Calibration of Forecast Trajectories

by
Christopher Heinz

in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Mathematics

at Heidelberg University
Faculty of Mathematics and Computer Science
May 8, 2012

Thesis Supervisors:
Dr. Thordis L. Thorarinsdottir
Prof. Dr. T. Gneiting

Abstract

In this paper we introduce a new concept of calibration for functional data, the modified band depth rank. It is based on a concept of ordering functions center-outwards based on their graphs, the modified band depth. We then compare the modified band depth rank with the existing multivariate rank in a simulation study as both concepts can be applied to multivariate forecasts in $\mathbb{R}^d$. In a second simulation study we evaluate the robustness of the modified band depth rank. We then apply it to two real data examples, temperature forecasting and inflation forecasting.

In meiner Bachelorarbeit wird ein neues Konzept für die Kalibrierung von functionalen Daten eingeführt, der *Modified Band Depth Rang.* Er basiert auf einer Methode um Funktionen an Hand ihrer Graphen von innen nach aussen anzuordnen. Nach der Einführung des neuen Rang-Konzeptes wird es mit einem bereits bestehenden Konzept, dem multivariaten Rang, verglichen, denn beide Rang-Konzepte können auf multivariate Vorhersagen im $\mathbb{R}^d$ angewendet werden. In einer zweiten Simulationsstudie wird die Robustheit des Modified Band Depth Ranges untersucht. Die neue Methode wird dann auch für zwei reale Datenbeispiele benutzt, und zwar für Temperaturvorhersagen und für Inflationsvorhersagen.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Heidelberg, May 8, 2012 ____________________

Christopher Heinz

Contents

Chapter 1

Introduction

How should I value a payment that I will receive in one year? Is it wise to have a picnic tomorrow afternoon? Accurate forecasts can yield big economic and social benefits in these and numerous other situations. Because of this it only seems natural that forecasting plays such an important role in modern statistics.

Forecasts can have different types of data structures ranging from classical point predictions to newer distributional predictions (Gneiting 2008) or ensemble forecasts, which will be of main interest in this paper. Ensemble forecasts are, for example, in use for variables that are especially hard to predict such as inflation and in medium-range weather forecasting. In these cases it makes sense to not only look at one forecast but rather try to draw information from a number of forecasts which may differ in their numerical methods or they may be forecasts by different forecasters. In the two examples mentioned above there is also often an interest in having a prediction not only for one point in time but a prediction trajectory for several successive points in time so that one can see how the variable of interest may behave or evolve over time. This type of data can be seen as an important special case of a more complex data type, the functional data. In functional data analysis each observation is a real function on an Interval I or a discrete set $\{1, \ldots, d\}$ (López-Pintado and Romo 2009).

This paper focuses on a new concept of calibration for forecasts of functional data. Calibration, in a broader sense, is a concept of evaluating the quality of past predictions, which also takes an important role in statistics. If a forecast is calibrated then it has in a certain way the same properties as the observed value, which is known for past predictions. A point forecast is in particular calibrated, if the forecast and the observed value derive from the same distribution. A forecast distribution F for the observation y is called calibrated, if $F(y)$ is drawn from the uniform distribution $U([0,1])$.

There are several situations in applied statistics in which this concept has to be generalized, for example multivariate (mv) rank for continuous vector-valued forecasts in $\mathbb{R}^d$ (Gneiting et al. 2008). This paper introduces a generalisation of calibration for functional forecasts, especially the finite-dimensional case, for which the concept of the multivariate rank (mv-rank) also applies. For this, we first introduce an ordering for functional data. Here, we choose the concept of modified band depth, introduced by López-Pintado and Romo (2009), which orders real functions by their graphs center-outwards.

The remainder of the paper is organized as follows. In Chapter 2 we will review the concept of calibration and review the generalisation to vector-valued forecasts based on the mv-rank (Gneiting et al. 2008). Then we review the concept of band depth for functional data (López-Pintado and Romo 2009) and use it to introduce a concept of calibration for functional data based on the band depth rank and also compare it to the mv-rank in simulation studies. We also shortly review another application of the band depth, functional boxplots (Sun and Genton 2011). In Chapter 3 we perform a simulation study to evaluate robustness of the proposed concept. Chapter 4 contains two real data examples. In Chapter 5 we summarize the results.

Chapter 2

Theory

2.1 Calibration

The probability integral transform (Dawid 1984; Gneiting et al. 2007) uses the fact that, if X is a random variable with continuous cumulative distribution function (CDF) F, then $Y = F(X)$ is uniformly distributed, i.e. $Y \sim U([0,1])$. Based on this result a forecast distribution F for the observation y is called calibrated, if $F(y)$ is drawn from the uniform distribution $U([0,1])$. If the forecast distribution F is the empirical distribution of m point forecasts $x_1, \ldots, x_m$ then the forecast is called calibrated, if the rank of the observed value y in $\{x_1, \ldots, x_m, y\}$ is uniformly distributed, i.e. derives from the discrete uniform distribution of the on $\{1, \ldots, m+1\}$. This is obviously the case, if y as well as the forecasts $x_1, \ldots, x_m$ are independent realisations of the same random variable X with continuous CDF F. A common way to check whether the forecasts $x_1, \ldots, x_m$ are calibrated is to use a rank histogram (Anderson 1996; Hamill and Colucci 1997). For this, the rank $\mathrm{rank}(y_i)$ of the observed value y_i in $\{x_{i1}, \ldots, x_{im}, y_i\}$ is computed for $i = 1, \ldots, n$, where n is sufficiently large. Then we compute the histogram of the points $\{\mathrm{rank}(y_i), \quad i = 1, \ldots, n\}$ and check for deviations from the uniform histogram. An alternative way of checking for deviation from uniformity is the discrepancy or reliability index

$$\mathit{\Delta} = \sum_{j=1}^{m+1} |f_j - \frac{1}{m+1}|, \tag{2.1}$$

where f_j is the relative frequency of rank j (Delle Monache et al. 2006). If the ranks of the observed values y_i are uniformly distributed then every rank occurs with a probability of $\frac{1}{m+1}$ and the reliability index should thus be close to zero.

2.2 The multivariate rank (mv-rank)

If we have two vectors in $\mathbb{R}^d$, $\mathbf{x} = (x_1, \ldots, x_d)'$ and $\mathbf{y} = (y_1, \ldots, y_d)'$, then we write $\mathbf{x} \preceq \mathbf{y}$ if and only if $x_k \leq y_k$ for $k = 1, \ldots, d$.

We now consider an ensemble forecast of size m $\{\mathbf{x}_j \in \mathbb{R}^d, \quad j = 1, \ldots, m\}$ and the observation $\mathbf{x}_0 \in \mathbb{R}^d$. With the definition above we can assign *pre-ranks*

$$\rho_j = \sum_{k=0}^{m} \mathbf{1}\{\mathbf{x}_k \preceq \mathbf{x}_j\}, \quad j = 1, \ldots, m.$$

Since every pre-rank is an integer between 1 and $m+1$ the pre-ranks can be used to assign ranks, the so called multivariate rank, to the vectors $\mathbf{x}_0, \ldots, \mathbf{x}_m$. The multivariate rank r of the observation vector $\mathbf{x}_0$ is the rank of ρ_0 in $\{\rho_0, \ldots, \rho_m\}$ with ties resolved at random. It is straightforward to see that the multivariate rank is uniform if the ensemble members and the verifying observation are exchangeable (Gneiting et al. 2008).

2.3 Modified band depth for functional data

López-Pintado and Romo (2009) introduce a band depth and a modified band depth (mbd), which allow centre-outward orderings of curves that are computationally feasible. We will focus on the modified version since it is more suited to our real data examples.[1] The introduced ordering is based on graphic representations of the curves. Mbd is defined both for continuous functions on an interval and for the finite dimensional version, which is the one we use in the remainder of the paper.

From now on let $S = \{x_1, \ldots, x_m\}$ be a set of discrete functions on $\{1, \ldots, d\}$,[2] i.e.

[1] Mbd is more suited for irregular curves and so should be better suited for e.g. inflation data

[2] In our examples the d values of one function would be d forecasts for d different time horizons by one forecaster or forecasting-method.

interpretable as points in $\mathbb{R}^d$. Let $x = (x(1), \dots, x(d))$ be the expression in $\mathbb{R}^d$ for one of these functions $x \in S$. The modified band depth, $\text{mbd}_m(x)$, of $x \in S$ is defined as the proportion of coordinates of x inside the coordinates of any pair of functions in S:

$$\text{mbd}_m(x) = \binom{m}{2}^{-1} \sum_{1 \le i_1 \le i_2 < m} \frac{1}{d} \sum_{k=1}^{d} \mathbf{1}\{\min\{x_{i_1}(k), x_{i_2}(k)\} \le x(k) \le \max\{x_{i_1}(k), x_{i_2}(k)\}\}.$$

Obviously mbd satisfies $0 \le \text{mbd}_m(x) \le 1$ and gets the closer to 1 the deeper the curve x is in S.

López-Pintado and Romo (2009) also point out that the modified band depth can be brought into relation with another depth concept, the simplicial depth (sd).[3] The simplicial depth (Liu 1990) is a measure of depth for multivariate data points. The univariate simplicial depth of $x(k)$ can be written as

$$\text{sd}_{m,k}(x(k)) = \binom{m}{2}^{-1} \sum_{1 \le i_1 \le i_2 < m} \mathbf{1}\{\min\{x_{i_1}(k), x_{i_2}(k)\} \le x(k) \le \max\{x_{i_1}(k), x_{i_2}(k)\}\}.$$

The sums in the definition of $\text{mbd}_m(x)$ are finite and thus we can rearrange the terms, such that mbd_m can be interpreted as the univariate simplicial depth averaged over the d coordinates, i.e.

$$\text{mbd}_m(x) = \frac{1}{d} \sum_{k=1}^{d} \text{sd}_{m,k}(x(k)). \tag{2.2}$$

2.4 The modified band depth rank (mbd-rank)

From the concept of band depth it is straightforward to generalise the definition of calibration to the finite dimensional version of functional data. If we have m forecast curves $x_1, \dots, x_m$ and an observation curve y, each with length d, then the *modified band depth rank* (mbd-rank) is the rank of $\text{mbd}_m(\mathbf{y})$ in $\{\text{mbd}_m(\mathbf{x}_1), \dots, \text{mbd}_m(\mathbf{x}_m), \text{mbd}_m(\mathbf{y})\}$ with ties resolved at random. We call an ensemble of forecasting curves $x_1, \dots, x_m$ calibrated, if $\text{mbd}_m(\mathbf{y})$ derives from the discrete uniform distribution on the set $\{1, \dots, m+1\}$.

For $d = 1$ it is straightforward to see that the mbd-rank of y is uniformly distributed, if y and $x_1, \dots, x_m$ are independent realisations of the same random variable X with continuous

[3] But contrary to the modified band depth, the simplicial depth is computationally intensive.

CDF F and the forecasts are thus calibrated. The calibration property further holds for every $d \in \mathbb{N}$, if for every $k \in \{1, \ldots, d\}$ the forecasts for the time horizon k, $x_1(k), \ldots, x_m(k)$, and $y(k)$ derive from the same random variable X.

We have only looked at the finite dimensional version so far, since it is more relevant in practice, but we should keep in mind that the modified band depth rank can be computed the same way for a more complex type of data, i.e. real functions on an interval, if we use the corresponding definition of band depth for this kind of data, see Section 5 in López-Pintado and Romo (2009).

2.5 Functional boxplots

Sun and Genton (2011) use the mbd and the ordering it induces on graphs from the centre outwards to create functional boxplots. If we have m functions $x_1, \ldots, x_m$ of length d then let $x_{[1]}, \ldots, x_{[m]}$ be the order statistics induced by the modified band depth with ties resolved at random as in the previous chapter. Then $x_{[m]}$ is the most outlying curve and $x_{[1]}$ is the deepest or most central curve and is also named the median curve. We can also look at the α proportion ($0 < \alpha < 1$) of the deepest curves. For the finite dimensional version the region with the 50% deepest curves is the region obtained by

$$C_{0.5} = \{(k, x(k)) : \min_{r=1,\ldots,\lceil n/2 \rceil} x_{[r]}(k) \leq x(k) \leq \max_{r=1,\ldots,\lceil n/2 \rceil} x_{[r]}(k), k = \{1, \ldots, d\}\},$$

where $\lceil n/2 \rceil$ is the smallest integer not less than $n/2$. If the region obtained by $C_{0.5}$ is used in analogy to the middle 50 % of data that is represented by the box in the classical boxplot, then one can use the median curve and the region of the 50% deepest curves to create a functional boxplot in analogy to the classical one. As in the classical boxplot, one can define outliers by getting a maximum non-outlying envelope by inflating the 50 % deepest curves area by 1.5 Every curve that crosses this maximum non-outlying envelope is marked as outlier.

2.6 Comparison of mbd-rank and mv-rank

Gneiting et al. (2008) introduce the mv-rank for forecasts in $\mathbb{R}^d$ for applications in which the dimension d is small. However the concept of mv-rank also applies for high dimensions. In comparison, the mbd-rank was introduced for continuous functions as well as the finite dimensional version in which the curves can also be interpreted as points in $\mathbb{R}^d$. In this case it is viable to compute both the mv-rank and the mbd-rank. The difference between these two concepts is the way the pre-ranks are assigned. For the mv-rank the number of points with every entry being smaller or equal to the respective point is used for ordering. For the mbd-rank the ordering is induced by the deepness of the graphs with the most central curve having the highest mbd-rank. If we also take the graphical approach for computing the mv-prerank then the mv-preranks are based on the ordering of the curves from the bottom to the top with the curve at the top having the highest prerank. The main difference between the two concepts is that for the mv-prerank a point or a curve $\mathbf{x}_1 \in \mathbb{R}^d$ only counts as "lower" than another curve $\mathbf{x}_2 \in \mathbb{R}^d$ if $\mathbf{x}_1$ is "lower" (i.e. smaller) in every component than $\mathbf{x}_2$. In contrast in the concept of mbd-rank the proportion of a curve being in a band of two other curves is considered. Because of this we expect that the assigned mv-preranks are more often the same than the mbd-preranks as the dimension d increases. Thus we can also expect the mv-rank concept having problems indicating a lack of calibration for higher values of d. We now compare these two rank concepts regarding to the shape of their rank histograms and their ability to detect non-calibration for $d \in \{2, 5, 20\}$.

For $d = 2$ we expect the mv-rank concept to give good results, since it was introduced and tested for small d (Gneiting et al. 2008). Thus we also want to compare the shapes of the histograms based on the two different concepts. The rank histogram of non-calibrated forecasts can have different shapes. For example the non-calibration of the forecasts can be caused by overdispersion or bias. We now examine how these sources of non-calibration affect the look of the mv-rank histogram in comparison to the mbd-rank histogram. We illustrate this by simulating 10,000 times $m = 9$ forecasting curves with $d = 2$ and compute the rank of a simulated observation curve which might follow a different distribution than the forecasting curves. Figure 2.1 shows the results for computing the mv-rank and Figure 2.2 for the mbd-rank.

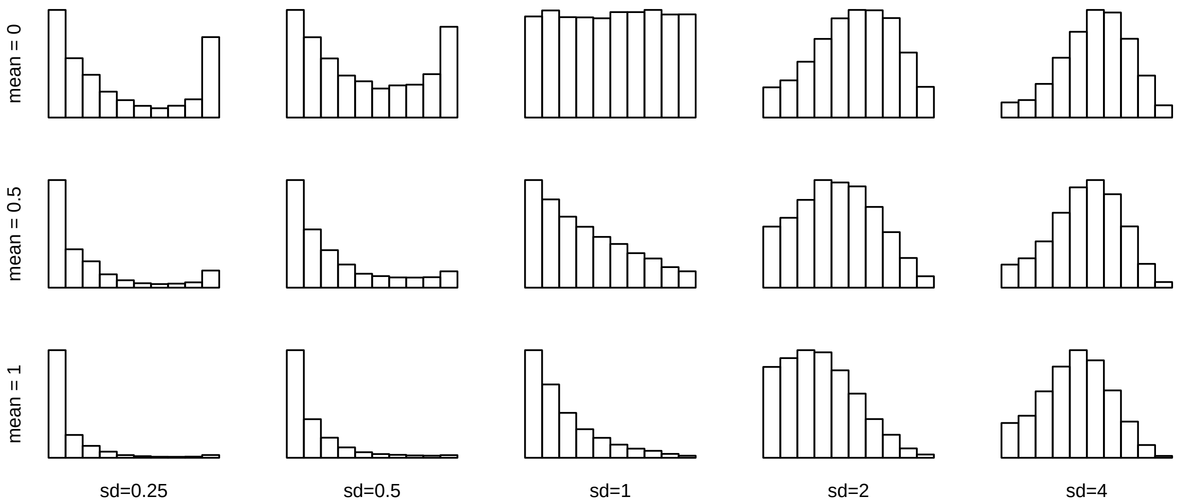

Figure 2.1: Simulation study for the mv-rank histogram with $d = 2$ dimensions. The points of the observation curve are iid standard normal. The points of the 9 forecasting curves are iid normally distributed with mean and standard deviation (sd) as specified. Repetitions have been simulated 10,000 times.

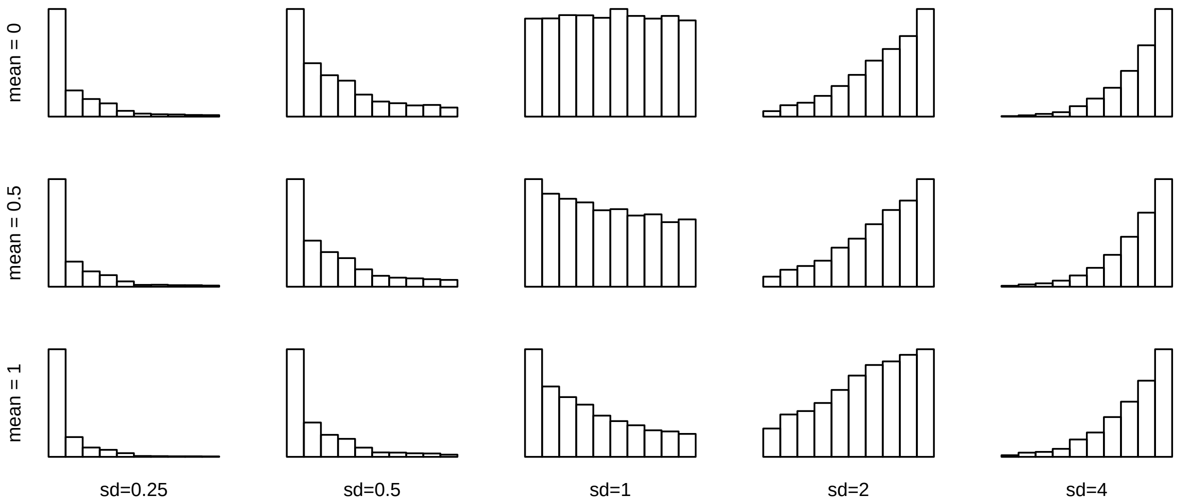

Figure 2.2: Simulation study for the mbd-rank histogram with $d = 2$ dimensions. The points of the observation curve are iid standard normal. The points of the 9 forecasting curves are iid normally distributed with mean and standard deviation (sd) as specified. Repetitions have been simulated 10,000 times.

We see that both concepts are able to reject calibration correctly for $d = 2$ dimensions and 10,000 repetitions. From Figure 2.1 we see that, for the mv-rank, when central tendencies are biased, the histogram is skewed. Overdispersion leads to an inverse U-shaped histogram and underdispersion to a U-shaped histogram. In comparison, the corresponding mbd-rank histograms do not show any sign of a U-shape. Instead, all deviations from the standard normal distribution result in skewed histograms. We see that we have to be careful when interpreting rank histograms based on different concepts and that less information can be drawn from the shape of the mbd-histogram in comparison to a mv-rank histogram when d is small.
We repeat this simulation for $d = 5$.

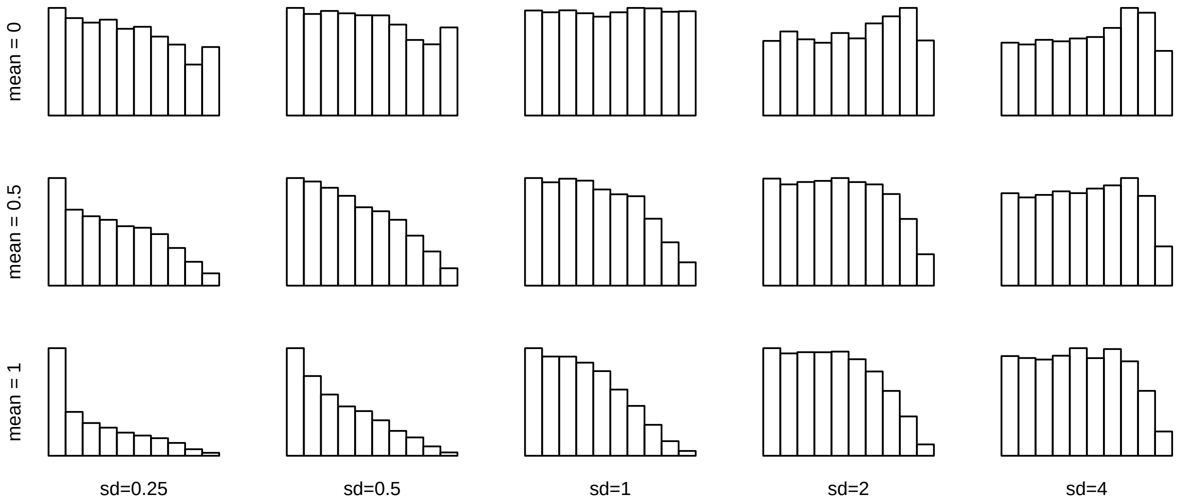

Figure 2.3: Simulation study for the mv-rank histogram for $d = 5$ dimensions. The points of the observation curve are iid standard normal. The points of the 9 forecasting curves are iid normally distributed with mean and standard deviation (sd) as specified. Repetitions have been simulated 10,000 times.

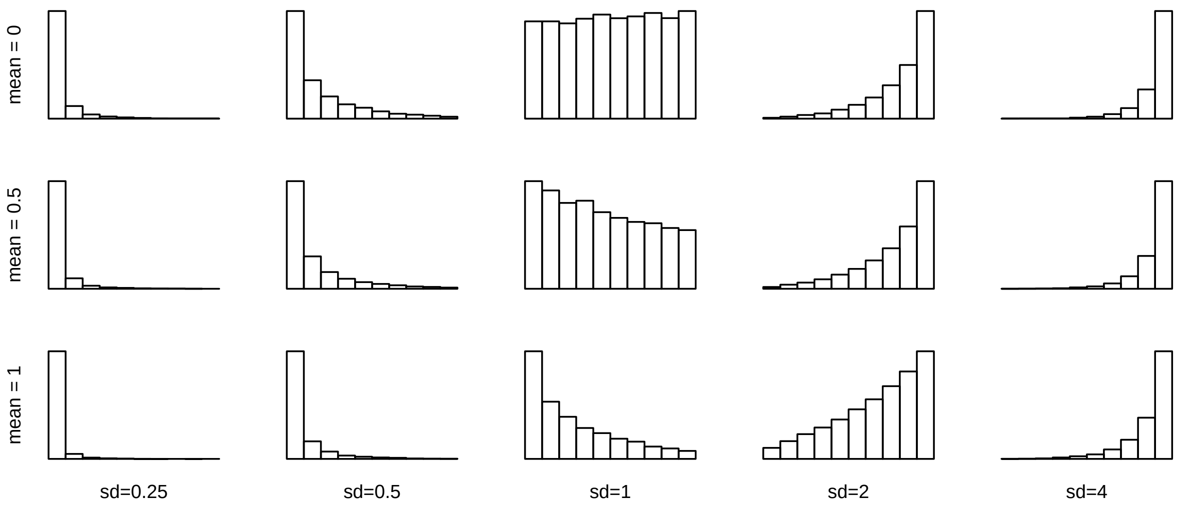

Figure 2.4: Simulation study for the mbd-rank histogram for $d = 5$ dimensions. The points of the observation curve are iid standard normal. The points of the 9 forecasting curves are iid normally distributed with mean and standard deviation (sd) as specified. Repetitions have been simulated 10,000 times.

We see in Figure 2.4 that for the mbd-rank with $d = 5$ detection of non-calibrated forecasts did not deteriorate but rather improved slightly overall, especially in the last two columns, in comparison to Figure 2.2 with $d = 2$. On the other hand in the mv-rank setting with $d = 5$ of Figure 2.3 all histograms look considerably more uniform with some even having problems to clearly detect non-calibration at all in comparison to the smaller total number of prediction horizons in Figure 2.1. Also, the types of U-shapes have disappeared in Figure 2.3 and thus for a medium number of total prediction horizons, here $d = 5$, mv-rank not only performs visibly worse than mbd-rank in detecting non-calibration, but also the advantage of giving more information about the type of non-calibration is not existent any more for the mv-rank concept.

We now repeat this simulation for $d = 20$ prediction horizons.

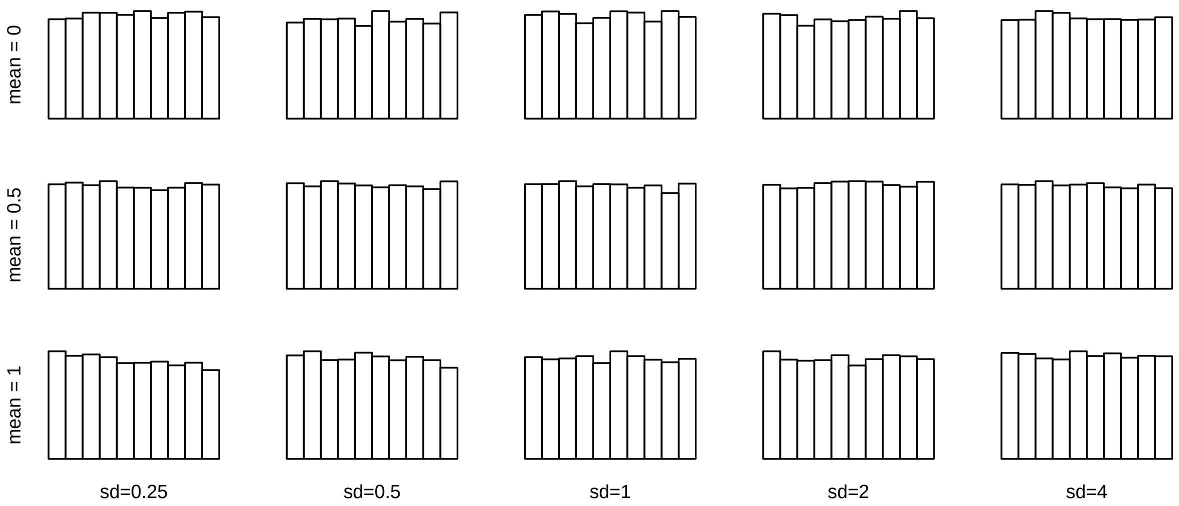

Figure 2.5: Simulation study for the mv-rank histogram for $d = 20$ dimensions. The points of the observation curve are iid standard normal. The points of the 9 forecasting curves are iid normally distributed with mean and standard deviation (sd) as specified. Repetitions have been simulated 10,000 times.

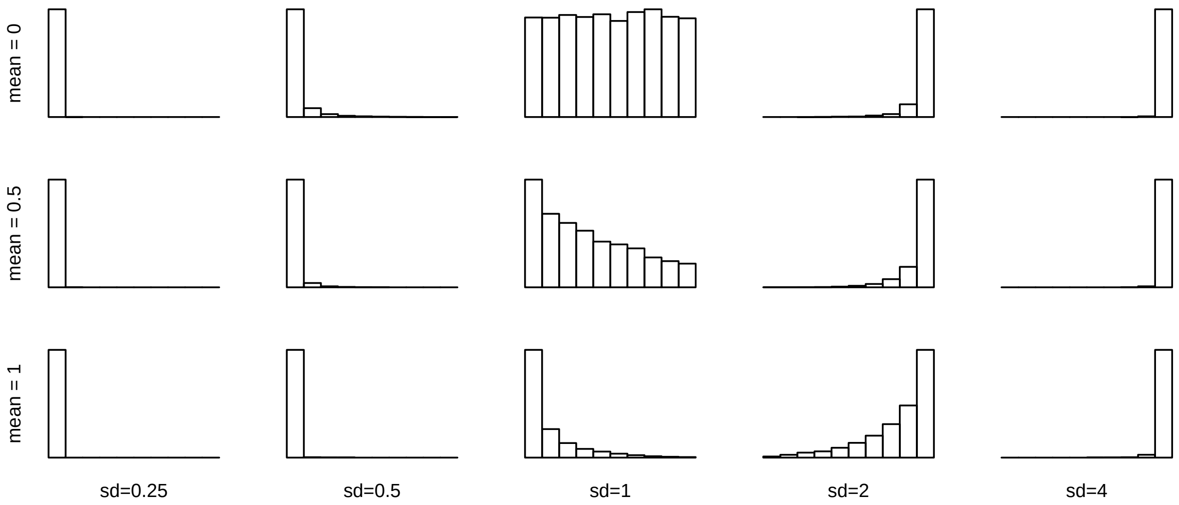

Figure 2.6: Simulation study for the mbd-rank histogram for $d = 20$ dimensions. The points of the observation curve are iid standard normal. The points of the 9 forecasting curves are iid normally distributed with mean and standard deviation (sd) as specified. Repetitions have been simulated 10,000 times.

For a higher total number of prediction horizons, here $d = 20$, we see in Figure 2.5 and 2.6 that the mv-rank is not able to detect non-calibration at all, whereas the mbd-rank performs very well and seems only to improve with increasing number of prediction horizons.

Chapter 3

Robustness of the mbd-rank

By simulating mbd-ranks for different number of repetitions, different number of curves, m, and different number of prediction horizons d, and whose ranks should theoretically be uniformly distributed, we check whether and how these parameters affect how easy it is to see that the mbd-ranks are indeed uniformly distributed. For this we computed the reliability index $\varDelta$ (1) which should be closer to zero the more apparently uniform the ranks of the curves are. The data was generated by drawing every point on each curve independently from the standard normal distribution independently. Because of that the mbd-ranks should derive from a uniform distribution. The results can be seen in Tables 3.1,3.2 and 3.3. The number of repetitions increases with every row. The number of curves increases with the columns. In the simulation process the "last" generated curve was simply treated as the observation curve. The length of the curves, or the number of prediction horizons, differ by the tables; we use $d \in \{5, 10, 50\}$. A typical characteristic of ensemble forecasts with different prediction horizons is that the variance of the underlying forecasts increases with increasing lead times. Because of that we also included this scenario into our simulations. The grey numbers in brackets are the reliability index $\varDelta$ with the underlying simulated curves having linearly increasing variances σ^2 in every point of the curves,

$$\sigma_k^2 = k, \qquad \forall k \in \{1, \ldots, d\}$$

Table 3.1: Reliability index $\mathit{\Delta}$ for simulated mbd-ranks with 5 prediction horizons with every point being standard normal (black number) and with linearly increasing variances (grey number) in every prediction horizon of each curve, i.e. $\sigma_k^2 = k, \quad \forall k \in \{1, \ldots, 5\}$.

	Ensemble size + 1		
Repetitions	10	20	100
50	0.28 (0.32)	0.46 (0.5)	1.22 (1.24)
100	0.32 (0.34)	0.36 (0.28)	0.74 (0.64)
500	0.11 (0.12)	0.18 (0.16)	0.33 (0.36)

Table 3.2: Reliability index $\mathit{\Delta}$ for simulated mbd-ranks with 10 prediction horizons with every point being standard normal (black number) and with linearly increasing variances (grey number) in every prediction horizon of each curve, i.e. $\sigma_k^2 = k, \quad \forall k \in \{1, \ldots, 10\}$.

	Ensemble size + 1		
Repetitions	10	20	100
50	0.36 (0.4)	0.38 (0.5)	1.26 (1.16)
100	0.18 (0.12)	0.36 (0.4)	0.64 (0.7)
500	0.08 (0.12)	0.15 (0.18)	0.35 (0.33)

Table 3.3: Reliability index $\mathit{\Delta}$ for simulated mbd-ranks with 50 prediction horizons with every point being standard normal (black number) and with linearly increasing variances (grey number) in every prediction horizon of each curve, i.e. $\sigma_k^2 = k, \quad \forall k \in \{1, \dots, 50\}$.

	Ensemble size + 1		
Repetitions	10	20	100
50	0.28 (0.28)	0.56 (0.44)	1.22 (1.24)
100	0.26 (0.32)	0.40 (0.22)	0.84 (0.66)
500	0.07 (0.9)	0.15 (0.13)	0.36 (0.31)

The results are as expected in that the reliability index $\mathit{\Delta}$ approaches zero for increasing number of repetitions as it should. Seemingly, only the number of repetitions and the ensemble size affect divergence from uniformity. This is just the same question as how many times do I have to draw independently from a discrete uniform distribution for the outcome to look uniformly distributed, which also depends on the cardinal number of the set from which is drawn, which in this case is the number of curves. If we compare the same entry in each of the 3 tables then we can conclude that the length of the curves does not seem to have an impact on $\mathit{\Delta}$.[1] The same holds for having the same variance at every point of the curves compared to having increasing variances respectively. If we look at the diagonal of each of the three tables we could also conclude that only the proportion of repetitions to curves matters and not the actual amounts of both.

[1] This seems to be counterintuitive in the light of relation (2.2).

Chapter 4

Real data examples

4.1 Temperature forecasting

The first real data example concerns weather forecasts, specifically the 50-member ECMWF ensemble forecast of the European Centre for Medium-Range Weather Forecasts (Molteni et al. 1996). The forecasts are issued twice a day on a 0.25 by 0.25 degree grid over the entire globe for a large number of weather variables. Each time, the current weather and the future weather in steps of 3 hours with a total lead time of 49 prediction horizons is issued. The single ensemble members differ only in random perturbations in the boundary conditions and they are thus exchangeable. For this analysis we use the temperature forecasts initialised at 0000 UTC for the period from February 1, 2010 until April 30, 2011, for a total of 444 days. For the observations, we have the data from 500 stations in Germany provided by the German Weather Service. Out of the 500 stations we choose 100 at random with full data. The weather stations for the observed data do not usually lie on the forecasting grid. Thus, we use bilinear interpolation for the forecasts to have forecasts at the locations of the weather stations. For the actual calculations of the ranks we only use the last year (365 days) of data in our data set.

This means for our setting that we have $m = 50$ ensemble members with $d = 49$ prediction horizons and $n = 36500$ observations. Figure 4.1 shows two examples of observation and forecasting curves.

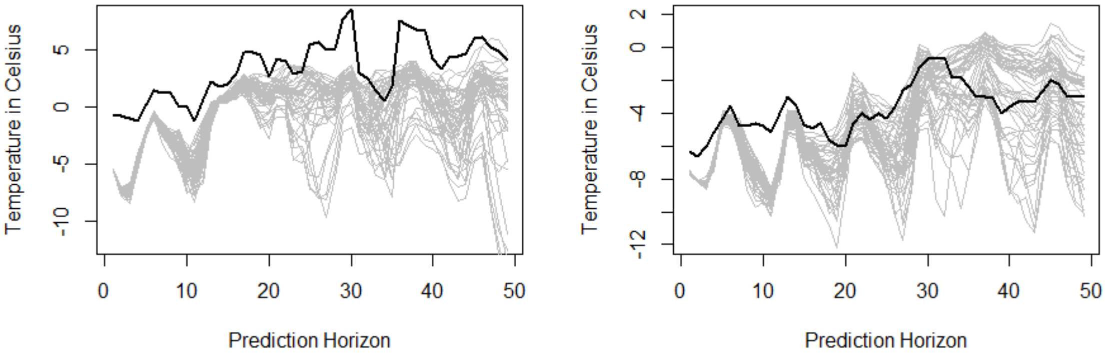

Figure 4.1: Examples of temperature forecasts (grey) and corresponding observation curves (bold black).

In the left part of Figure 4.1, the forecasting curves are significantly biased such that the observation curve is the most outlying curve in the mbd-concept. In the right figure the forecast is biased for the first 20 time points and overdispersed afterwards. After around time period 20 the observation curve is no longer clearly the most outlying curve. However overall we would expect the observation curve to be one of the most outlying curves. After seeing these two examples we would not expect the forecasts to be calibrated. Figure 4.2 shows the mbd-rank histogram for all 36500 observations.

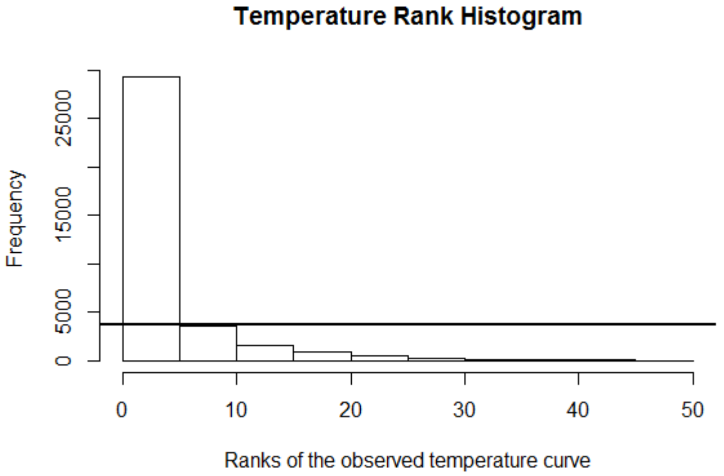

Figure 4.2: Mbd-rank-histogram of the temperature data. The horizontal black line indicates the height of a hypothetical histogram of 36500 observations drawn from a uniform distribution.

The histogram in Figure 4.2 is clearly not uniform and thus we reject calibration of the ECMWF temperature ensemble forecasts. The reliability index $\varDelta$ has a value of 0.99 which is far from 0 which is a further evidence for the non-uniformity.

The examples in Figure 4.1 indicate that the diurnal temperature changes are correctly predicted while there seems to be a fairly constant bias in the forecast. For this reason, we also investigated the calibration of bias-corrected forecasts where we bias-corrected each lead-time independently based on the 30 most recent forecast and observation pairs at each location. However, the resulting mbd-rank histogram did not differ much from the results of Figure 4.2. Therefore, we have chosen not to show these results here.

Finally, we compare the mbd-rank and the mv-rank for this data set. We have seen that the mv-rank performs considerably worse for higher numbers of total prediction horizons. Because of that we first only look at the forecasts reaching 1 day into the future. This means there are only 8 prediction horizons, since we have a prediction for every 3 hours. We compute the mv-ranks and the mbd-ranks for the first 8 and then for the last 8 prediction horizons out of all 49 prediction horizons. Figure 4.3 shows the rank histograms.

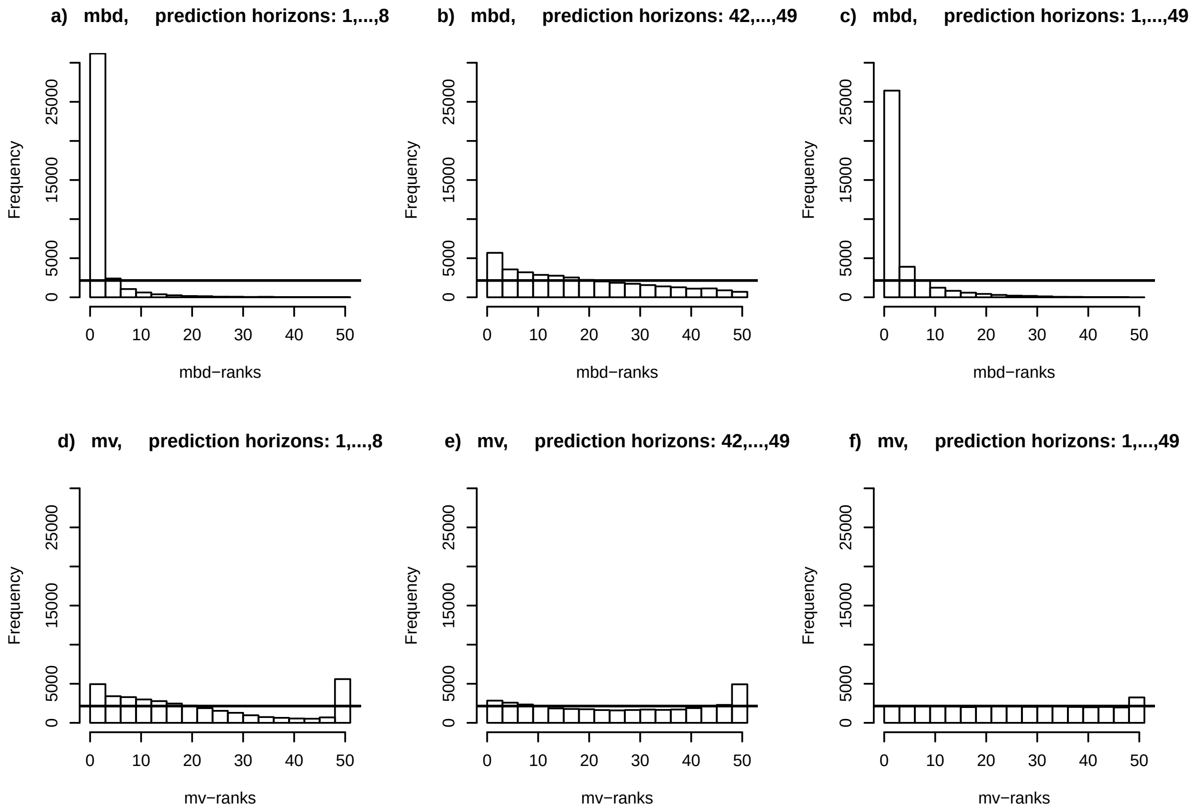

Figure 4.3: Mbd- and mv-rank histograms for different prediction horizons for the temperature data. In the first row ranks have been computed using the mbd-concept whereas ranks in the second row have been computed using the mv-concept. In the first two columns only the 8 prediction horizons, as specified, have been used, out of the total 49 prediction horizons.

The variability in the forecasts based on the different numerical methods grows with growing prediction horizons, which can also be seen in the two examples in Figure 4.2. Because of that we expect the histograms in Figure 4.3 b) and e) to look more uniform than a) and d). This is indeed the case. Furthermore for 8 prediction horizons we see that the histograms d) and e) based on the mv-rank are able to visually reject calibration but the histogram in f) for all prediction horizons does look uniform except for a rise for the very highest ranks. As in our simulation study in Chapter 2.6 we see that the mv-rank concept is not suitable for detecting non-calibration for high numbers of prediction horizons. However, due to the rise in the histogram for mv-ranks between 49 and 51 we could argue that the forecasts tend to underpredict temperature.

4.2 Inflation forecasting

The second real data example are U.S. inflation forecasts. We refer to inflation as quarter to quarter change of the consumer price index expressed in annualised percentage points. The forecasting data is obtained from The Survey of Professional Forecasters (SPF) of the Federal Reserve Bank of Philadelphia. SPF is a quarterly survey of a wide range of macroeconomic variables and the oldest survey of quarterly macroeconomic forecasts in the United States ranging back to 1968 (Federal Reserve Bank of Philadelphia 2012). Each quarter, university professors as well as private sector economists are asked to give predictions of a number of macroeconomic variables for the current quarter and each of the following 4 quarters (Gneiting and Thorarinsdottir 2010). We only use the inflation forecast data from the third quarter of 1981 to the second quarter of 2010. The number of forecasters changes every quarter ranging from 10 to 54 with a mean close to 33 for the time period of interest.

Because of this the interpretation is slightly different compared to temperature forecasting. There we had the forecast data of the 50 different ensemble members for temperature forecasting for 100 stations on 365 days. In our inflation setting the different ensemble members correspond to the different forecasters, but these forecasters change over time. So we are not able to check whether the forecasts of a certain number of economic forecasters is calibrated, i.e. - in a certain way - how good their economic forecasts are, but rather whether this setting of forecasts - professionals giving economic forecasts - is calibrated. Difficulties could also arise if the quality of the forecasts is highly dependent on the specific forecasters which vary over time, so that the forecasts at different points in time could not be interpreted as identical repetitions. Because of this we need the assumption that the overall or collective quality of the forecasts given by the professional forecasters does not vary over time, which is a rather strong assumption.

We have a varying amount of forecasting curves each with a length of 5. The mbd-rank of the observed inflation is computed at 111 points in time. Due to the varying amount of forecasting curves, we need to normalize the ranks for them to be comparable for different time points. If γ is the mbd-rank of the observation curve y with m forecasting curves then $\gamma^* = \frac{\gamma}{m+1}$ is the normalized rank. Figure 4.4 shows the forecasting curves and the curve with the observed values for the last 3 quarters used in the dataset.

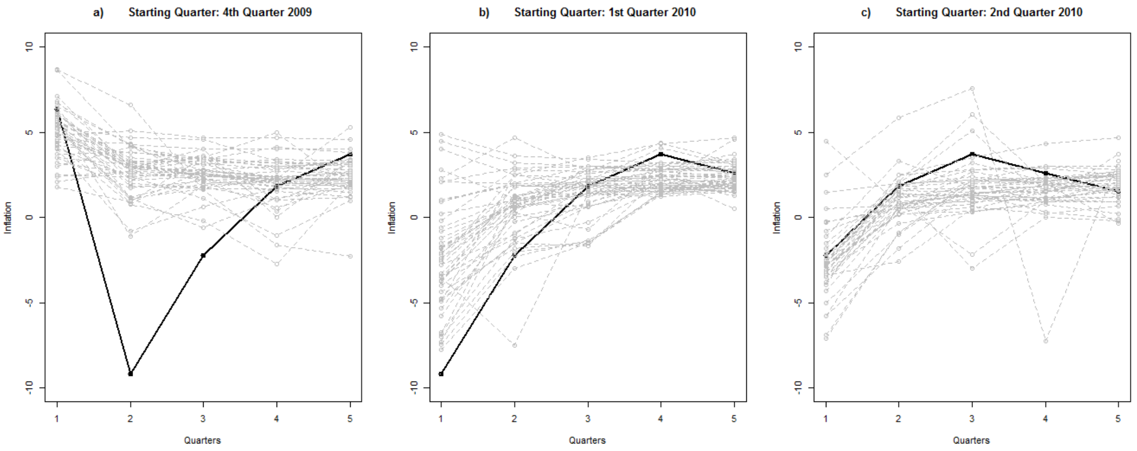

Figure 4.4: Example of the inflation forecasting and observation curves. Forecasting curves (grey curves) and observed inflation curve (bold black curve). Predicted and observed inflation values are shown beginning with the quarter in the headline as the quarter the forecasts were issued and the following four quarters.

Figure 4.4 a) shows that the very unusual high deflation of nearly 9% in the first quarter of 2010 could not be predicted by any of the forecasters during the 4th quarter of 2009. Figure 4.4 b) shows that all in all the forecasters got the direction of the future inflation right but the observation curve still seems to be the most non-central curve. From Figure 4.4 c) we should expect that this time the observation curve does not have te be the most outlying curve but it is still rather outlying. From these three examples we should not be surprised if the forecasts are not calibrated with low ranks appearing more frequently than they should in a uniform distribution. Figure 4.5 shows the rank-histogram for the normalized ranks.

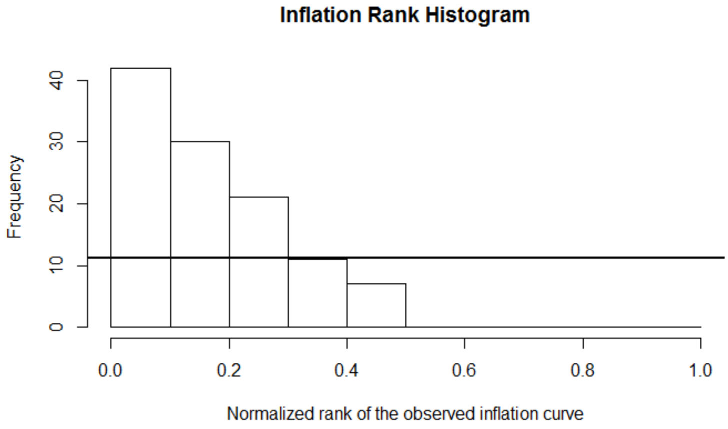

Figure 4.5: Mbd-rank histogram of the inflation data with normalized ranks. The horizontal black line indicates the height of a hypothetical histogram of 111 observations drawn from a uniform distribution.

The rank histogram clearly suggests that the inflation forecasts of the SPF are not calibrated. Lower ranks occur more often than higher ranks, which do not occur at all for normalized ranks above 0.5. A normalized rank of 0.5 means that half of the forecasting curves are deeper and half of them are more outlying than the observation curve. This means that the observation curve never was one of the 50% most deepest curves and that it belonged most frequently to the most outlying curves.

Because of the varying amount of forecasting curves it is not possible to simply compute the reliability index defined as in Chapter 2. Instead of comparing the frequency f_j of rank j with the theoretical value, i.e. the probability that rank j occurs if the ranks are drawn from a discrete uniform distribution, we now compare the frequencies we get from the histogram with 10 breaks to the theoretical probability of $\frac{1}{10}$. The result $\Delta = 1.07$ also indicates that the inflation forecasts are not calibrated.

Chapter 5

Conclusions

We have introduced a new concept of calibration for functional data, based on the mbd-rank. It is based on an center-outwards ordering of functions, the modified band depth (mbd), introduced by López-Pintado and Romo (2009).

The finite dimensional functions are also interpretable as points in $\mathbb{R}^d$. For points in $\mathbb{R}^d$ Gneiting et al. (2008) already introduced a concept of calibration, based on the multivariate rank (mv-rank). The mv-rank was introduced for applications with small dimension d in mind, but the concept also holds for higher dimensions. This means that both concepts can be applied for finite dimensional functions, or in other words points in $\mathbb{R}^d$. Because of that we have compared both rank concepts in simulation studies regarding the correct rejection of calibration. For each d ($d = 2, 5, 20$) the mbd-rank was able to detect non-calibration successfully and clearly with the clearness only increasing with increasing d. This means that the mbd-rank concept is suited for all dimensions from very low to high. The mv-rank concept on the other hand was not able to detect miss-calibration at all in our simulation study for higher dimensions (in simulation study: $d = 20$) and had some problems for medium-sized dimensions (in simulation study: $d = 5$). For very small dimensions (in simulation study: $d = 2$) the mv-rank concept was able to detect miss-calibration successfully. For dimensions of five or higher the mbd-concept should be chosen when checking for calibration. Both concepts work fine for very small dimensions, but the mv-rank can be preferred because more information, regarding the type of non-calibration, can be drawn out of the histograms based on the mv-rank.

In a second simulation study we have evaluated how good the theoretical uniformity of the histograms, measured by the reliability index, holds in practice when the forecasts are indeed calibrated. The result shows that the deviation from uniformity only depends on the number of repetitions and the ensemble size and not on the total number of prediction horizons or from which distribution we draw the simulated data for a single prediction horizon.

We have also used the new mbd-rank concept on two real data examples, inflation forecasting and temperature forecasting, to check for calibration. In both cases we had to reject calibration clearly. We have also computed the mv-ranks for the temperature forecasting example and, as expected from the simulation study, the mv-rank had problems detecting that the temperature forecasts are not calibrated.

Knüppel (2011) also focuses on calibration of multi-step-ahead forecasts. But unlike our approach it is based on multi-step-ahead density forecasts. Jordà and Marcellino (2010) focus on path forecasts and a confidence band for those paths constructed with the joint predictive density, which can also be used to evaluate local internal consistency.

Bibliography

[1] Anderson JL, 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518-1530.

[2] Dawid AP, 1984: Statistical theory: The prequential approach (with discussion and rejoinder). *J. Roy. Stat. Soc.*, **147A**, 278-292.

[3] Delle Monache L, Hacker JP, Zhou Y, Deng X, Stull RB, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res*, **111**, D24307.

[4] Federal Reserve Bank of Philadelphia, 05.04.2012: Survey of Professional Forecasting. URL: `http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/`

[5] Gneiting T, 2008: Editorial: probabilistic forecasting. *J. Roy. Stat. Soc.*, **171 A**, 319-321.

[6] Gneiting T, Balabdaoui F, Raftery AE, 2007: Probalistic forecasts, calibration and sharpness. *J. Roy. stat. Soc.*, **69B**, 243-268.

[7] Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA, 2008: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, **17**, 211-235.

[8] Hamill TM, and Colucci SJ, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327.

[9] Jordà O, and Marcellino M, 2010: Path forecast evaluation. *J. Appl. Econ.*, **25**, 635-662.

[10] Knüpell M, 2011: Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *working paper.*

[11] Liu R, 1990: On a notion of data depth based on random simplices. *Ann. Stat.*, **18**, 405-414.

[12] López-Pintado S, and Romo J, 2009: On the concept of depth for functional data. *J. Ame. Stat. Assoc.*, **104**, 718-734.

[13] Molteni F, Buizza R, Palmer TN, Petroliagis T, 1996: The new ECMWF ensemble prediction system: methodology and validation. *Q. J. Roy. Meteor. Soc.*, **122**, 73-119.

[14] Sun Y, Genton MG, 2011: Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics*, **23**, 54-64.