Evaluating Density Forecasting Models

Michael Carney, Pádraig Cunningham

Trinity College Dublin, Dublin 2, Ireland,
firstname.surname@cs.tcd.ie

Abstract. Density forecasting in regression is gaining popularity as real world applications demand an estimate of the level of uncertainty in predictions. In this paper we describe the two goals of density forecasting[1] *sharpness* and *calibration*. We review the evaluation methods available to a density forecaster to assess each of these goals and we introduce a new evaluation method that allows modelers to compare and evaluate their models across both of these goals simultaneously and identify the optimal model.

1 Introduction

Daily, we use and accept probability estimates for common prediction tasks; the weatherman tells you there is a 70% chance of rain, or medical experts say a patient has a 40% chance of being alive five years after a cancer operation. But, if it doesn't rain, or if the patient doesn't die, were those predictions wrong?

Evaluation of predictions is an important step in any forecasting process. For point estimates this is a straightforward process that typically involves determining the Euclidean distance between the predicted and observed points. There is a vast literature on evaluation metrics for point forecasting models, for a review of the most popular methods see [1]. However, there are conspicuously less papers available that describe methods for evaluating density forecasting models. In fact, one must turn to the meteorological and financial literature to find any papers that focus on the evaluation of density forecasts with any degree of rigour. This is in spite of density forecast evaluation being a considerably more complex problem than point estimation. Diebold et al. [2] suggest that there might be three reasons for this neglect.

1. *Restrictive assumptions* - until recently, due to the computational complexity of making density forecasts, very restrictive assumptions were required in terms of the number of parameters that could be estimated and the distributions that had to be assumed.
2. *Lack of demand* - in the past there was seemingly less demand for density forecasts, this is particularly true in the financial domain on which Diebold et al. focus. However, the recent growth in the area of risk management has focused the attention of people on this problem.

[1] For convenience and simplicity in this document the term "density forecasting" will refer to "probability density forecasting for regression" unless stated otherwise.

3. *Difficulty of the problem* - it is possible to adapt methods that are used in the point forecasting and interval forecasting literature to evaluate density forecasts, however, these adaptations lead to incomplete evaluations.

The defining difference between density forecast and point forecast evaluation is the fact that the performance of a density forecasting model can not be summarised meaningfully by one metric. This can be attributed to the richer information produced by a density forecasting model. The popular Mean Squared Error and Root Mean Squared Error scores are sufficient evaluation metrics for most when assessing the quality of point forecasting models. On the other hand, density forecasting models must both produce estimates that give a high density at the observation and produce probability estimates that are correct. The first requirement, a high density at the observation, relates to the predicted density having minimum variance about the observation, this is commonly termed the sharpness of an estimate. The second requirement, to produce probability estimates that are correct, refers to the empirical validity of the predicted probabilities and is commonly called *calibration*.

As mentioned above we aim to address the regression problem of estimating the parameters for a model given a set of training data $\{(x_i, t_i)\}_{i=1}^{m}$, where the i^{th} example is described by the pattern $x_i \in \Re^p$ and the associated response $t_i \in \Re$. Point forecasting attempts to estimate, $\langle t_i | x_i \rangle$, the conditional mean of the target variable given an input pattern[2]. Density forecasting models attempt to estimate, $p(t_i|x_i)$, the conditional probability density that the target is drawn from, a considerably more complex task.

The aim of this paper is to provide a review of evaluation techniques for density forecasting in regression and present a new way of combining the two main evaluation approaches used in the literature. The paper is organised as follows. In Section 2 we introduce the various terminology and high level concepts behind the two main approaches to evaluate density forecasting models (sharpness and calibration). Section 3 outlines methods for assessing sharpness and Section 4 reviews approaches of assessing calibration. In Section 5 we introduce our new method of combining and comparing these two evaluation approaches in a meaningful manner. Finally, in Section 6 we briefly conclude the paper.

2 Calibration, Sharpness, Refinement, Empirical Validity

The literature in point forecasting makes the suggestion that the "closer" a forecast is to the observation the better. Similarly, this intuition transfers to probability forecasting, for example, a forecast of 90% for an event that occurs will appear better, after the fact, than a forecast of 80% for the same event [3]. This property of density forecasts is known as sharpness or refinement [4]. Put simply, sharpness assesses how spread out or how "sharp" a forecaster's predictions are. In the binary sense this refers to the concentration of the probability estimates near the values 0 and 1. A sharp forecaster will have a high concentration of its probability estimates around these two extreme values. In the continuous domain, sharpness relates to the amount of density assigned

[2] At points in this document we refer to sequenced or time series data, in these cases the reader can assume that $x_i = t_{i-1}$

to the actual observation. High density at the observation necessitates a low variance around the observation.

Probability forecasts are unique insofar as they not only provide a prediction of the location/class of the observation but they also give a measure of the uncertainty in that prediction. Sharpness rewards models in terms of the location/class accuracy but gives no real indication of the correctness of probability estimates. Calibration, also known as reliability in meteorology or empirical validity in statistics [5], refers to the ability of a model to make good probabilistic predictions. A model is said to be well calibrated if for those events the model assigns a probability of P%, the long-run proportion that actually occur turns out to be P%. Intuitively, this is a desirable characteristic of any probabilistic forecast; in fact, it could be argued that probability forecasts that are not well calibrated are of no more use than point forecasts because the probabilistic aspect of the prediction is incorrect.

The two objectives outlined above are the requirements for any good probabilistic forecaster. Both sharpness and calibration evaluations are necessary to ascertain the quality of a probabilistic forecaster. Calibration evaluation must be accompanied by an estimate of sharpness in order to ascertain the usefulness and predictiveness of the forecaster.

3 Assessing Sharpness

In this section we identify three methods of evaluation of the sharpness criterion.

3.1 Negative Log-likelihood

The Negative Log-likelihood (NLL), also known as the Ignorance Score (Good, 1952) or Negative Log Predictive Density, is a method of assessing the sharpness of a predicted density function. It is specified as follows:

$$NLL_i = -\log(p(t_i|x_i)) \tag{1}$$

The NLL can be easily and cheaply calculated and is by far the most popular error function in the density forecasting literature for this reason. The NLL or a variation on the NLL is almost exclusively used as the optimisation error function of density forecasting models[3]. This is because of the relationship between NLL and Maximum Likelihood Estimation (MLE). MLE theory can be easily adapted to optimise density functions rather than point values. Figure 1 plots the relationship between the NLL value for a sample prediction density against all possible outcomes in the interval [0, 10]. The NLL is a negatively oriented score, meaning the more density the actual observation has been awarded by the prediction the smaller the NLL value.

The major weakness of the NLL is that it evaluates density estimates based solely on the probability density at the observation and does not take the calibration of the forecast into consideration. These problems manifest themselves in erroneous probability estimates. For examples of these see [6] or [7].

[3] The exceptions to this are, for example, the indirect density forecasting techniques such as ensemble methods.

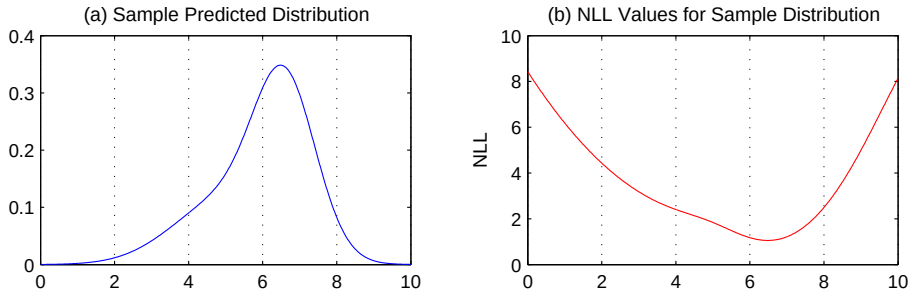

Fig. 1. The left plot depicts a sample predicted density. The right plot shows the NLL for a target at each point on the x axis for the given sample density. The relationship between density and contribution to error can be clearly seen.

A further weakness of NLL is its sensitivity to outliers. This is due to the fact that a change of x in the NLL relates to a change of $\exp(x)$ in the observation values. This effect can be seen in Figure 1. Weigend & Shi suggest using a trimmed mean to get around this issue [8] . This ameliorates the situation but does not solve the problem.

3.2 Continuous Rank Probability Score

The Continuous Ranked Probability Score (CRPS) [9] is a verification method for probabilistic forecasts of continuous variables. It is equivalent to the Brier Score [10] integrated over all possible values and is a generalisation of the Ranked Probability Score [11] that is used to evaluate probabilistic predictions over ordinal variables. The CRPS is sensitive to distance i.e. it is capable of penalising predictions that are far away from the actual observation. In essence, the CRPS measures the difference between the predicted and the occurred cumulative distributions, see Figure 2. In order for the score to be sensitive to distance, the squared errors are computed with respect to the cumulative probabilities of the forecast and observation. The CRPS is calculated as follows:

$$CRPS_i = \int_{-\infty}^{\infty} (p(u|x_i) - H(u, t_i))du \tag{2}$$

Where, H, is the Heaviside function,

$$H(l, m) = 1\{l \geq m\} \tag{3}$$

Again, like the NLL, the mean CRPS is calculated over all predictions to determine the average error.

Hersbach [9] shows that the CRPS reduces to the Mean Absolute Error for deterministic forecasts. Therefore, this evaluation technique is a means of comparing deterministic and probabilistic forecasting models and is also easy to interpret as an error measure. This interpretability is further aided by the fact that it is in the same units as the target variable.

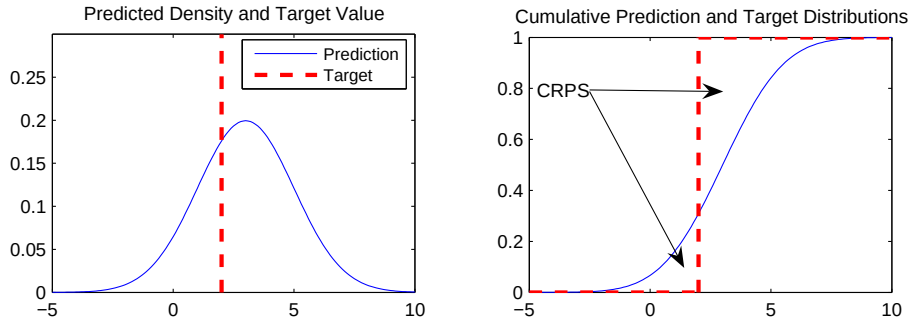

Fig. 2. The left plot in this figure is the predicted probability density function for some regression task. The target value is at 2. The right plot shows the cumulative distributions for the target and prediction. The area between the target and predicted cumulative (as shown by the arrows) is the value returned by the CRPS score. The sensitivity of the CRPS to distance is linear relative to the predicted densities error. Sharpness (small spread) is rewarded if the forecast is accurate. A perfect CRPS score is 0.

A major disadvantage of the CRPS score is that it is not of closed form. It is possible to derive a closed form version of the CRPS for normal distributions [12]; however, if your probabilistic model produces non-Gaussian solutions, such as mixture models, then determining numerical estimations of the integrals for every input pattern is a computationally intensive task.

3.3 Wilson Score

The Wilson Score [13] is a means of assessing the quality of forecasts of continuous variables in terms of an acceptable range. For example, predictions of temperature may only need to be accurate between $\pm 1°$C of the actual observation. This score evaluates predictions in terms of this range. Like the CRPS, the concept of the Wilson Score is derived from the Brier Score and Rank Probability Score. The Wilson Score determines the percentage of the forecasted probability that lies within the tolerable range of the observation/target. The equation for the Wilson Score is;

$$WS_i = \int_{t-\Delta t}^{t+\Delta t} p(u|x_i)du \tag{4}$$

where, Δt represents the threshold, or tolerated distance from t.

Figure 3 depicts the area contributing to the Wilson score for a sample prediction. This is a positively oriented score in the interval [0,1]. A perfect score receives a 1. The numerical score that it produces can be understood as a probability. It is sensitive to distance and to the spread of the forecasts.

A weakness of this score is that the modality of the predicted distribution affects the Wilson Score. The error score must be adapted to account for the different predicted modes. Wilson et al. describes how the score can be adapted by determining the modality of the target a priori [13]. However, this is not a trivial task and may not even be

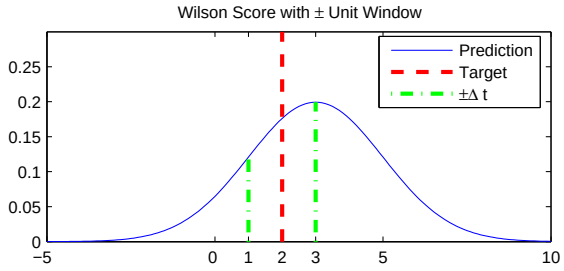

Fig. 3. Shows the area of the distribution that is considered by the Wilson Score $\pm\Delta t$. In this example a tolerance, or window size of ± 1 units is used.

achievable in certain scenarios. Another difficulty with this score, like the CRPS, is that it requires the calculation of an integral.

4 Evaluating Calibration

The Probability Integral Transform (PIT) as recommended by Dawid [14] is the most common method of evaluation of calibration. Within the PIT method of evaluation there exists a large number of approaches that can be used to interpret its results. Below we describe the PIT score and review the different methods of interpretation. Although other methods of assessing calibration exist they are generally presented in terms of the particular decision problem and are not general methods that can be applied in every case. This is an alternative and valid approach to forecast evaluation but we do not have space to address it in this short text, for more information see [15].

4.1 Probability Integral Transform

The PIT score is a popular method of evaluation across all forms of density forecasting, for example, Schervish uses it for evaluation of classification problems [3], Christoffessen adapts it for prediction intervals [16], and there is an equivalent method of evaluation for ensembles of point forecasters [17]. PIT overcomes the problem of loss function inconsistencies across problem domains by not depending on a specific user loss function. Instead, the PIT assumes that the prediction user is attempting to estimate the true data generating process. As the true data generating process is at least as good as, if not better than, any other possible model i.e. it weakly dominates all other models. Intuition also suggests that it is reasonable to assume as the correct density is always preferred to an incorrect density. In the case of the true data generating process, we know that the set of cumulative densities at the realisations will be uniform[4]. By making the assumption that we are striving to find the true data generating process we

[4] In the case of time series or sequential data the cumulative densities should also be independently and identically distributed. An excellent introduction to the PIT for sequential and time series data is provided in [2].

can assess the set of predicted cumulative densities at the observations for uniformity. The PIT is defined by:

$$z_i = \int_{-\infty}^{t_i} p(u|x_i)du \tag{5}$$

For a series of length M the probability of those events occurring in their predicted densities should result in a random sample as would appear using the true generating densities. Rosenblatt shows that this random sample will be U(0,1) and i.i.d. for the true generating density and any correctly specified density forecasting model [18].

Diebold et al. [2] suggest plotting a histogram of the PIT values and comparing this to a perfect U(0,1) distribution as a method of discerning the degree of calibration of a model. Crnkovic et al. [19] suggest using the Kuiper statistic for measuring uniformity. However, this is not a robust test for uniformity and requires a very large number of data points before it is consistent in its estimates. Berkowitz [20] suggests a Likelihood Ratio test for evaluating PIT values and developed a rigorous framework for evaluation of different aspects of the density forecast, even when evaluation data is sparse. Specifically, he proposes that the PIT values be transformed into a normally distributed series, N(0,1), via the inverse normal cumulative density function transform because tests for normality are more powerful than tests for uniformity. Using this transformed series he suggests that a one-degree-of-freedom test of independence against a first-order autoregressive structure and a three-degree-of-freedom test of zero mean, unit variance and independence. This approach is probably the most well developed and principled method of evaluation of PIT values to date; however, it is also the most time consuming and computationally intensive. Wallis [21] suggests using an adaptation of Pearson's chi-squared goodness-of-fit test for density forecasting evaluation. His suggested adaptation provides a means of extracting information from the PIT values that can diagnose more precisely where the predictions fall down e.g. location, scale or skewness. Somewhere between the suggestions of Crnkovic et al. [19], and Berkowitz and Wallis [20, 21], lies the research carried out by Noceti et al. [22]. In their paper, the Kolmogorov-Smirnov, Kuiper, Cramér-von Mises, Watson and the Anderson-Darling goodness-of-fit tests were compared. After analysis they concluded that the Anderson-Darling test was the most robust metric for this task.

At this point it is important to contextualise the problem again by referring back to our initial postulation that density forecasts should be both sharp and calibrated. Although, the PIT score identifies a well calibrated model, it is not sufficient to identify whether a density forecasting model is useful or not. In no way does the PIT score evaluate sharpness and so it should be used in conjunction with a sharpness score to identify models that are both well calibrated and sharp. In point of fact, determining the distribution of the observations in the training set and using this distribution as a prediction for every input will result in a uniform set of PIT values over the training set. This density is known as the unconditional distribution in finance or the climatology in meteorology. This trivial model will be well calibrated but will have very poor sharpness. In the context of time series data it is possible to argue that by determining whether the PIT values are independent, or not, one can identify if a model is simply predicting the unconditional distribution. However, Hamill [23] shows that in certain circumstances a biased model will return a uniform set of PIT values. He suggests that a uniform series

is a necessary but not sufficient criterion for determining that a model is calibrated. He shows that it is possible that an incorrect density model could have a uniform set of PIT values. Again, in this case, the sharpness score will highlight the fact that the model is incorrectly specified.

4.2 Interpreting PIT Histograms Visually

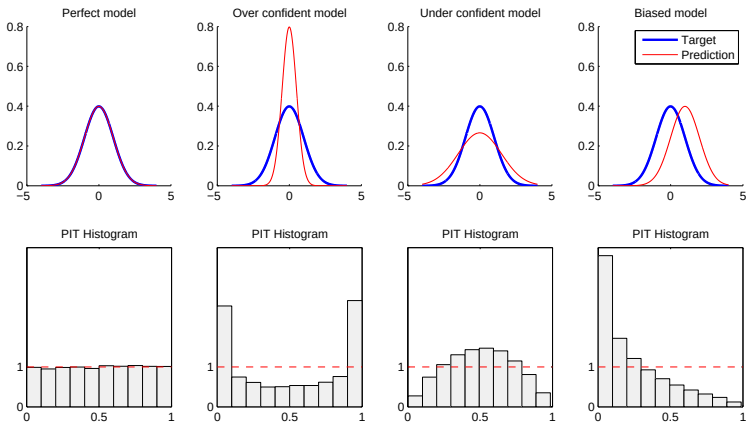

Fig. 4. The top row of plots shows the true or target generating distribution (N(0,1)) and the predicted distributions. From left to right the predicted distributions are, N(0,1), N(0, 0.5), N(0,2) and N(1,1). The bottom row of plots represents the resulting PIT histogram for each prediction distribution determined by evaluation on 10,000 points generated from the target distribution.

To further understand the PIT approach to determining calibration this section describes the effect of bias and variance[5] on the PIT histogram. To do this we simplify the problem; all data points in our test series are random samples from an N(0,1) distribution. The true conditional density at every point is, therefore, an N(0,1) density. Knowing the true density means we can artificially simulate bias and variance in the predicted densities by making all predicted densities N(μ, σ). Bias is simulated by varying the μ of the density and variance is simulated by varying the σ of the density. Figure 4 shows the distributions and resulting PIT histograms for each μ and σ pair. The first PIT histogram is the only correctly specified histogram because the predicted density is N(0,1), the same as the distribution used to generate the data. Bias and variance affect the PIT histogram in different ways. Too narrow a variance forms a “U” shaped PIT histogram signifying over-confident predictions. Too wide a variance creates a hump in the middle of the PIT, this can be thought of as under-confidence. Bias causes a sloping effect and in the extreme case it creates a “J” or “L” shaped PIT histogram depending on the direction of the bias.

[5] In this experiment, bias refers to the incorrect specification of the mean of the predicted density.

5 A Complete Evaluation Framework

In the preceding sections the argument for two metrics, sharpness and calibration, when evaluating density forecasting models was developed and a number of methods for assessing these were described. This is the generally accepted methodology for density forecasting evaluation by the literature and is seen as sufficient to fully evaluate the quality of a density forecasting model [12]. Given this fact, we propose a further diagnostic tool for evaluation and comparison of predictive performance of density forecasting models. We describe a simple method of comparison that can clearly and definitively identify the best models in terms of their sharpness and calibration objectives.

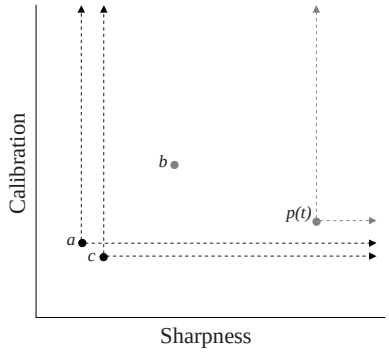

Fig. 5. Sample density forecasting evaluation plot. In this example the sharpness and calibration scores are negatively oriented i.e. the smaller the better. $p(t)$, represents the unconditional distribution of target values, a,b and c are sample models. Model b is dominated by both a and c. a and c do not dominate each other and so are optimal solutions. The region above and to the right of a model's point in objective space represents the region that that model dominates (dotted lines).

Vilfredo Pareto [24] was the first to discover that under multiple conflicting objectives there is no single optimum, instead there is a set of optimum trade-off solutions. These optimum solutions can be identified by their dominance or non-dominance amongst the other competing solutions. Therefore, Pareto dominance can be defined as the unique nontrivial partial order on the set of finite-dimensional real vectors satisfying a number of objectives [25]. This is precisely the ordering that is required in evaluation of the multiple objectives of density forecasting to find the optimal solutions. Formally, Pareto dominance can be described as follows: Assume, without loss of generality, k negatively oriented objectives and consider two sets of model parameters a, b. Then, a is said to dominate b iff:

$$\forall i \in \{1, ..., k\} : f_i(a) \leq f_i(b)$$
$$\exists j \in \{1, ..., k\} : f_j(a) < f_j(b)$$

Where, $f_i(x)$ returns for decision vector x the ith objective function [26].

The concept of Pareto optimal solutions is illustrated in Figure 5. The unconditional distribution is included to show the sharpness of models relative to this base model. Models that are dominated can be discarded. The forecaster can then select the model that best optimises their goals from the Pareto optimal set.

5.1 Sample Evaluation: Model Selection

A common application for evaluation techniques is model selection. Here, the goal is to determine the best model set up to use for your final prediction model. In this simple example we show how it is useful to plot a Pareto evaluation plot to determine the optimal models. For this experiment we used a Mixture Density Network (MDN) to make our density forecasts[27]. MDNs are an adaptation of the multi-layer perceptron that can accurately estimate conditional probability density functions by outputting a Gaussian Mixture Model (GMM). Like most neural networks there are a number of variables that must be decided upon by the modeler before the model can be trained. The two most important variables to be selected with this type of model are the number of hidden units in the network architecture and the number of Gaussians to be included in the GMM. In our experiment we use the Pareto optimality plot described above to determine the best model set up in terms of hidden units and Gaussian components for a simple inverse problem. Target variables, t, are uniformly drawn from the interval [0,1] and the input variables, x, are generated by $x = t + 0.3\sin(2\pi t) + \epsilon$ and ϵ is uniform noise drawn from the interval [-0.1, 0.1][6]. We created a training and test set of 1,000 points, from the training set we created 20 bootstrap training sets, this is so that we get a more robust estimate of the average model error score for each architecture. 18 different model architectures were tested, outputs of 1, 3 and 6 Gaussian components for the GMM were tested and for each output type we tested network architectures with 2,3,4,5,6 and 7 hidden units. All models were trained till termination or for 1,000 iterations of the Scaled Conjugate Gradient algorithm, whichever came first. The average error for each model over the 20 runs on the test set were calculated. We evaluate all models using the negative log-likelihood score (see Section 3.1) as our measure of sharpness and the Anderson-Darling goodness-of-fit test statistic on the model PIT values as our measure of calibration [28]. This calibration score is calculated by determining $A^2 = -m - \frac{1}{m}\sum_{j=1}^{m}(2j-1)[\log(z_i)+\log(1-z_{m-j})]$, where, m, is the number of z values, calculated by equation 5, and these z values are sorted in ascending order. This resulting quality score for calibration is negatively oriented. Figure 6 describes the results from this experiment. After analysis of the Pareto plot the modeler should have a good understanding of which model set up best suits their goals. However, if there is still uncertainty regarding the best model they can carry out further tests such as those described in [20].

Figure 6 can be augmented in a number of different ways, for example; the position of the unconditional distribution can be included to identify the position of the baseline model or in the model selection scenario at least, the objective function score may be noisy due to the specific data set being evaluated, by applying a technique such as

[6] As described in [27] pp 14.

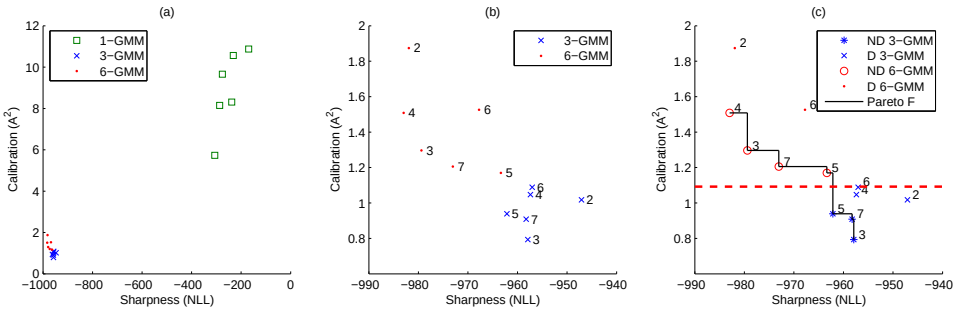

Fig. 6. (a) plots the results of all the models tested. The models with just 1 Gaussian component (1-GMM) clearly perform the worst and can be discarded. This is to be expected as the data is trimodal. (b) is a close up of the other models, the numbers next to the markers signify the number of hidden units used in that model. There are clearly two clusters, the 3-GMM models having good calibration, and the 6-GMM models having good sharpness. Finally (c) is the Pareto plot. The non-dominated (ND) Pareto optimal set is identified and the Pareto front is shown. Dominated (D) models can be discarded. The critical value (dashed line) for the A^2 score is included, below this line models are uniform at the 1% level.

that described in [29] it is possible to plot a Pareto optimal set that are mutually non-dominating with some known probability.

6 Conclusions

In this paper we described the goals of density forecasting as sharpness and calibration and identified approaches for evaluating models on both of these criteria. We introduced a new method of evaluation that allows the modeler to identify the best models from a set based on these two criteria.

References

1. Armstrong, J.S., Collopy, F.: Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting **8** (1992) 69–80
2. Diebold, F.X., Gunther, T.A., Tay, A.S.: Evaluating density forecasts with applications to financial risk management. International Economic Review **39**(4) (1998) 863–83
3. Schervish, M.J.: A general method for comparing probability assessors. The Annals of Statistics **17** (1989) 1856–1879
4. DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. The Statistician **32** (1983) 12–22
5. Seillier-Moiseiwitsch, F.and Dawid, A.P.: On testing the validity of sequential probability forecasts. Journal of the American Statistical Association **88** (1993) 355–359
6. Carney, M., Cunningham, P.: Calibrating probability density forecasts with multi-objective search. In: In Proceedings of European Conference in Artificial Intelligence. (2006)
7. Carney, M., Cunningham, P., Lucey, B.M.: Making density forecasting models statistically consistent. Technical Report TCD-CS-2006-04, Dept. of Comp. Sci., TCD (2006)

8. Weigend, A.S., Shi., S.: Predicting daily probability distributions of S&P500 returns. Journal of Forecasting **19**(4) (2000) 375–392
9. Hersbach, H.: Decomposition of the continuous ranked probability score for ensembles prediction systems. Weather Forecasting **15** (2002) 559–570
10. Brier, G.: Verification of forecasts epressed in terms of probability. Monthly Weather Review **78** (1950) 1–3
11. Epstein, E.: A scoring system for probabilities of ranked categories. Journal of Applied Meterology **8** (1969) 985–987
12. Gneiting, T., Raftery, A.: Strictly proper scoring rules, prediction, and estimation. Technical Report 463, University of Washington (2004)
13. Wilson, L.J., Burrows, W., Lanzinger, A.: A strategy for verification of weather element forecasts from an ensemble prediction system. Montly Weather Rev. **127** (1999) 956–70
14. Dawid, A.P.: Statistical theory: The prequential approach. Journal of the Royal Statistical Society A **147** (1984) 278–292
15. Granger, C.W.J., Pesaran, M.H.: Economic and statistical measures of forecast accuracy. Journal of Forecasting **19** (2000) 537–560
16. Christoffersen, P.F.: Evaluating interval forecasts. Intnl. Econ. Review **39** (1998) 841–862
17. Talagrand, O., Vautard, R., Strauss, B.: Evaluation of probabilistic prediction systems. In: In Proceedings, ECMWF Workshop on Predictability (20-22 October 1997) 1 - 25. (1997)
18. Rosenblatt: Remarks on a multivariate transformation. Annals Math. and Stat. **23** (1952) 470–472
19. Crnkovic, C., Drachman, J.: "quality control", in var: Understanding and applying value-at-risk. In: Risk Publications. London (1997)
20. Berkowitz, J.: Testing density forecasts, with applications to risk management. Journal of Business & Economic Statistics, American Statistical Association **19** (2001) 465–474
21. Wallis, K.F.: Chi-squared tests of interval and density forecasts, and the bank of englands fan charts. International Journal of Forecasting **19** (2003) 165–175
22. Noceti, P., Smith, J., Hodges, S.: An evaluation of tests of distributional forecasts. Journal of Forecasting **22**(6-7) (2003) 447–455
23. Hamill, T.M.: Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review **129** (2000) 550–561
24. Pareto, V.: Cours D'Economie Politique. Lausanne (1896)
25. Voorneveld, M.: Characterization of pareto dominance. Technical report, Department of Economics, Stockholm School of Economics (2002)
26. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. IEEE Trans on Evo. Comp. **3**(4) (1999) 257–271
27. Bishop, C.M.: Mixture density networks. Technical report, NCRG, Aston University (1994)
28. Anderson, T., Darling, D.: A test of goodness of fit. J. of Amer. Stat. Ass. **19** (1954) 765–769
29. Fieldsend, J., Everson, R.: Multi-objective optimisation in the presence of uncertainty. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'05). (2005)