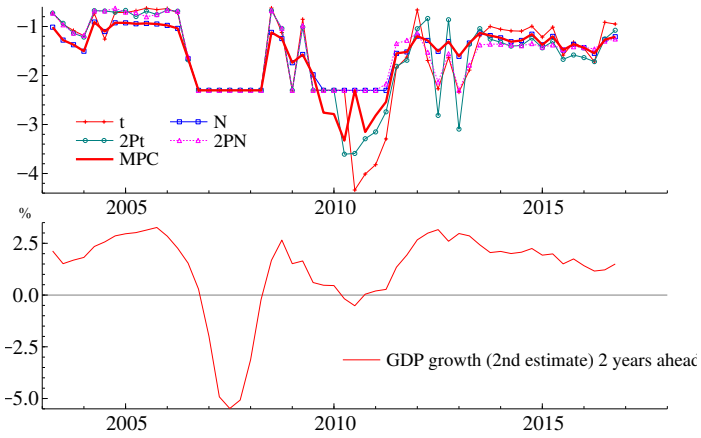
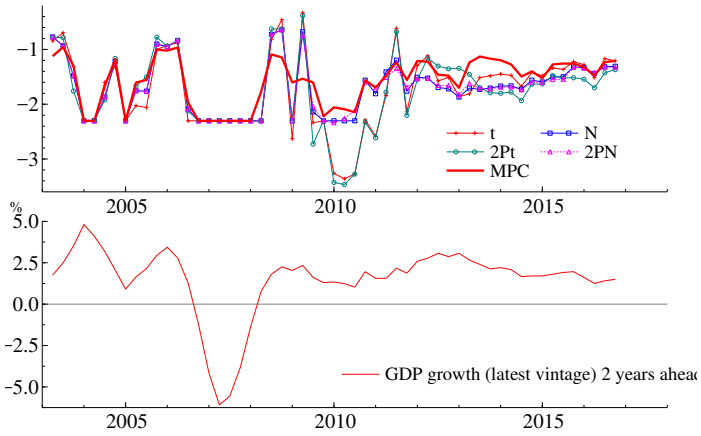


Figure A16: Censored log scores ($LS^C_{B,t}$) of MPC and data-based density forecasts of GDP growth (latest estimate)



A.4.3 Properties of the data-based censored forecasts

To understand the time-varying performance of the data-based forecasts when estimated recursively, as if in real-time, Figures A17 to A20 show the evolution over time of the BCRs, the underlying distributional parameters and the error associated with the modal point forecast. For comparative purposes, alongside the 2PN error plot we also indicate the Bank of England's (the MPC's) own forecasting error for its modal forecast. We show results both for inflation and GDP growth, for both 2Pt and 2PN, focusing on use of the latest or final vintage GDP to measure the outturns - given this is the MPC's stated preference.

Figure A17: Properties of the data-based censored forecasts for 2Pt GDP Growth (final vintage outturns): evolution over time of the 10% BCR censoring thresholds, distributional parameters and point forecasting error

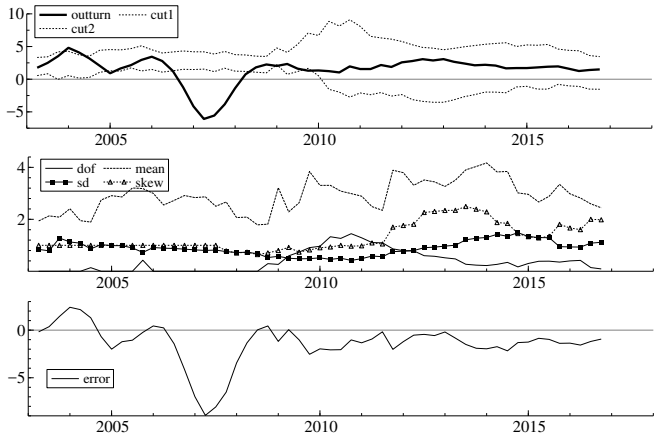


Figure A18: Properties of the data-based censored forecasts for 2PN GDP Growth (final vintage outturns): evolution over time of the 10% BCR censoring thresholds, distributional parameters and point forecasting error

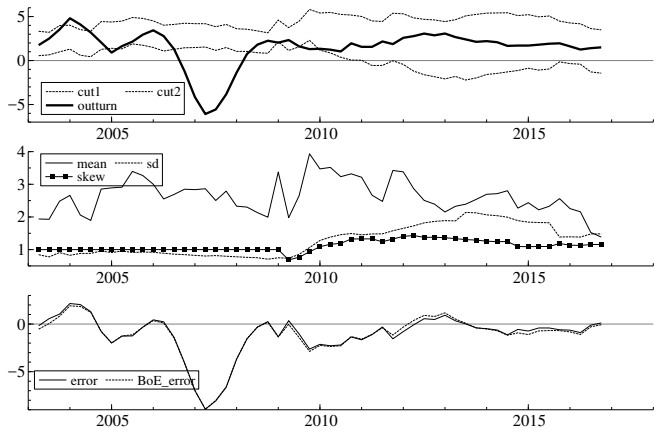


Figure A19: Properties of the data-based censored forecasts for 2Pt Inflation: evolution over time of the 10% BCR censoring thresholds, distributional parameters and point forecasting error

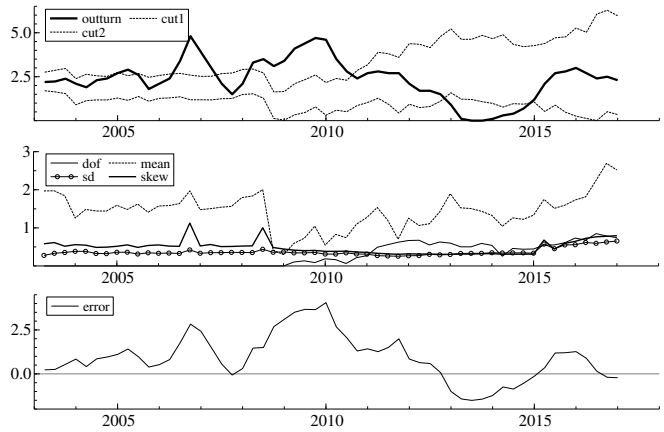
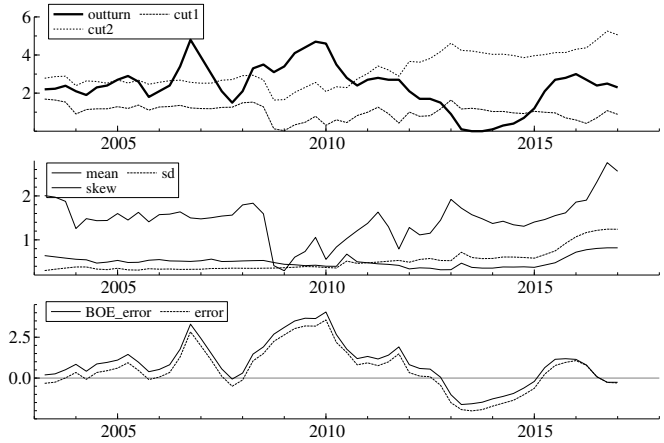


Figure A20: Properties of the data-based censored forecasts for 2PN Inflation: evolution over time of the 10% BCR censoring thresholds, distributional parameters and point forecasting error



A.5 Appendix References

Adrian, T., Boyarchenko, N. and Giannone, D. (2019), "Vulnerable Growth", *American Economic Review*, 109(4), 1263-89.

Arellano-Valle, R.B., Gómez, H. W. and Quintana, F. A. (2005), "Statistical inference for general class of asymmetric distributions". *Journal of Statistical Planning and Inference*, 128: 427-443.

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones", *Scandinavian Journal of Statistics*, 12(2), 171-178.

Azzalini, A. (2018), "Package 'sn' - The R Project for Statistical Computing". https://cran.r-project.org/web/packages/sn/sn.pdf

Azzalini, A. and Arellano-Valle, R.B. (2013), "Maximum penalized likelihood estimation for skew-normal and skew-t distributions". *Journal of Statistical Planning and Inference*, 143, 419-433.

Azzalini, A. and Dalla Valle, A. (1996), "The multivariate skew-normal Distribution". *Biometrika*, 83, 715-726.

Azzalini, A. and Capitanio, A. (1999), "Statistical applications of the multivariate skew-normal distribution". *Journal of the Royal Statistical Society: Series B*, 61(3), 579-602.

Azzalini, A. and Capitanio, A. (2003), "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution". *Journal of the Royal Statistical Society: Series B*, 65, 367-389.

Cabras, S., Racugno, W., Castellanos, M. E., and Ventura, L. (2012), "A matching prior for the shape parameter of the skew-normal distribution". *Scandinavian Journal of Statistics*, 39, 236–247.

Fechner, G.T. (1897), *Kollektivemasslehre*, Engelmann, Leipzig.

Fernandez, C. and Steel, M.F.J. (1998), "On Bayesian Modelling of Fat Tails and Skewness", *Journal of the American Statistical Association*, 93, 359-371.

Independent Evaluation Office (2015), "Evaluating Forecast Performance", Bank of England. https://www.bankofengla
/media/boe/files/independent-evaluation-office/2015/evaluating-forecast-performance-november-2015

Jones, M.C. and Pewsey, A. (2009), "Sinh-arcsinh distributions", *Biometrika*, 96, 761-780.

Mudholkar, G. S. and Hutson, A. D. (2000), "The epsilon-skew-normal distribution for analyzing near-normal data". *Journal of Statistical Planning and Inference*, 83, 291-309.

Ramirez-Cobo, P., Lillo, R.E., Wilson, S. and Wiper, M.P. (2010), "Bayesian inference for double Pareto lognormal queues". *The Annals of Applied Statistics*, 4(3), 1533-1557.

Rubio, F.J. and Steel, M.F.J. (2014), "Inference in Two-Piece Location-Scale models with Jeffreys Priors, with discussion". *Bayesian Analysis*, 9, 1-22.

Rubio, F.J. and Steel, M.F.J. (2015), "Bayesian modelling of skewness and kurtosis with two-piece scale and shape distributions". *Electronic Journal of Statistics*, 9, 1884-1912.

Wallis, K.F. (2003), "Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts". *International Journal of Forecasting*, 19, 165-175.

Zhu, D. and J. Galbraith (2010), "A Generalized Asymmetric t-distribution with Application to Financial Econometrics", *Journal of Econometrics*, 157, 297-305.