Evaluation of Probabilistic Forecasts: Conditional Auto-calibration

Alexander Tsyplakov

Department of Economics, Novosibirsk State University

January 31, 2020

Abstract

The paper explores theoretical foundations behind checking calibration of a probabilistic forecast. The emphasis is on a situation when the forecast examiner possesses only partially the information that was available and was used to produce the forecast. We argue that in such a situation the concept of ideal calibration is not directly applicable and the forecast should be judged by its conditional auto-calibration. Necessary and sufficient conditions of auto-calibration are discussed and expressed in the form of testable moment conditions. The paper also analyzes relationships between forecast calibration and forecast efficiency.

Key words: probabilistic forecast; forecast calibration; moment condition; probability integral transform; orthogonality condition; scoring rule; forecast encompassing.

JEL classification: C53; C52.

1 Introduction

There is little doubt that it is important for users of economic forecasts to have information on the degree of forecast uncertainty and probabilities of different scenarios. In general point forecasts give insufficient information to a user who needs to make a decision. This is the reason for the growing popularity of complete probabilistic forecasts in the form of an entire predictive probability distribution. Such a forecast can be represented by a probability density function (an example is the Bank of England density forecast of inflation; Britton, Fisher, and Whitley, 1998) or a probability mass function in the case of a discrete target variable. Other kinds of forecasts such as predictive quantiles can be viewed as derivatives of complete forecasts.

Real-life forecasts are not perfect and we want to be able to diagnose imperfections. This paper explores the theory which can help in evaluation of the quality of complete probabilistic forecasts. We

consider two key requirements for a probabilistic forecast: calibration and efficiency. *Calibration* means conformity between the forecast and the actual behavior of the target variable. This idea of conformity is vague and we give an accurate formulation below. *Efficiency* (also called optimality and rationality in different contexts) is a requirement that the forecast maximizes some performance measure which would typically reflect the objectives of potential forecast users. The current paper considers theoretical implications of these two requirements and provides ideas for designing the corresponding statistical tests which can be used for forecast evaluation. The goal of testing calibration (or efficiency) is to be able to improve forecasting methods. Even if immediate improvement is not feasible, we would like to be aware of the problems and to have some information about the nature and the degree of miscalibration.

Several empirical procedures were used for testing forecast calibration and efficiency in the literature on probabilistic forecasts and closely related interval/quantile forecasts. For example, see Kupiec (1995), Diebold, Gunther, and Tay (1998), Christoffersen (1998), Diebold, Tay, and Wallis (1999), Berkowitz (2001), Clements and Taylor (2003), Wallis (2003), Engle and Manganelli (2004), Clements (2006), Mitchell and Wallis (2011), Chen (2011), Galbraith and van Norden (2011), Knüppel (2015). However, most of these procedures are applicable only in a narrow class of forecasting situations, primarily when one-step-ahead forecasts of a time series are made given the full previous history of this series. The literature does not provide a comprehensive general picture of testable implications of forecast calibration/efficiency. Even the conditions under which one can call a forecast calibrated or efficient are not yet fully understood and formally stated.

In this paper we approach to the task of calibration testing from the fundamentals. We believe that before considering statistical procedures for forecast evaluation it is necessary to state clear definitions and derive important implications. Such approach differs from most of the existing literature, which concentrates on statistical testing (Corradi and Swanson, 2006c is a vivid example). It turns out that for many illuminating theoretical results one can treat forecasting as a one-shot activity, rather than repeated one, which is subject to statistical procedures. The focus is on a situation when a forecast is evaluated by a partially informed external examiner. This is an important circumstance since forecasters are not always impartial to their own forecasts and sequences of forecasts are often evaluated externally (some examples are Wallis, 2003; Clements, 2004, 2006; Engelberg, Manski, and Williams, 2009; The National Institute of Economic Research, 2013).

To catch the idea of the approach employed in this paper consider the example of familiar point forecasting under quadratic loss. It is well known that the conditional mean with respect to an information set Ψ is the best in mean-square sense of all the Ψ-measurable point forecasts. From the properties of

conditional mean it follows that the efficient forecast must be unbiased and the forecast error must be uncorrelated with any Ψ-measurable variables. These theoretical properties lead to corresponding test procedures, for example, Mincer-Zarnovitz-type regression-based tests (Mincer and Zarnowitz, 1969). A similar approach can be applied to complete probabilistic forecasts.

We consider forecasting of some target outcome y, which is a real-valued random variable. A complete probabilistic forecast of y is represented by a random function $\hat{F}$ defined on the entire real line and possessing the usual properties of a cumulative distribution function (CDF).[1] Forecast evaluation must rely on some relevant information represented by an information set Ψ. Formally, Ψ is a sub-σ-algebra in the underlying probability space.

For a point forecast judged by the quadratic loss it is important to correctly represent the central point of the conditional distribution of y given the relevant information set, which is achieved when the forecast coincides with the conditional mean. Similarly, for a probabilistic forecast it is important to be calibrated (Diebold, Hahn, and Tay, 1999; Gneiting, Balabdaoui, and Raftery, 2007). Several different modes of calibration were considered in the literature: probabilistic calibration (PIT uniformity), marginal calibration and ideal calibration with respect to an information set (Gneiting, Balabdaoui, and Raftery, 2007; Gneiting and Ranjan, 2013). Also very popular is the condition of uniformity and independence of PIT values (Diebold, Gunther, and Tay, 1998; Mitchell and Wallis, 2011).

A fundamental mode of calibration is *ideal calibration* requiring that $\hat{F}$ is the conditional CDF of y given the information set Ψ. In a sense, it is comprehensive and underlies the current econometric literature on probabilistic forecasting with its reliance on model-based forecasts. However, methodological and practical considerations suggest a different (though closely related) concept of calibration.

In a forecast evaluation situation one should distinguish (at least) two different parties: the forecaster and the individual who evaluates the forecast. The later party will be called the *examiner* here. The information sets of the forecaster and the forecast examiner can be distinct, say, Ψ^* and Ψ. The concept of ideal calibration is ambiguous without specifying the information set. The forecast, which is calibrated with respect to Ψ^*, can be miscalibrated with respect to Ψ and vice versa.

If, for example, an external examiner tries to evaluate forecasts produced by a government agency, he usually can use only the published statistics, but not the internal information available to the agency. To be comprehensive enough, the theory of forecast evaluation should not exclude the possibility that a forecaster uses not publicly accessible, non-obvious or irrelevant information, which is not in the examiner's

[1] We work with CDF, because it fully characterizes the distribution of any real-valued target variable, whether continuous or discrete.

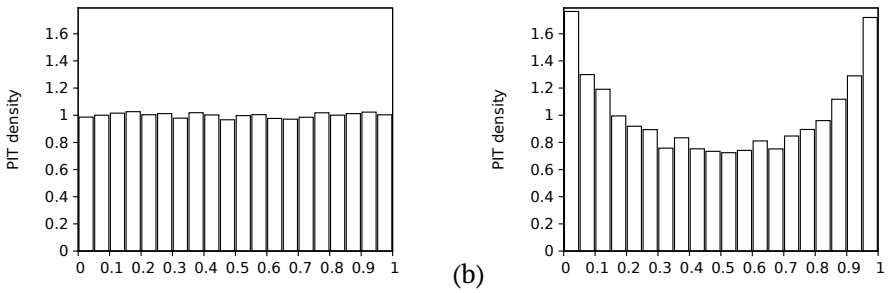

Figure 1: Histograms of PIT values for forecast $\hat{F}_{r2}$ of subsection 4.3: (a) unconditional; (b) conditional given $|x| < 1$ and $\mathrm{IQR}(\hat{F}_{r2}) > 1.4$. Sample size is 100,000, the sub-sample size given the condition is approximately 30,000.

information set. Even in the case of academic forecasters, who accurately describe their models and data sources, an external examiner is not always able to ascertain the quality and integrity of the forecasting process. In general, an examiner needs procedures, which can convincingly demonstrate imperfections of a miscalibrated forecast irrespectively of its origin.

Our main idea is that it is convenient and illuminating to introduce a notion of calibration, which explicitly takes into account the fact that not only the examiner's information set, but also the forecast itself can be the source of information for the forecast examiner. We call this mode of calibration *conditional auto-calibration*: forecast $\hat{F}$ is auto-calibrated if it coincides with the conditional CDF of y given the information set Ψ *and the forecast itself.* In the case of point forecasting under quadratic loss a similar idea can be expressed as follows: a point forecast $\hat{y}$ is efficient (or rational) if it coincides with the conditional mean of y given the relevant information set Ψ and itself, that is $\mathsf{E}[y|\Psi, \hat{y}] = \hat{y}$. Only when the examiner is fully informed and thus $\hat{y}$ is Ψ-measurable we can safely condition on the information set alone and reduce the definition to $\mathsf{E}[y|\Psi] = \hat{y}$.

Figure 1 illustrates the importance of using adequate conditioning in forecast evaluation. The illustration is based on the simulation example from subsection 4.3. The unconditional histogram of PIT values does not show apparent signs of non-uniformity even with sample size as large as 100,000. However, for observations with relatively small values of x and wide predictive distributions (as measured by interquartile ranges) the histogram is evidently non-uniform and signals underdispersed predictive distributions. This demonstrates that an examiner can detect forecast miscalibration by combining his own information (variable x) with information contained in the forecast.

For practical reasons it is convenient to express the implications of calibration in terms of moment conditions in the spirit of the generalized method of moments (GMM). Such conditions relate two different kinds of moments: the moments defined on the underlying probability space and the moments

calculated from the forecast-based probability measures. The paper states and discusses various necessary and sufficient conditions of calibration and expresses them in the form of moment conditions, which can further be used to design calibration tests. Moment conditions approach provides a more general and accurate way of conducting conditional analysis similar to the graphical analysis of Figure 1(b).

If a forecast examiner tests moment conditions using functions from some inadequately narrow class, then some aspects of miscalibration, which are potentially detectable, cannot be revealed using any function from the class. Thus, it is important (1) to formulate a comprehensive class of functions producing moment conditions of auto-calibration, and (2) to find out which contractions of this class seeming reasonable lead to non-comprehensiveness, and which do not. For example, using a general class of functions corresponding to *orthogonality conditions* does not lead to a loss of comprehensiveness, while other functions correspond only to *conditional marginal* or *probabilistic calibration*, which, even together, do not entail conditional auto-calibration.

Calibration is a property of probabilistic forecast, which is impersonal and easier to test, but there are also objectives of a forecast user. It is important to link the notion of calibration with the notion of forecast efficiency. We call a forecast efficient if it maximizes a pertinent performance measure based on a *scoring rule*. First of all, miscalibration can be an indication of inefficiency. However, one can also formulate more direct conditions of efficiency and relate them to calibration. Additionally, the *sharpness principle* of forecasting conjectured in Gneiting, Balabdaoui, and Raftery (2007) happens to rely on the notions of scoring rules and auto-calibration.

When there are multiple forecasts, there is no need for calibration testing to choose between them as they can be compared on the basis of their average scores. However, existence of several alternatives gives a possibility to test calibration of one forecast using the information from another one. This leads to the notion of *forecast encompassing* and the corresponding moment conditions and tests.

Section 2 analyzes in detail the notion of calibration and characterize it by moment conditions. Calibration testing is not of primary interest in this paper; however, subsection 2.8 gives a reader basic ideas about relevant statistical procedures. Section 3 discusses forecast efficiency from the point of view of proper scoring rules and analyzes the links between calibration and efficiency. Section 4 provides illustrative examples. Section 5 concludes. Theorems are placed in Appendix A.

2 Forecast calibration

2.1 Basic definitions

In a typical forecast evaluation situation we have a sequence of predictive distributions $\hat{F}_1, \dots, \hat{F}_N$ (in the form of CDFs) and a sequence of points $y_1, \dots, y_N$ corresponding to actual realizations of the target variable. The task is to compare one to the other and make a conclusion about the conformity between the two (i. e. about forecast calibration). Attached to each forecast-outcome pair $\hat{F}_i, y_i$ there is the information set Ψ_i, which is relevant for this pair and can be used by the forecast examiner for evaluation purposes.

As an example consider one-year-ahead forecasts of the Swedish CPI inflation which were published by the Riksbank (figure 2). The forecasts are in the form of the two-piece normal distribution. The actual inflation is shown by triangles. Behind each forecast is the information used by the Riksbank to obtain it. An external forecast examiner cannot know all this information. He can only use the information which is both available to him and was potentially accessible to the Riksbank's staff, such as any official statistics published before the date of forecast issue.

The current paper approaches the task of calibration testing from the fundamentals. Before considering any applied procedures for calibration testing, which contrast forecasts with outcomes, we give a formal definition of calibration for a one-shot forecasting situation (a single forecast-outcome pair). Then the definition is rendered into testable conditions. Finally, such conditions form the basis for statistical procedures for observed sequences of forecast-outcome pairs.

We start by defining ideal calibration for a target variable y, a CDF-valued variable $\hat{F}$ (forecast) and an information set Ψ.[2] Assume that Ψ includes all the relevant information which can be used. The intuition is that for a given information set Ψ the ideally calibrated forecast, first, is based only on Ψ without employing any other information (formally, the forecast $\hat{F}$ is Ψ-measurable) and, second, fully utilizes Ψ. In other words, it is the best achievable forecast among forecasts based on Ψ (see subsection 3.1 for a formal discussion). Here and below we use $\mathbb{F}$ to denote the (conditional) distribution function of y.

Definition 1. A forecast $\hat{F}$ is *ideally calibrated given* Ψ if it is the conditional distribution function of y given Ψ, that is, $\hat{F}(q) = \mathbb{F}(q|\Psi)$ for $q \in \mathbb{R}$.

There are many real-life situations in which the information sets of the forecaster and the forecast

[2] The concept of ideal calibration with respect to an information set is quite natural and is implicit in the literature on probabilistic forecasting, albeit, possibly, in a non-direct fashion—like "conditional density governing a series", "true data generating process" in Diebold, Gunther, and Tay (1998). Explicit definitions can be found in Tsyplakov (2011) and Gneiting and Ranjan (2013). It is also similar to the definition of interval forecast efficiency with respect to an information set in Christoffersen (1998).

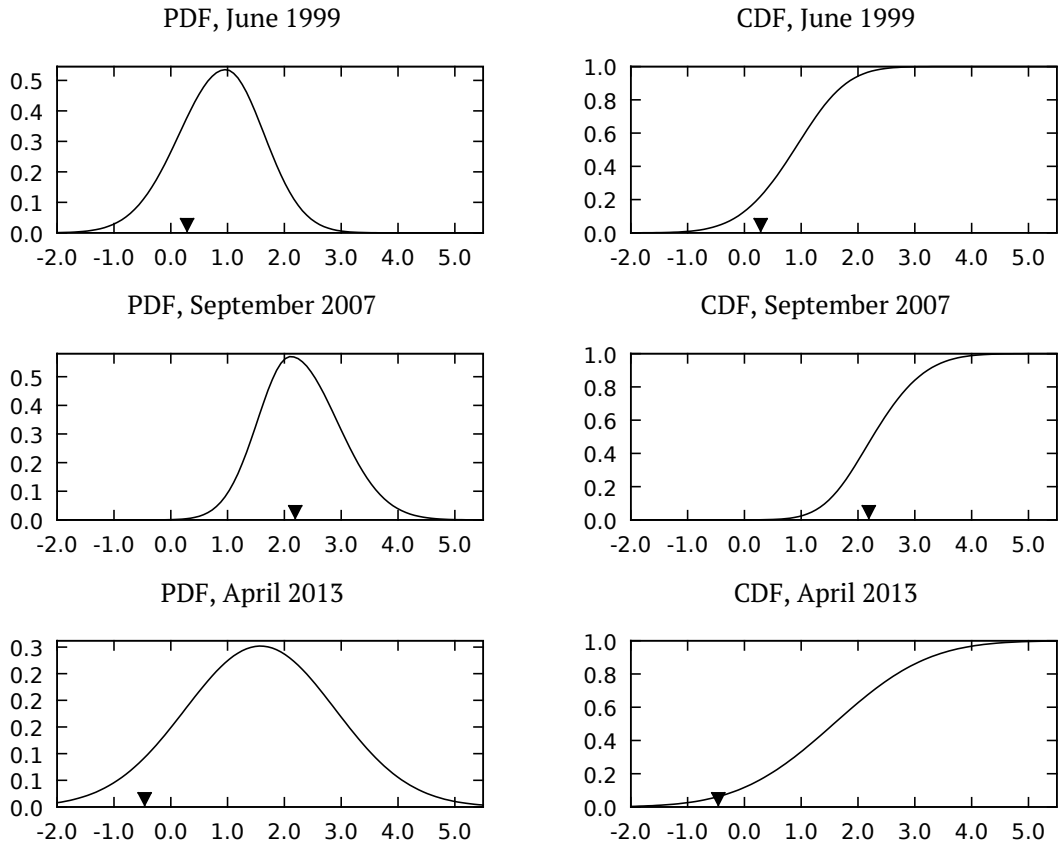

Figure 2: Riksbank's one-year-ahead CPI inflation forecasts. The forecasts are for the specified dates. Left graphs show the forecasts in the form of probability density function, right ones—the same forecasts in the form of cumulative distribution functions. Actual outcomes of the CPI inflation are marked by triangles.

examiner are not the same. For example, the central bank or the government can use internal information which is not publicly available. A better informed forecaster can produce forecasts which are in some sense better than the ideal forecast given examiner's information set, while still not fully calibrated. Sometimes partially informed examiner can be able to detect this miscalibration.

It is also not uncommon for a forecast to include subjective judgment of the forecaster or utilize non-obvious sources of information. The leading motivating example is that of survey forecasts (like Survey of Professional Forecasters, see Diebold, Tay, and Wallis, 1999, Clements, 2006, Engelberg, Manski, and Williams, 2009). As a more exotic example consider a forecaster using sunspot activity to predict stock prices. One can even recall Roman augurs using observation of birds' behavior for foretelling. There also exists a more scientific reason for using extraneous noise, namely, the use of Monte Carlo methods in estimation and/or prediction.

If the forecaster makes use of some information which is not available to the examiner, then the

examiner can potentially derive some new information from the forecast itself.[3] Thus, in this case the information set relevant to forecast evaluation combines the examiner's prior information with the information delivered by the forecast, and we can state that forecast $\hat{F}$ is calibrated from the examiner's point of view if it coincides with $\mathbb{F}(\cdot|\Psi, \hat{F})$, which is the conditional distribution function of y given Ψ and $\hat{F}$. In the following definition and below $\sigma(\cdot)$ is an operator which combines information sets and random elements into a properly constructed information set (the generated σ-algebra).

Definition 2. A forecast $\hat{F}$ is *conditionally auto-calibrated given* Ψ if it is ideally calibrated with respect to $\sigma(\Psi, \hat{F})$, that is, $\hat{F}(q) = \mathbb{F}(q|\Psi, \hat{F})$ for $q \in \mathbb{R}$.

By definition any auto-calibrated forecast is ideally calibrated with respect to an appropriate information set. Moreover, a Ψ-measurable forecast, which is auto-calibrated given Ψ, must be ideally calibrated given Ψ. Conversely, it can be stated that any forecast, which is ideally calibrated with respect to some information set Ψ^* including Ψ, is auto-calibrated with respect to Ψ (Theorem 2). Intuitively, the information set generated by an ideally calibrated forecast contains all the contents of the original information set which is relevant for predicting the target variable. It follows from Theorem 2 that if $\hat{F}$ is auto-calibrated given Ψ_1 and Ψ_1 is a "richer" information set than Ψ_2 (that is, Ψ_1 contains all the information of Ψ_2 and maybe some additional useful information; formally, $\Psi_2 \subseteq \Psi_1$), then it is auto-calibrated given Ψ_2. In particular, conditional auto-calibration implies unconditional auto-calibration (auto-calibration with respect to the trivial information set).

Of course, one can base the theory of forecast evaluation on the definition of ideal forecast calibration. However, it is more clear and natural to concentrate on the property of conditional auto-calibration instead. The main cause for introducing this along with ideal calibration is that if forecaster's information set is not known to the examiner, the later has no way to verify that the forecast is ideally calibrated with respect to this information set. Hence the examiner in fact can only test auto-calibration (with respect to his own information set).

As discussed below, in general it is not sufficient to use popular conditions of PIT uniformity and lack of correlation between functions of PIT values and some observable variables based on Ψ to test calibration. Thus, a partially informed examiner confronted with a black-box forecast has to use specific instruments and construct peculiar variables based on both $\hat{F}$ and Ψ, which can be used in calibration testing.

[3] The use of information contained in the forecast itself to define calibration is not new to the literature (cf. Lichtenstein, Fischhoff, and Phillips, 1982; Galbraith and van Norden, 2011), but existing analysis is mostly limited to the case of the trivial Ψ and dichotomous target variable. Lichtenstein, Fischhoff, and Phillips (1982), p. 307: "Formally, a judge is calibrated if, over the long run, for all propositions assigned a given probability, the proportion true equals the probability assigned." Bröcker (2009) considers a more general case of a discrete target variable with a finite support; he uses an alternative term "reliability".

Even if the forecast under examination is known to be Ψ-measurable, these specific instruments can be utilized with benefit, because for the examiner it might not be clear how the forecast is constructed from Ψ. Moreover, even if the forecast is not a black-box one, these specific instruments can be useful, because at the technical side a forecast $\hat{F}$ is not a finite-dimensional variable which is a typical object of analysis in econometrics. There are specific aspects of using CDF-valued variables, and we illustrate these in examples below.

A final remark is pertinent here. It goes without saying that adequate choice of information set is crucial for testing calibration in applications. If an examiner wants to evaluate a forecast, then he must consider information Ψ which is available at the time the forecast was made. Further, judgments about forecaster's *rationality* can only be based on the information known to be available to this forecaster.

2.2 General moment conditions of forecast calibration

The definition of conditional auto-calibration with respect to an information set, although methodologically appealing, is too abstract. For the purpose of forecast evaluation one would like to have some functions of y, which are directly observable and could be compared to something, which is based on the forecast and the information contained in Ψ. In the case of point forecasting we can directly compare the forecasts and the actual realizations of y. One would like to have something similar in the case of probabilistic forecasting.

We start by noting that any predictive distribution $\hat{F}$ can be readily used to produce a point forecast of y. Such a forecast is given by the corresponding mean, that is,

$$\text{mean}(\hat{F}) = \int_{-\infty}^{\infty} t d\hat{F}(t).$$

This is a reasonable forecast since under assumption of conditional auto-calibration $\hat{F}$ is the conditional distribution function of y given Ψ and $\hat{F}$ and thus $\text{mean}(\hat{F})$ is the corresponding conditional mean. Then for a series of predictive distributions we can obtain a series of point forecasts. The series can be directly compared to a series of corresponding actual realizations of the target variable y using some statistical procedure. Figure 3(a) plots such series for the Riksbank's inflation forecasts mentioned above.

This observation is rather trivial. However, note further, that the same procedure can be applied to any suitable function of y. For example, let y_L be the last observed value of the annual inflation before the forecast issue. We can calculate the theoretical mean of the squared increase in inflation $(y - y_L)^2$

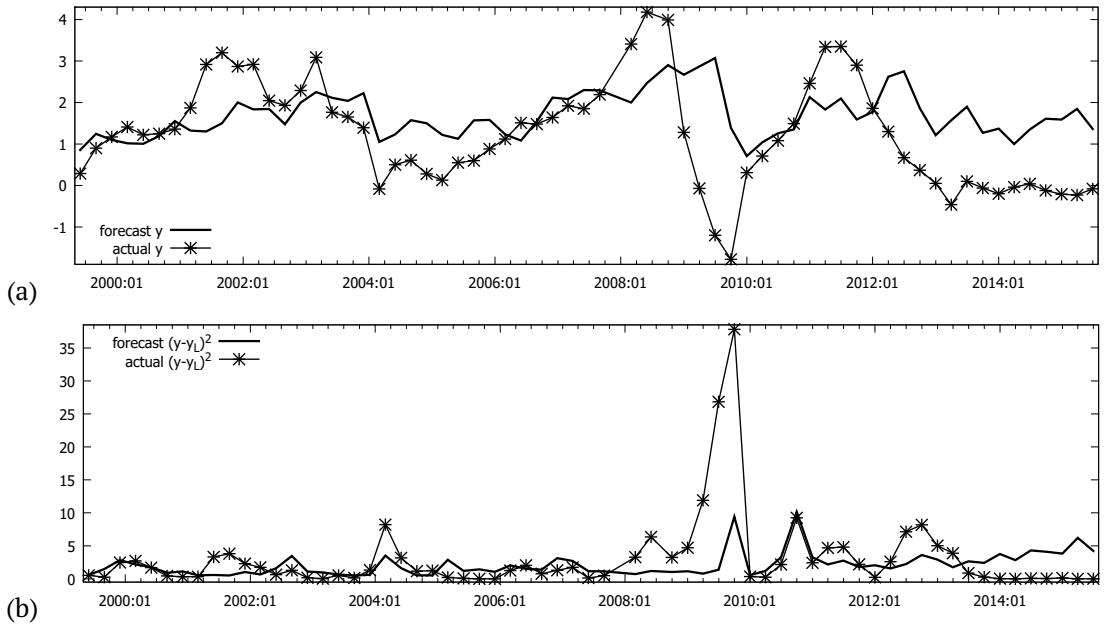

Figure 3: Forecasts of annual inflation (a) and squared increase in inflation (b) derived from the Riksbank's inflation forecasts compared to the actual outcomes.

implied by $\hat{F}$ as $\int_{-\infty}^{\infty}(t-y_L)^2 d\hat{F}(t)$, and use it as a point forecast for $(y-y_L)^2$.[4] See figure 3(b), which compares the forecasts to the actual outcomes.

Continuing this line of reasoning, if $\hat{F}$ is a good predictive distribution for y, then it should be able to predict the behavior of $g(y, w, \hat{F})$, where g is some function and w is some Ψ-measurable random element. We can treat $\hat{F}$ and w as fixed, since they are assumed to be already known at the forecasting time, and use the expectation of g under the assumption that y is distributed according to $\hat{F}$ as our forecast of g. Such a forecast must coincide with the conditional expectation in the underlying probability space, because under conditional auto-calibration $\hat{F}$ is the conditional CDF of y given Ψ and $\hat{F}$.[5] That is,

$$\mathsf{E}[g(y, w, \hat{F})|\Psi, \hat{F}] = \int_{-\infty}^{\infty} g(t, w, \hat{F}) d\hat{F}(t). \tag{1}$$

By the law of iterated expectations this suggests a very general type of moment conditions of calibration:

$$\mathsf{E}g(y, w, \hat{F}) = \mathsf{E}\hat{g}(w, \hat{F}), \qquad \text{where} \quad \hat{g}(w, F) = \int_{-\infty}^{\infty} g(t, w, F) dF(t), \tag{2}$$

for any g and any Ψ-measurable w. Here we use $\hat{F}$ to obtain a prediction $\hat{g}$ of function g and the expectation of this prediction must be the same as the actual expectation of g.

[4] The idea of this is that y_L is a naive forecast of future inflation and $(y-y_L)^2$ is the corresponding squared forecast error. One can test whether $\hat{F}$ can predict this loss value for y_L. Thus, the comparison of actual and predicted $(y-y_L)^2$ is a kind of forecast encompassing considered in subsections 2.7 and 3.4.

[5] Cf. Theorem 6.4 (*disintegration theorem*) in Kallenberg (2002), p. 108.

For the example of Figures 3(a) and (b) we have $g = y$ and $g = (y-w)^2$ for $w = y_L$ with corresponding functions $\hat{g} = \text{mean}(\hat{F})$ and $\hat{g} = \int_{-\infty}^{\infty}(t-w)^2 d\hat{F}(t)$.

In practice one can represent a CDF F in function $g(y, w, F)$ by some characteristics of the corresponding distribution such as the mean, median or interquartile range. Ψ-measurable variable w is some univariate or multivariate variable which could be used by the forecaster and is known to the examiner.

In fact, (2) are the most general moment conditions of calibration since they enclose all the aspects of a forecasting situation: the target variable, the forecast and the information set. As discussed below, they are not only necessary, but also sufficient for conditional auto-calibration (see subsection 2.6). The moment conditions of this kind can be a basis for statistical procedures implementing calibration testing (see subsection 2.8).

Conditions (2) use a rather general class of functions g. It is both theoretically and practically interesting to find out whether some natural contractions of this class can lead to a loss of sufficiency or not. Dropping w from g produces a condition of unconditional auto-calibration, which in general is not sufficient for conditional auto-calibration. In a density forecasting situation dropping y and $\hat{F}$ from g and replacing them by the PIT variable $\hat{F}(y)$, that is, letting $g = c(\hat{F}(y), w)$, produces a condition of conditional probabilistic calibration introduced below in subsection 2.3. Dropping $\hat{F}$ from g, that is, letting $g = n(y, w)$, produces a condition of conditional marginal calibration from subsection 2.4. Subsection 2.5 discusses orthogonality conditions, which are produced by letting $g = r(y, \hat{F})a(w, \hat{F})$. In subsection 2.6 it is explained that probabilistic and marginal calibration given Ψ are not sufficient for auto-calibration given Ψ, while general orthogonality conditions are sufficient. We can even replace g in (2) by $g = r(y, \hat{F})w$, $g = m(y)a(w, \hat{F})$ or $g = k(\hat{F}(y))a(w, \hat{F})$ without losing comprehensiveness.

2.3 Conditional probabilistic calibration (PIT uniformity)

Consider a situation when forecasts are such that their values have a constant support $[a, b]$ with possibly infinite bounds, continuous and strictly increasing over $[a, b]$. (The target variable y is implicitly assumed to have conditional CDFs with similar properties). We loosely call this setting a density forecasting situation. For a random variable y with the cumulative distribution function F the probability integral transform (PIT) value is defined as $F(y)$. It has the $U[0, 1]$ distribution if F is continuous. In the same manner one can define the PIT value for a probabilistic forecast $\hat{F}$ as $\hat{F}(y)$.[6] The PIT value $\hat{F}(y)$ is a random variable which, given the forecast, carries the same information as the target variable y. This

[6]The notion of PIT can be extended to arbitrary real-valued distributions by introducing randomization (e.g. Ferguson, 1967; Brockwell, 2007). One can extend the results of the current paper in this direction, but we prefer not to do so in order to keep the exposition more transparent.

is true in a density forecasting situation irrespective of the forecast $\hat{F}$. (While this property seems very natural, its proof requires some manipulations of σ-algebras; see Theorem 1 in Appendix A.)

The quantity $\hat{F}(y)$ is the one that is used most often for calibration diagnostics in econometrics; e.g. Diebold, Gunther, and Tay (1998); Mitchell and Wallis (2011); Chen (2011). Probabilistic calibration is a mode of calibration based on these PIT values. The term "probabilistic calibration" for the condition of PIT uniformity was suggested in Gneiting, Balabdaoui, and Raftery (2007); see also the reformulation of this definition in Gneiting and Ranjan (2013).[7] Here we introduce a conditional version of this mode of calibration.

Definition 3. A forecast $\hat{F}$ is *probabilistically calibrated given* Ψ if $\hat{F}(y)|\Psi \sim U[0,1]$.

This condition can be decomposed into two conditions, namely, that, first, PIT values $\hat{F}(y)$ are unconditionally distributed as $U[0,1]$ and, second, $\hat{F}(y)$ and Ψ are independent.

Unconditional PIT uniformity can be assessed, for example, with the help of a histogram of the PIT values on the $[0,1]$ interval. The histogram should be almost flat (e.g. Diebold, Gunther, and Tay, 1998). In Wallis (2003) a corresponding formal goodness-of-fit test is proposed. See also Knüppel (2015), where a family of useful tests of unconditional probabilistic calibration is proposed.

It can be seen that the concept of probabilistic calibration is closely connected to interval forecasting and quantile forecasting (e.g. value-at-risk forecasting).[8] For a forecast $\hat{F}$ we can define the corresponding α-quantile forecast $Q_\alpha = \hat{F}^{-1}(\alpha)$. Under correct calibration the probability that $\hat{F}(y)$ does not exceed α should be equal to α. Consequently the probability that y does not exceed Q_α should also be equal to α.

More generally, under conditional probabilistic calibration in a density forecasting situation we have for any suitable function c and any Ψ-measurable w

$$\mathsf{E}c(\hat{F}(y), w) = \mathsf{E}\int_0^1 c(p, w)\,dp. \tag{3}$$

This property is not only necessary, but also sufficient, which can be seen from setting $c(p, w) = \mathrm{I}\{p \leq \alpha\}w$ and using indicator random variables for events $A \in \Psi$ as w: $w = I_A$.[9] Moment conditions (3) are

[7] The definitions of probabilistic and marginal calibration proposed in Gneiting, Balabdaoui, and Raftery (2007) are formulated from a prequential perspective due to Dawid (1984) for sequences of forecasts. In Gneiting and Ranjan (2013) the one-shot view on the forecasting theory is employed, similar to that of the current paper.

[8] That is why the literature in this area such as Kupiec (1995), Christoffersen (1998), Lopez (1998), Clements and Taylor (2003), Engle and Manganelli (2004) can be considered as a part of the literature on probabilistic forecasting.

[9] Recall that by definition $\hat{x}$ is the conditional expectation of x given Ψ, if $\hat{x}$ is Ψ-measurable and $\mathsf{E}[(x - \hat{x})I_A] = 0$ for any $A \in \Psi$. If $\mathsf{E}[(\mathrm{I}\{\hat{F}(y) \leq \alpha\} - \alpha)I_A] = 0$ for any real $\alpha \in [0,1]$ and any $A \in \Psi$, then the conditional probability of $\hat{F}(y) \leq \alpha$ given Ψ is α for any $\alpha \in [0,1]$ and thus by Definition 3 forecast $\hat{F}$ is probabilistically calibrated given Ψ.

weaker than the general moment conditions of conditional auto-calibration (2). Unconditional probabilistic calibration correspond to $c(p, w) = k(p)$ in (3) (dropping the conditioning variable w).

Further references and examples of moment conditions of probabilistic calibration can be found in subsection 2.5 and in Chen (2011).

2.4 Conditional marginal calibration

The concept of probabilistic calibration implicitly assumes a situation when probabilities are fixed while the bounds are reported by the forecaster. A reversed situation is when bounds are fixed while the forecaster reports probabilities as in the Survey of Professional Forecasters. A calibrated forecast must supply probabilities which are in accordance with the true ones. Probabilities for all possible bounds are summarized by a CDF. Thus, another mode of calibration is defined in terms of CDFs. The definition is given in Gneiting, Balabdaoui, and Raftery (2007) and reformulated in Gneiting and Ranjan (2013). Again, here we provide a conditional version of this definition.

Definition 4. A forecast $\hat{F}$ is marginally calibrated given Ψ if $\mathsf{E}(\hat{F}(q)|\Psi) = \mathbb{F}(q|\Psi)$ for $q \in \mathbb{R}$.

Similarly to conditional probabilistic calibration conditional marginal calibration can be characterized by moment conditions, which are weaker than the general moments conditions of conditional auto-calibration (2). If a forecast $\hat{F}$ is marginally calibrated with respect to Ψ, then for any function n and any Ψ-measurable w it satisfies

$$\mathsf{E}n(y, w) = \mathsf{E}\hat{n}(\hat{F}, w), \qquad \text{where} \quad \hat{n}(F, w) = \int_{-\infty}^{\infty} n(t, w)dF(t) \tag{4}$$

(see Theorem 4). That is, conditional marginal calibration implies that the prediction $\hat{n}$ produced from $\hat{F}$ is an unbiased forecast of n. This condition is not only necessary, but also sufficient, which can be seen from setting $n(y, w) = \mathrm{I}\{y \le q\}\, w$, $\hat{n} = \hat{F}(q)\, w$ and using indicator random variables for events $A \in \Psi$ as w: $w = I_A$.[10] Unconditional marginal calibration corresponds to $n(y, w) = m(y)$ in (4) (dropping the conditioning variable w).

In particular, for $n = y$ we obtain the condition of mean unbiasedness of $\hat{F}$: $\mathsf{E}y = \mathsf{E}\mathrm{mean}(\hat{F})$. Gneiting, Balabdaoui, and Raftery (2007) propose a diagnostic diagram for unconditional marginal calibration based on the difference between the empirical CDF and the mean forecast CDF, which correspond to the moment condition $\mathsf{E}\mathrm{I}\{y \le q]\} = \mathsf{E}[\hat{F}(q)]$. Note also that figure 3 above can be considered as an illustration

[10] If $\mathsf{E}[(\mathrm{I}\{y \le q\} - \hat{F}(q))I_A] = 0$ for any real q and any $A \in \Psi$, then $\mathsf{E}[(\mathrm{I}\{y \le q\} - \hat{F}(q))|\Psi] = 0$ and by Definition 4 forecast $\hat{F}$ is marginally calibrated given Ψ.

of a testing procedure for marginal calibration.

2.5 Orthogonality conditions of calibration

From the theory of point forecasting it is known that the expectation conditional on information set Ψ is the forecast which is optimal in mean-square sense among the forecast based on Ψ (e.g. Bierens, 2004, pp. 80–81). This forecast satisfies orthogonality conditions: the prediction error is uncorrelated with any random variable based on Ψ.[11] There are also extensions to the case of general loss functions (e.g. Granger, 1999). In Mitchell and Wallis (2011) an idea was put forward that calibration of probabilistic forecasts can be tested by verifying similar orthogonality conditions. It can be demonstrated that this idea lends itself to further generalization.

By letting $g = ra$ in the general moment conditions of calibration (2) we obtain the following general orthogonality conditions of calibration:[12] if a forecast $\hat{F}$ is conditionally auto-calibrated with respect to Ψ, then for any functions r and a and any Ψ-measurable w it satisfies

$$\mathsf{E}[(r(y,\hat{F}) - \hat{r}(\hat{F}))a(w,\hat{F})] = 0, \qquad \text{where} \quad \hat{r}(F) = \int_{-\infty}^{\infty} r(t,F)\,dF(t). \tag{5}$$

According to these conditions a point forecast of r derived from a probabilistic forecast $\hat{F}$ must be unbiased and the corresponding forecast error must not be correlated with any function of a Ψ-measurable w and the forecast $\hat{F}$.

An example of this type of orthogonality conditions can be found in Clements (2006), where in the context of evaluating the SPF probabilistic forecasts it was noted that $\mathsf{E}[(I-p)p] = 0$, where I is an indicator variable for the event that y is in some interval and p is the predicted probability of this event. Mincer and Zarnowitz (1969) regressions correspond to $r = y$ and $a = \text{mean}(\hat{F})$.

Note that conditional probabilistic and marginal calibration can also be characterized by orthogonality conditions, but these conditions are less general. Under conditional probabilistic calibration with respect to Ψ for any function $k(p)$ defined on the $[0,1]$ interval and any Ψ-measurable w we must have

$$\mathsf{E}\left[\left(k(\hat{F}(y)) - \int_0^1 k(p)\,dp\right)w\right] = 0. \tag{6}$$

[11] These conditions were utilized in the rational expectations literature. Shiller (1978), p. 7: "…Expected forecast errors conditional on any subset of the information available when the forecast was made, are zero… Hence, the forecast error … is uncorrelated with any element of I_t [the set of public information available at time t]".

The term "orthogonality conditions" is known from the GMM literature (cf. Hansen, 1982).

[12] Note that any random variable A which is measurable with respect to $\sigma(\Psi,\hat{F})$ can be represented as $A = a(w,\hat{F})$ for some function a, where w is a Ψ-measurable variable.

Similarly under conditional marginal calibration with respect to Ψ for any function $m(y)$ and any Ψ-measurable w we must have

$$\mathsf{E}\left[\left(m(y) - \hat{m}(\hat{F})\right) w\right] = 0, \qquad \text{where} \quad \hat{m}(F) = \int_{-\infty}^{\infty} m(t)\, dF(t). \tag{7}$$

As an example of orthogonality conditions for probabilistic calibration consider an autoregression from Berkowitz (2001). See also a regression from subsection 4.3 of Christoffersen (1998) used for testing conditional coverage of an interval forecast. A similar regression representing orthogonality conditions for marginal calibration can, for example, be found in Clements (2006). Ordinary orthogonality conditions for point forecasts can be considered in many cases as conditions of orthogonality between $y - \text{mean}(\hat{F})$ and w and thus also relate to marginal calibration.

2.6 Sufficient conditions of auto-calibration

In some sense general moment conditions (2) are complete. That is, it can be proved (see discussion of conditions (8) and (9) below) that the requirement that they are satisfied for any g and any w is sufficient for conditional auto-calibration. (In the same sense conditions (3) and (4) are sufficient for conditional probabilistic and marginal calibration respectively.) However, it is tempting to narrow these general moment conditions somehow. Both theoretically and practically interesting question is how "narrow" one can be in calibration testing without a fundamental sacrifice of comprehensiveness.

According to Theorem 3 auto-calibration given Ψ implies both probabilistic and marginal conditional calibration given Ψ. Probabilistic calibration and marginal calibration are different concepts. Neither of them generalizes the other one. Counterexamples[13] for a density forecasting situation can be found in Gneiting, Balabdaoui, and Raftery (2007) (Examples 3, 5, 6) and in Mitchell and Wallis (2011) (combined and unfocused forecasts in the AR(2) example).[14]

It can be seen that neither probabilistic, nor marginal calibration given Ψ is sufficient for auto-calibration given Ψ. Theorem 5 in Appendix A demonstrates that even when a forecast is simultaneously (unconditionally) probabilistically and marginally calibrated, it can fail to be auto-calibrated.

Of course, if we do not want to assume that the forecast examiner is only partially informed, then the

[13] One can easily generate other counterexamples. In theory an arbitrary forecast can be readily *recalibrated* (that is, modified to improve calibration) to achieve either probabilistic or marginal calibration relative to Ψ. If $G(p)$ is the conditional distribution function of PIT values $\hat{F}(y)$ given Ψ, then $G(\hat{F}(\cdot))$ is a probabilistically recalibrated version of $\hat{F}(\cdot)$. Similarly $\hat{F}(H^{-1}(\mathbb{F}(\cdot|\Psi)))$ is its marginally recalibrated version, where $\mathbb{F}(\cdot|\Psi)$ is the conditional CDF of y and $H(q) = \mathsf{E}(\hat{F}(q)|\Psi)$. In general, so recalibrated imperfect forecast is either probabilistically or marginally calibrated, but not both.

[14] Probabilistic and marginal calibration are also distinct concepts for discrete y assuming more than two values; see an example in Table 2 of Gneiting and Ranjan (2013).

distinction above is not important. If a Ψ-measurable forecast $\hat{F}$ is either marginally calibrated or (in a density forecasting situation) probabilistically calibrated with respect to Ψ, then $\hat{F}$ is ideally calibrated with respect to Ψ (see Theorem 6).

From these arguments it can be seen that both marginal and probabilistic calibration with respect to $\sigma(\Psi, \hat{F})$ are equivalent to auto-calibration given Ψ. Thus, sufficient conditions of marginal and probabilistic calibration with respect to $\sigma(\Psi, \hat{F})$ are sufficient conditions of auto-calibration. Marginal calibration with respect to $\sigma(\Psi, \hat{F})$ can be expressed in terms of orthogonality conditions between $\mathrm{I}\{y \le q\} - \hat{F}(q)$ and any function $a(w, \hat{F})$ of a Ψ-measurable w and forecast $\hat{F}$ (for any real q and any a). Similarly (in a density forecasting situation) probabilistic calibration with respect to $\sigma(\Psi, \hat{F})$ can be expressed in terms of orthogonality conditions between $\mathrm{I}\{\hat{F}(y) \le p\} - p$ and any $a(w, \hat{F})$. Hence, we have two different sufficient moment conditions of auto-calibration with respect to Ψ:

$$\mathsf{E}[(\mathrm{I}\{y \le q\} - \hat{F}(q))\, a(w, \hat{F})] = 0 \tag{8}$$

for any real q, any a and any Ψ-measurable w and (in a density forecasting situation)

$$\mathsf{E}[(\mathrm{I}\{\hat{F}(y) \le p\} - p)\, a(w, \hat{F})] = 0 \tag{9}$$

for any $p \in [0, 1]$, any a and any Ψ-measurable w.

Theorem 7 states less obvious sufficient moment conditions of conditional auto-calibration:

$$\mathsf{E}[(r(y, \hat{F}) - \hat{r}(\hat{F}))\, w] = 0, \tag{10}$$

for any r and any Ψ-measurable w. These are also orthogonality conditions, where orthogonality is between prediction errors of $r(y, \hat{F})$ and Ψ-measurable variables.

Orthogonality conditions (8), (9) and (10) are not the narrowest sufficient conditions of conditional auto-calibration, as one can further contract the class of functions a and r, but we stop here, because further contraction could be of little practical significance.[15]

2.7 Forecast encompassing

Next we consider forecast encompassing as an example and extension of conditions of general type (2). The idea is to verify calibration of one forecasting method against another one.

[15] For example, in (8) and (9) we can use indicator variables $a = I_A$ for $A \in \sigma(\Psi, \hat{F})$.

Suppose that we want to test whether $\hat{F}_1$ is calibrated and $\hat{F}_2$ is an alternative forecast. Forecast examiner can use an information set Ψ and information contained in forecast $\hat{F}_2$ for forecast evaluation purposes. Then for two forecasts $\hat{F}_1$, $\hat{F}_2$ under the assumption that $\hat{F}_1$ is auto-calibrated with respect to $\sigma(\Psi, \hat{F}_2)$ we have for any g and any Ψ-measurable w

$$\mathsf{E} g(y, w, \hat{F}_2) = \mathsf{E} \int_{-\infty}^{\infty} g(t, w, \hat{F}_2) d\hat{F}_1(t). \tag{11}$$

This can be called a *forecast encompassing* condition. The idea of applying encompassing principle to forecasts is due to Chong and Hendry (1986). The principle states that "models which claim to congruently represent a data generation process must be able to account for the findings of rival models" (Chong and Hendry, 1986, p. 676).

We can note here that predictive distributions are particularly suited for application of the encompassing principle since they provide *complete* distribution functions, so that given one probabilistic forecast we can derive forecast of *any* calibration-related characteristics of another probabilistic forecast.

Another form of forecast encompassing contrasts results of one forecast with results of another one. The idea is that a calibrated forecast $\hat{F}_1$ must be able to explain the differential in some function g for forecasts $\hat{F}_1$ and $\hat{F}_2$. When $\hat{F}_1$ is well-calibrated we have

$$\mathsf{E}[g(y, w, \hat{F}_2) - g(y, w, \hat{F}_1)] = \mathsf{E}\left[\int_{-\infty}^{\infty} g(t, w, \hat{F}_2) d\hat{F}_1(t) - \int_{-\infty}^{\infty} g(t, w, \hat{F}_1) d\hat{F}_1(t)\right]. \tag{12}$$

The two forms of forecast encompassing conditions, (11) and (12), roughly correspond to FE(2) and FE(3) regressions in Clements and Harvey (2010) where forecast encompassing is applied to probability forecasts of 0/1 events. More generally, one can put $\hat{F}_1$ and $\hat{F}_2$ into the moment function in a non-separable way (see, for example, (18) below).

2.8 A general idea of moment-based calibration testing

As calibration tests in the existing literature mostly pertain to a situation when one-step-ahead forecasts of a time series are made given the full previous history of this series, these tests often rely on the uniformity and independence condition. Moreover, under this assumption any functions of PIT values are also independent and have known distribution; for example, this is true of the tick (indicator) variables for interval/quantile forecasts. Therefore, under the uniformity and independence the distribution of the vector of observations is fully known, which facilitates construction of the corresponding tests. For example, likelihood ratio tests are often used (e.g. Kupiec, 1995; Christoffersen, 1998; Berkowitz, 2001;

Clements and Taylor, 2003).

In general, we do not know the complete distribution of observations. The conditional distribution of a single y_i given Ψ_i and $\hat{F}_i$ is under the null of conditional auto-calibration with respect to Ψ_i fully described by the forecast $\hat{F}_i$. However, to design tests we have to make assumptions on the dependence structure in the observations $i = 1, \ldots, N$.

Given a moment condition of calibration one can replace theoretical moments by sample ones based on a series of forecasts and outcomes of the target variable and see how far the result is from what should be in theory. This allows to develop various types of diagnostic tests for forecast calibration. In Chen (2011) it is observed that many of the forecast evaluation tests developed in the literature fall within this approach.

Suppose that in theory the expectation of d must be zero under the null of calibration: $\mathsf{E}d = 0$. (Typically $d = g - \hat{g}$ as defined in (2).) We can obtain the values of d for a series of realizations of predictive distributions $\hat{F}_1, \ldots, \hat{F}_N$ and a series of outcomes $y_1, \ldots, y_N$ and calculate the corresponding sample moment $\bar{d} = \sum_{i=1}^{N} d_i / N$. If $\bar{d}$ is far from zero, then we can conclude that the forecast is miscalibrated.

Under appropriate assumptions on the distribution of the sequence of d_i, discussion of which is beyond the limits of this paper, we can use the usual t-ratios $\bar{d} / se(\bar{d})$.[16] The most subtle aspect here is adequate calculation of the standard error $se(\bar{d})$ for dependent d_i. For a multistep forecast the series of d_i values can be autocorrelated; for example, in the two plots of Figure 3 the differences between actual outcomes and forecasts series show strong persistence. In Example 2 below the usual heteroskedasticity and autocorrelation consistent standard errors are used. If this is done correctly and the series of forecasts is well-calibrated, then this statistic is asymptotically distributed as $N(0, 1)$. In order to reduce size distortion in finite samples and/or increase power we can use the fact that under the null of calibration not only conditional means, but also conditional variances are known. This knowledge can be utilized in the formula for $se(\bar{d})$.

An extension to the multivariate case—simultaneous testing of several moment conditions—is straightforward and is familiar from the GMM framework: a t-ratio is replaced by a quadratic form (J-statistic) and the distribution is chi-square. Testing of orthogonality conditions sometimes could be conveniently done by means of F-statistics and Wald statistics from auxiliary regressions (with robust covariance matrices if needed).

Note that in order to test the moment conditions of calibration it is not necessary to assume that the data are described by some parametric model and that forecasts follow that model (as in e.g. Corradi and

[16] For example, for the rolling forecasting scheme the idea of Giacomini and White (2006) can be used (see Comment 6 there).

Swanson, 2006a; Corradi and Swanson, 2006c; Chen, 2011). It is more realistic to consider forecasting *methods* rather than forecasting *models* (cf. Giacomini and White, 2006). If a forecast is obtained from a parametric model, then one can take into account the parameter uncertainty using one of the available methods (e. g, Cooley and Parke, 1990; Barndorff-Nielsen and Cox, 1996; Pascual, Romo, and Ruiz, 2001).

3 Forecast efficiency

3.1 Forecast efficiency, proper scoring rules and ideal calibration

Calibration is one important aspect of probabilistic forecasting and another is forecast efficiency. Theoretically one can call a forecast $\hat{F}$ efficient if it achieves the maximum of the expected score $\mathsf{E}S(y,\hat{F})$ according to some *scoring rule*, which is a function $S(y,F)$ of an outcome value y and a CDF F. In our analysis we can allow the scoring rule to depend on some additional variable w (a Ψ-measurable random variable for a relevant information set Ψ), which represents the environment conditions: $S = S(y,F;w)$. In applications a forecast is considered as more successful if it has a higher average score $\sum_{i=1}^{N} S(y_i,\hat{F}_i)/N$ for a sequence of predictive distributions and realized outcomes.

The general requirement is that scoring rules used for forecast evaluation must be *proper*. One can define the expected score function as the expected score of F_2 given that y is distributed as F_1:

$$\mathscr{S}(F_1,F_2) = \int_{-\infty}^{\infty} S(t,F_2)\,dF_1(t).$$

The scoring rule S is called proper if the expected score is maximized with respect to F_2 when F_2 coincides with F_1:

$$\mathscr{S}(F_1,F_1) \geq \mathscr{S}(F_1,F_2),$$

and it is *strictly proper* (within a suitable class of distributions), if the inequality is strict for $F_2 \neq F_1$. (Both F_1 and F_2 are non-random CDFs in these definitions.) A detailed review of this topic can be found in Gneiting and Raftery (2007) and Bröcker and Smith (2007). Economic applications of scoring rules can be found in Diebold and Rudebusch (1989), Corradi and Swanson (2006b), Clements and Harvey (2010), Boero, Smith, and Wallis (2011), Diks, Panchenko, and van Dijk (2011), Mitchell and Wallis (2011), Lahiri and Yang (2015).

By definition proper scoring rules encourage truthful forecast statement. Moreover, if a proper scoring rule is used, then the forecast, which is ideally calibrated with respect to Ψ, attains the highest expected score among the Ψ-measurable forecasts (Tsyplakov, 2011; also cf. Diebold, Gunther, and Tay,

1998; Granger and Pesaran, 2000). Such a forecast can be called efficient or optimal. With a strictly proper scoring rule a miscalibrated Ψ-measurable forecast must be strictly suboptimal (Holzmann and Eulert, 2014, Theorem 3).

If a forecast is not auto-calibrated given Ψ, which is signaled by a violation of some necessary moment condition, then it is not ideally calibrated given Ψ and $\hat{F}$ and there is a potential for its improvement with the help of the information contained in Ψ and $\hat{F}$. An improvement is measured by an increase in the mean score.

On the other hand, if Ψ^* is the information set containing all available information, then an efficient forecast based on Ψ^* must be ideally calibrated given Ψ^* and thus auto-calibrated given any $\Psi \subseteq \Psi^*$ as long as the scoring rule used is strictly proper. Such forecast would not be dismissed by the auto-calibration criterion. If the scoring rule used is proper, but not strictly proper, then there can be miscalibrated forecasts among efficient ones, but the forecast, which is ideally calibrated given Ψ^*, is still efficient and auto-calibrated.

Subject to these reservations, it can be said that in a certain sense the concept of calibration is intrinsically based on proper scoring rules and score maximization.

3.2 Moment conditions of forecast efficiency

One can derive moment conditions of efficiency from the first-order conditions of score maximization. Suppose that $\hat{F}$ is an efficient forecast and Ψ is the relevant information set. Consider a CDF-to-CDF transformation $T(F, w, \delta)$ depending on a real vector of parameters δ and an additional Ψ-measurable random variable w. We require that $F = T(F, w, 0)$. The transformation T produces a family of forecasts $\hat{F}_\delta = T(\hat{F}, w, \delta)$ parametrized by δ, which includes the efficient forecast $\hat{F}$ with $\delta = 0$. If $\mathsf{E}S(y, \hat{F}_\delta)$ is differentiable as a function of δ, then

$$\frac{d}{d\delta}\mathsf{E}S(y, \hat{F}_\delta)\Big|_{\delta=0} = 0.$$

Under appropriate regularity conditions the differentiation and expectation operations are interchangeable. Hence,

$$\mathsf{E}\frac{d}{d\delta}S(y, \hat{F}_\delta)\Big|_{\delta=0} = 0. \tag{13}$$

By the same logic, if S is proper, then the maximum of $\mathscr{S}(F, F_\delta)$ is achieved at $\delta = 0$, where $F_\delta =$

$T(F, w, \delta)$. Thus, for an arbitrary non-random CDF F

$$\frac{d}{d\delta}\mathscr{S}(F, F_\delta) = \int_{-\infty}^{\infty} \frac{d}{d\delta} S(t, F_\delta)\Big|_{\delta=0} dF(t) = 0.$$

It can be seen that efficiency conditions (13) can be considered as auto-calibration conditions of general type (2) with

$$g = \frac{d}{d\delta} S(y, F_\delta)\Big|_{\delta=0}.$$

The idea here is that we can extend a forecast $\hat{F}$ in a parametric way (irrespectively of a possible parametric model on which $\hat{F}$ could be based) and then derive moment conditions, which follow from the the first-order conditions of efficiency. It turns out, that these moment conditions are at the same time necessary conditions of auto-calibration.

Location A simple transformation of a CDF is a shift by $w'\delta$ where w is a real vector (which would typically include a constant element 1):

$$F_\delta(y) = F(y - w'\delta). \tag{14}$$

For example, consider a density forecast with the log-density

$$\hat{\ell}(y) = \log \hat{F}'(y)$$

and the logarithmic scoring rule

$$S(y, F) = \log F'(y).$$

In this case (necessary) moment conditions of forecast efficiency are given by

$$\mathsf{E}[-\hat{\ell}'(y)\, w] = 0,$$

where w is a Ψ-measurable vector.

Scale Another simple transformation is scaling of CDF F around some central point $c(F)$. Natural central points are the median $c = F^{-1}(1/2)$ and the mean $c = \mathrm{mean}(F)$:

$$F_\delta(y) = F((y - c(F)) \exp(-w'\delta) + c(F)). \tag{15}$$

For the logarithmic scoring rule the corresponding conditions of forecast efficiency are given by

$$\mathsf{E}[(-\hat{\ell}'(y)(y - c(\hat{F})) - 1)\, w] = 0.$$

Inverse normal transform: location and scale Alternatively, we can employ transformations based on the inverse normal transform (INT) of CDF F defined as $\Phi^{-1} \circ F$, where $\Phi(\cdot)$ is the standard normal CDF:

$$F_\delta(y) = \Phi(\Phi^{-1}(F(y)) - w'\delta)$$

and

$$F_\delta(y) = \Phi(\Phi^{-1}(F(y)) \exp(-w'\delta)).$$

These transformations correspond to the location and scale and suggest the following conditions of forecast efficiency with the logarithmic scoring rule:

$$\mathsf{E}[\mathrm{INT}\, w] = 0$$

and

$$\mathsf{E}[(\mathrm{INT}^2 - 1)\, w] = 0,$$

where $\mathrm{INT} = \Phi^{-1}(\hat{F}(y))$. It can be seen that the two conditions are orthogonality conditions for probabilistic calibration of type (6). This demonstrates that some known calibration tests based on PIT and INT values (e.g. Berkowitz, 2001) can be motivated by their connection with moment conditions of forecast efficiency.

Note, that a forecast calibration test is similar in structure to an ordinary model diagnostic test. That is, upon rejection of the null of calibration we do not necessary have a well-defined alternative forecasting method to be applied instead of the rejected one. However, forecast efficiency tests based on parametric extensions of the evaluated forecast introduced here provide us with a direction of a reasonable forecast modification (recalibration). If an efficiency test diagnoses miscalibration, we could estimate the corresponding parameters δ by maximizing the average score

$$\max_\delta \frac{1}{N} \sum_{i=1}^{N} S(y_i, \hat{F}_{\delta,i})$$

and replace the rejected forecast $\hat{F}$ by a recalibrated one $\hat{F}_{\delta^*}$, where δ^* is a value giving the maximum.

For example, if the location test with $w = 1$ rejects the null of calibration, then the conclusion is

that the predictive distributions are systematically located lower or higher than needed. Then one can estimate the constant δ and shift predictive distributions accordingly. If a scale test with $w = 1$ rejects calibration, then the predictive distributions are too narrow or too wide and we can scale them by $\exp(\delta)$. In Example 3 below theoretical recalibration using model (19) for INT values is mentioned. In practice one can estimate the model using available observations and recalibrate the forecast as suggested there.

An alternative way to test forecast efficiency is to use the average score differential $\frac{1}{N}\sum_{i=1}^{N}(S(y_i, \hat{F}_{\delta^*,i}) - S(y_i, \hat{F}_{0,i}))$ as the basis for a test statistic. In Berkowitz (2001) and Bao, Lee, and Saltoğlu (2007) (see also Mitchell and Hall, 2005) this was suggested for the logarithmic score. The difference with the current approach is similar to the difference of a likelihood ratio test with a score (Lagrange multiplier) test in the context of maximum likelihood inference.

3.3 Auto-calibration, efficiency and sharpness

Another link between calibration and efficiency is provided by "the sharpness principle" of probabilistic forecasting conjectured in Gneiting, Balabdaoui, and Raftery (2007). It states that the problem of finding a good forecast can be viewed as the problem of maximizing sharpness subject to calibration. (Forecast sharpness is a characteristic which reflects the degree of forecast definiteness, concentration of the forecast distribution; Murphy and Winkler, 1987; Gneiting, Balabdaoui, and Raftery, 2007). It can be shown that the conjecture is true provided that a vague "calibration" notion is replaced by (conditional or unconditional) auto-calibration.[17]

First, for a proper scoring rule $\mathscr{S}(F, F)$ can be viewed as a measure of sharpness of a distribution F. For a proper scoring rule $-\mathscr{S}(F, F)$ is a concave function of F and thus, according to DeGroot (1962), can be viewed as a measure of uncertainty of a probability distribution with CDF F; see also Bröcker (2009). Second, for a forecast $\hat{F}$ which is auto-calibrated given Ψ and a Ψ-measurable variable w we have

$$\mathsf{E}S(y, \hat{F}; w) = \mathsf{E}\mathscr{S}(\hat{F}, \hat{F}; w), \tag{16}$$

i. e. the expected score of an auto-calibrated forecast equals its expected sharpness. The fact follows from (2) for $g = S(y, F; w)$.

Users may prefer sharp forecast as they are more definite and informative. However, forecast sharpness can be deceptive and it is not a good idea to make a choice between forecasts solely on the basis of their sharpness. We see from (16) that sharpness is no more a deceptive characteristic when only

[17] See Bröcker (2009) for an alternative interpretation of this principle.

auto-calibrated forecasts are considered. Such forecasts can be compared on the basis of the levels of their expected sharpness. The ideally calibrated forecast given Ψ^* is the sharpest of all Ψ^*-measurable forecasts, which are auto-calibrated given $\Psi \subseteq \Psi^*$, because it is characterized by the greatest expected score.

3.4 Predicted efficiency conditions

Finally in this section we consider calibration conditions, which relate to forecast efficiency indirectly, through the use of proper scoring rules.

As was already noted, the expected sharpness of an auto-calibrated forecast equals its expected score. In general if $\hat{F}$ is auto-calibrated given Ψ, then from (5) with $r = S(y, F)$ we have that for any function $a(w, F)$ and any Ψ-measurable w

$$\mathsf{E}[(S(y, \hat{F}) - \mathscr{S}(\hat{F}, \hat{F}))\, a(w, \hat{F})] = 0. \tag{17}$$

An interesting extension of this idea is to use forecast encompassing conditions based on predicted efficiency. If $\hat{F}_1$ is auto-calibrated given $\sigma(\Psi, \hat{F}_2)$, then $\hat{F}_1$ should be able to give conditionally unbiased prediction of the score of $\hat{F}_2$. That is, we have for any function b and any Ψ-measurable w that

$$\mathsf{E}[(S(y, \hat{F}_2) - \mathscr{S}(\hat{F}_1, \hat{F}_2))\, b(w, \hat{F}_1, \hat{F}_2)] = 0. \tag{18}$$

One can also implement encompassing on the bases of the score differential between $\hat{F}_1$ and $\hat{F}_2$. Unconditional encompassing for score differential parallels the generalization of the likelihood ratio test for non-nested models developed in Cox (1961), Cox (1962).

4 Examples

4.1 Example 1, analytical illustration

Let

$$y = z + \epsilon = x + \xi + \epsilon$$

be the target variable for forecasting, where ξ, ϵ, x are jointly independent standard normal variables and

$$z = x + \xi.$$

Table 1: Moment conditions (3) check for Example 1 with $c = \Phi^{-1}(p)\,w$

	$\mathsf{E}[\Phi^{-1}(\hat{F}_x(y))\,w]$	$\mathsf{E}[\Phi^{-1}(\hat{F}_z(y))\,w]$
$w = x$	$\mathsf{E}\left[\frac{\xi+\epsilon}{\sqrt{2}}x\right] = 0$	$\mathsf{E}[\epsilon x] = 0$
$w = z$	$\mathsf{E}\left[\frac{\xi+\epsilon}{\sqrt{2}}(x+\xi)\right] = \frac{1}{\sqrt{2}}$	$\mathsf{E}[\epsilon(x+\xi)] = 0$

Consider two forecasts of y based on $y|x \sim N(x,2)$ and $y|z \sim N(z,1)$. Forecast $\hat{F}_x$ is ideally calibrated given x:

$$\hat{F}_x(q) = \mathbb{F}(q|x) = \Phi\left(\frac{q-x}{\sqrt{2}}\right)$$

and forecast $\hat{F}_z$ is ideally calibrated given z:

$$\hat{F}_z(q) = \mathbb{F}(q|z) = \Phi(q-z).$$

Here x does not provide additional information for predicting y once z is known, because $y|x,z \sim N(z,1)$. We use this setting to illustrate some key concepts and results of this paper analytically. One can start by observing that neither $\hat{F}_z$ is ideally calibrated given x, nor $\hat{F}_x$ is ideally calibrated given z.

In the following we use $\text{mean}(\hat{F}_x) = x$, $\text{var}(\hat{F}_x) = 2$, $\text{mean}(\hat{F}_z) = z$, $\text{var}(\hat{F}_z) = 1$, where

$$\text{mean}(F) = \int_{-\infty}^{\infty} t\,dF(t) \qquad \text{and} \qquad \text{var}(F) = \int_{-\infty}^{\infty} t^2\,dF(t) - \text{mean}(F)^2.$$

Auto-calibration Since $\sigma(x,\hat{F}_z) = \sigma(x,z)$ and thus $y|(x,\hat{F}_z) \sim N(z,1)$, it follows that $\hat{F}_z$ is auto-calibrated given x. Further, $\sigma(z,\hat{F}_x) = \sigma(x,z)$, $y|(z,\hat{F}_x) \sim N(z,1)$ and thus $\hat{F}_x$ is not auto-calibrated given z.

Conditional probabilistic calibration The PIT values of the two forecasts are $\hat{F}_x(y) = \Phi\left(\frac{y-x}{\sqrt{2}}\right) = \Phi\left(\frac{\xi+\epsilon}{\sqrt{2}}\right)$ and $\hat{F}_z(y) = \Phi(y-z) = \Phi(\epsilon)$. Both are unconditionally uniformly distributed at $[0,1]$. Since ϵ and $z = x+\xi$ are independent, we have $\hat{F}_z(y)|z \sim U[0,1]$. Similarly, $\hat{F}_z(y)|x \sim U[0,1]$. Independence of $\xi+\epsilon$ and x implies $\hat{F}_x(y)|x \sim U[0,1]$. Thus, according to Definition 3 $\hat{F}_x$ is probabilistically calibrated with respect to both x and z, and $\hat{F}_z$ is probabilistically calibrated with respect to z. Finally, $\xi+\epsilon$ and $z = x+\xi$ are correlated and $\hat{F}_x$ is not probabilistically calibrated with respect to z.

These conclusions are corroborated by checking the moment conditions (3) for $c(p,w) = \Phi^{-1}(p)\,w$. For this function the prediction is zero: $\int_0^1 c(p,w)\,dp = w\int_0^1 \Phi^{-1}(p)\,dp = 0$. Note that $\Phi^{-1}(\hat{F}_x(y)) = \frac{\xi+\epsilon}{\sqrt{2}}$ and $\Phi^{-1}(\hat{F}_z(y)) = \epsilon$. Table 1 shows that $\hat{F}_x$ is not probabilistically calibrated with respect to z.

This is an example of orthogonality conditions for conditional probabilistic calibration (6) with

Table 2: Conditional marginal calibration (Definition 4) check for Example 1

	$\mathbb{F}(q\|w)$	$\mathsf{E}(\hat{F}_x(q)\|w) = \mathsf{E}\left(\Phi\left(\frac{q-x}{\sqrt{2}}\right)\middle\|w\right)$	$\mathsf{E}(\hat{F}_z(q)\|w) = \mathsf{E}(\Phi(q-z)\|w)$
$w=x$	$\Phi\left(\frac{q-x}{\sqrt{2}}\right)$	$\Phi\left(\frac{q-x}{\sqrt{2}}\right)$	$\Phi\left(\frac{q-x}{\sqrt{2}}\right)$
$w=z$	$\Phi(q-z)$	$\Phi\left(\frac{q-z/2}{\sqrt{5/2}}\right)$	$\Phi(q-z)$

Table 3: Moment conditions (4) check for Example 1 with $n = e^{y-w}$

	$\mathsf{E}e^{y-w}$	$\mathsf{E}\hat{n}(\hat{F}_x, w) = \mathsf{E}e^{x+1-w}$	$\mathsf{E}\hat{n}(\hat{F}_z, w) = \mathsf{E}e^{z+1/2-w}$
$w=x$	$\mathsf{E}e^{\xi+\epsilon} = e$	e	$\mathsf{E}e^{\xi+1/2} = e$
$w=z$	$\mathsf{E}e^{\epsilon} = e^{1/2}$	$\mathsf{E}e^{-\xi+1} = e^{3/2}$	$e^{1/2}$

$k(p) = \Phi^{-1}(p)$ and $w = x$ (or $w = z$).

Conditional marginal calibration Table 2 allows to check conditional marginal calibration. The formulas for $\mathsf{E}(\hat{F}_x(q)|z)$ and $\mathsf{E}(\hat{F}_z(q)|x)$ can be derived using the following identities for the standard normal CDF:

$$\Phi(-t) = 1 - \Phi(t)$$

and

$$\int_{-\infty}^{\infty} \Phi((t-m_1)/s_1)\, d\Phi((t-m_2)/s_2) = \Phi\left((m_2-m_1)/\sqrt{s_1^2+s_2^2}\right), \qquad s_1 > 0, s_2 > 0.$$

It can be seen that according to Definition 4 $\hat{F}_x$ is not marginally calibrated given z, since $\mathsf{E}(\hat{F}_x(q)|z)$ does not coincide with $\mathbb{F}(q|z)$. For other cases the conditional marginal calibration is confirmed.

These conclusions can be corroborated by checking the moment conditions (4) for $n(y,w) = e^{y-w}$. If F is the CDF of $N(\mu,\sigma^2)$, then by the properties of log-normal distribution[18] $\hat{n}(F,w) = e^{\mu+\sigma^2/2-w}$. Using the same formula for log-normal distribution one can obtain the expectations of n and $\hat{n}$ (Table 3). It is seen from the table that $\hat{F}_x$ is not marginally calibrated with respect to z.

This is an example of orthogonality conditions for conditional marginal calibration (7) with $m(y) = e^y$ and $w = e^{-x}$ (or $w = e^{-z}$).

Orthogonality conditions of auto-calibration and predicted efficiency Let $r = (y - \mathrm{mean}(\hat{F}))^2$ and $a = w^2$ in orthogonality conditions (5). Note that $\hat{r} = \mathrm{var}(\hat{F})$, $y - \mathrm{mean}(\hat{F}_x) = y - x = \xi + \epsilon$ and $y - \mathrm{mean}(\hat{F}_z) = y - z = \epsilon$. Table 4 demonstrates that the orthogonality condition is violated only for the forecast $\hat{F}_x$ and $w = z$, which confirms that $\hat{F}_x$ is not auto-calibrated given z.

This orthogonality condition is a predicted efficiency condition (17) for a (non-strictly) proper scoring

[18] If $\nu \sim N(\mu,\sigma^2)$ (e^ν is log-normal), then $\mathsf{E}e^\nu = e^{\mu+\sigma^2/2}$.

Table 4: Orthogonality conditions (5) check for Example 1 with $r = (y - \text{mean}(\hat{F}))^2$, $a = w^2$

	$\mathsf{E}[((y - \text{mean}(\hat{F}_x))^2 - \text{var}(\hat{F}_x))\, w^2]$	$\mathsf{E}[(y - \text{mean}(\hat{F}_z))^2 w^2]$
$w = x$	$\mathsf{E}[((\xi + \epsilon)^2 - 2)x^2] = 0$	$\mathsf{E}[(\epsilon^2 - 1)x^2] = 0$
$w = z$	$\mathsf{E}[((\xi + \epsilon)^2 - 2)z^2] = 2$	$\mathsf{E}[(\epsilon^2 - 1)z^2] = 0$

Table 5: Forecast encompassing conditions (11) check for Example 1 with $g = (y - \text{mean}(\hat{F}))^2$

	$\mathsf{E}g_2(y)$	$\mathsf{E}\left[\int_{-\infty}^{\infty} g_2(t)\, d\hat{F}_1(t)\right]$
$\hat{F}_1 = \hat{F}_z$, $\hat{F}_2 = \hat{F}_x$	$\mathsf{E}[(y - x)^2] = \mathsf{E}[(\xi + \epsilon)^2] = 2$	$\mathsf{E}[1 + (z - x)^2] = \mathsf{E}[1 + \xi^2] = 2$
$\hat{F}_1 = \hat{F}_x$, $\hat{F}_2 = \hat{F}_z$	$\mathsf{E}[(y - z)^2] = \mathsf{E}[\epsilon^2] = 1$	$\mathsf{E}[2 + (x - z)^2] = \mathsf{E}[2 + \xi^2] = 3$

rule $S(y, F) = -(y - \text{mean}(F))^2$.

Forecast encompassing Consider forecast encompassing conditions (11) for $g = (y - \text{mean}(\hat{F}))^2$. Denote $g_2(q) = (q - \text{mean}(\hat{F}_2))^2$ and observe that

$$\int_{-\infty}^{\infty} g_2(t)\, d\hat{F}_1(t) = \text{var}(\hat{F}_1) + (\text{mean}(\hat{F}_1) - \text{mean}(\hat{F}_2))^2.$$

Since $\hat{F}_z$ is auto-calibrated given $\sigma(\hat{F}_x) = \sigma(x)$, we have that (11) is fulfilled for $\hat{F}_1 = \hat{F}_z$, $\hat{F}_2 = \hat{F}_x$ (Table 5). For $\hat{F}_1 = \hat{F}_x$, $\hat{F}_2 = \hat{F}_z$ the encompassing moment condition (11) is violated. That is, $\hat{F}_x$ is not able to explain the behavior of $\hat{F}_z$, and we observe that $\hat{F}_x$ is not auto-calibrated given $\sigma(\hat{F}_z) = \sigma(z)$.

These forecast encompassing conditions are of the predicted efficiency form (18) if one takes $S(y, F) = -(y - \text{mean}(F))^2$ and $b = 1$.

4.2 Example 2, evaluation of the Swedish Riksbank's inflation forecasts

The next example illustrates an applied use of location and scale tests based on forecast efficiency conditions introduced in subsection 3.2. We want to evaluate the Swedish Riksbank's forecasts of CPI inflation, which appeared above as illustrations of density forecasts.

Sweden's central bank (Riksbank) started to publish its density forecasts of inflation in June 1998. The forecasts are in the form of two-piece normal distribution. The target variable is the yearly CPI inflation. We evaluate only one-year-ahead forecasts. There are 64 forecasts available for evaluation for the period 1999–2015. They were issued 4 times a year with approximately quarterly frequency. Note that the forecasts before 2007 are conditional, assuming a predetermined trajectory of the repo rate. Nevertheless, in this forecast evaluation example we take them "as is", thus representing a point of view of a user, who considers the possibility of employing the forecasts in the unmodified form. Additional details about the Riksbank's forecasts can be found in Appendix C.

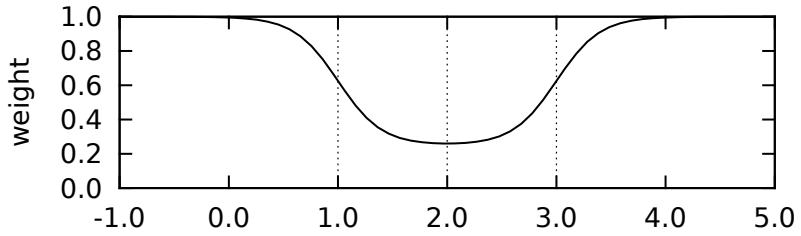

Figure 4: The weighting function $\gamma(y)$ for evaluation of the Riksbank's inflation forecasts.

The Riksbank's forecasts are aimed to provide an informational support to the bank's inflation targeting policy. The target was fixed at 2% yearly inflation, initially with ± 1 percentage point tolerance interval. Subsequently the tolerance interval was abandoned as unrealistic. However, the interval from 1% to 3% of yearly inflation can still be considered as a good reference for evaluation of the monetary policy. That is, inflation outside this interval is an important warning to the monetary authorities.

To take into account this background, one can build efficiency test on a specific proper scoring rule, which puts more weight to the areas outside the $2 \pm 1\%$ band. Two varieties of suitable scoring rules can be found in Diks, Panchenko, and van Dijk (2011). They modify the ordinary logarithmic score by a weighting function $\gamma(y)$ ($0 \le \gamma(y) \le 1$), which emphasizes certain areas of the target variable range. We choose one of the rules of Diks, Panchenko, and van Dijk (2011), the *generalized censored likelihood* (GCsL) scoring rule, defined as

$$S(y,F) = \gamma(y)\log F'(y) + (1-\gamma(y))\log\left(\int_{-\infty}^{\infty}(1-\gamma(t))\,dF(t)\right).$$

The formulas for the location and scale moment functions corresponding to the GCsL score are given is Appendix B. The weighting function $\gamma(y)$ used is constructed in such a way, that the "dangerous" regions outside the $2 \pm 1\%$ band have higher weight (figure 4):

$$\gamma(y) = 0.25 + \frac{0.75}{1+\exp(5(y-1))} + \frac{0.75}{1+\exp(5(3-y))}.$$

The test statistics below are Wald statistics from the regressions corresponding to the orthogonality conditions. Newey–West covariance matrices with 4 lags are used throughout.

Location test 1 Our first test is an unconditional location test based on transformation (14) with $w = 1$. It gives $\chi^2_1 = 1.55$ with a p-value of 21%.

Location test 2 The test is a conditional location test. The conditioning variables are the mean of the Riksbank's forecast and a point forecast of Swedish inflation $\hat{y}_{\text{NIER}}$ produced by the National Institute of Economic Research (NIER; the details are in Appendix C). Since the available history of NIER's forecasts starts from 2001, there are only 53 observations. The test is again based on transformation (14) with $w = (1, \text{mean}(\hat{F}), \hat{y}_{\text{NIER}})$. It gives $\chi^2_3 = 5.08$ with a p-value of 17%.

Scale test 1 A scale test can be based on transformation (15) with $c = \mu$ (the mode of $\hat{F}$). The unconditional test (with $w = 1$) gives $\chi^2_1 = 1.78$ with a p-value of 18%.

Scale test 2 Finally, we run a conditional scale test using the scale of the predictive distribution itself as the conditioning variable. The test is again based on transformation (15) with $c = \mu$ and $w = (1, \log(\hat{\sigma}^2))$, where $\hat{\sigma}^2$ is the variance of the predictive distribution $\hat{F}$. It gives $\chi^2_2 = 1.89$ with a p-value of 39%.

We are not able to find any signs of miscalibration in the Riksbank's forecasts. Both unconditional and conditional location and scale tests do not reveal miscalibration at the 15% significance level. Alternative NIER's forecasts does not seem to provide information, which can lead to significant improvement. The conclusion is limited by using the one year horizon and a peculiar weighting function related to inflation targeting.

Note that the tests used here refer to orthogonality conditions of general type (10) and are not reducible to conditional marginal or probabilistic calibration.

4.3 Example 3, power comparison of calibration tests by simulation

Our third example concerns the power of calibration tests. Suppose that y is given by

$$y = x + \epsilon/\sqrt{z},$$

where $x \sim N(0,1)$, $z \sim \chi^2_8/8$ (scaled chi-squared distribution) and $\epsilon \sim N(0,1)$ are independent. Also denote $\hat{F}_x$, $\hat{F}_z$, $\hat{F}_{xz}$ conditional CDFs which correspond to $y|x \sim x + t_8$ (shifted Student's distribution), $y|z \sim N(0, 1 + 1/z)$ and $y|x, z \sim N(x, 1/z)$.

We run simulations for four forecasts which are combinations of two partial conditional CDFs, $\hat{F}_x$ and $\hat{F}_z$. The first is the equal-weight linear pool[19]: $\hat{F}_c = \frac{1}{2}\hat{F}_x + \frac{1}{2}\hat{F}_z$. The second and third forecasts are recalibrated versions of $\hat{F}_c$. The recalibration (improving in order to achieve better calibration) is

[19]This is a popular method of combining predictive distributions; e. g. Geweke and Amisano (2011), Gneiting and Ranjan (2013).

Table 6: Statistics for the forecasts of Example 3

	$\hat{F}_c$	$\hat{F}_{r1}$	$\hat{F}_{r2}$	$\hat{F}_{xz}$
Expected log. score	−1.612	−1.596	−1.525	−1.485
% best	0	0	1.41	98.59
Test 1, $\mathrm{INT}^2 - 1 \perp 1$	71.35	5.07	4.59	5.06
Test 2, $\mathrm{INT} \perp 1, x$	99.68	99.92	4.30	5.11
Test 3, $y - \hat{y} \perp 1, x$	99.71	99.88	4.99	5.22
Test 4, $S - \hat{S} \perp 1, \hat{S}$	94.03	55.17	54.47	4.66
Test 5, $S_x - \hat{S}_x \perp 1, \hat{S}, \hat{S}_x$	100.0	98.06	55.23	4.94

Note: Logarithmic scoring rule is used throughout. The table is based on 10000 simulations. The expected logarithmic score $\mathsf{E}S(y, \hat{F})$ is in the first row. The test statistics are quadratic forms for the moment conditions. The figures for the tests are rejection frequencies at 5% asymptotic significance level using the corresponding chi-squared quantiles. The number of observations in the tests is 200.

implemented via an INT-based model:

$$\mathrm{INT}_c = \beta x + \xi, \qquad \mathsf{Var}\xi = \sigma, \tag{19}$$

where $\mathrm{INT}_c = \Phi^{-1}(\hat{F}_c(y))$ is the inverse normal transform corresponding to $\hat{F}_c$. The recalibrated forecast is given by $\hat{F}_r(q) = \Phi((\Phi^{-1}(\hat{F}_c(q)) - \beta x)/\sigma)$. Forecast $\hat{F}_{r1}$ with $\beta = 0$, $\sigma = 0.874$ repairs only the incorrect unconditional dispersiveness of $\hat{F}_c$. Forecast $\hat{F}_{r2}$ with $\beta = 0.316$, $\sigma = 0.814$ also repairs the conditional location. The parameters are approximations to the corresponding theoretical models. Finally, forecast $\hat{F}_{xz}$ is known to be ideally calibrated with respect to $\sigma(x, z)$ and conditionally auto-calibrated with respect to $\sigma(x)$. It can be regarded as a perfectly recalibrated variant of $\hat{F}_c$ since $\sigma(\hat{F}_c) = \sigma(x, z)$).

We reproduce a situation where a forecast examiner can observe x, but not z or ϵ. Some forecaster(s) submitted him forecasts $\hat{F}_c$, $\hat{F}_{r1}$ and $\hat{F}_{r2}$. (We are adding $\hat{F}_{xz}$ for control purposes.) From examiner's point of view the suitable mode of calibration is conditional auto-calibration with respect to $\Psi = \sigma(x)$. Five different tests are used, which are based on the following moment conditions.

Test 1 $\mathsf{E}[\mathrm{INT}^2 - 1] = 0$.

Test 2 $\mathsf{E}\mathrm{INT} = 0$ and $\mathsf{E}[\mathrm{INT}x] = 0$, where $\mathrm{INT} = \Phi^{-1}(\hat{F}(y))$.

Test 3 $\mathsf{E}[y - \hat{y}] = 0$ and $\mathsf{E}[(y - \hat{y})x] = 0$, where $\hat{y} = \mathrm{mean}(\hat{F})$.

Test 4 $\mathsf{E}[S - \hat{S}] = 0$ and $\mathsf{E}[(S - \hat{S})\hat{S}] = 0$, where $S = S(y, \hat{F})$ and $\hat{S} = \mathscr{S}(\hat{F}, \hat{F})$ for the logarithmic scoring rule $S(y, F) = \log F'(y)$.

Test 5 $\mathsf{E}[S_x - \hat{S}_x] = 0$, $\mathsf{E}[(S_x - \hat{S}_x)\hat{S}] = 0$ and $\mathsf{E}[(S_x - \hat{S}_x)\hat{S}_x] = 0$, where $S_x = S(y, \hat{F}_x)$ and $\hat{S}_x = \mathscr{S}(\hat{F}, \hat{F}_x)$.

Test 1 is an unconditional scale test based on INT values. Test 2 is a conditional location test based on INT values. Test 3 is a conditional location test based directly on the target variable. Test 4 is a predicted efficiency test and relates to the condition that the expected score is equal to the expected sharpness and can be interpreted as a test of correct sharpness. Finally Test 5 is a forecast encompassing test against $\hat{F}_x$ based on predicted efficiency (conditions (18) with $b = 1$, $b = \hat{S}$, $b = \hat{S}_x$). Tests 1 and 2 relate to probabilistic calibration given x, Test 3 relates to marginal calibration given x, while Tests 4 and 5 are more general tests of unconditional and conditional (given x) auto-calibration respectively.

Test statistics are quadratic forms based on plain sample moments with weighting matrices calculated from the corresponding predicted variance-covariance matrices suggested by the forecasts. The statistics are distributed asymptotically as chi-squared under auto-calibration.

We used simulations with 200 observations. The required moments in intractable cases were calculated by numerical integration.[20] Table 6 shows the results on the expected logarithmic score comparison of average logarithmic scores and rejection rates for the five calibration tests.

The expected logarithmic score shows the asymptotic potential of a forecast, which becomes visible when the number of observations tends to infinity. When a series of forecasts is not very long, imperfect forecasts can obtain higher average scores than the ideal forecast. So "% best" row shows the percentage of experiments in which the corresponding forecast had the highest average logarithmic score.

All of the forecasts $\hat{F}_c$, $\hat{F}_{r1}$ and $\hat{F}_{r2}$ are both probabilistically and marginally miscalibrated. However, with 200 observations the power of some tests is low.

The basic combined forecast $\hat{F}_c$ is overdispersed and $\hat{F}_{r1}$ corrects this well-known problem (cf. Gneiting and Ranjan, 2013 where the beta CDF is used for the same purpose). Test 1, indeed, frequently signals inadequate unconditional dispersiveness in $\hat{F}_c$, while it shows rejection rate close to 5% for the three recalibrated forecasts.

Recalibration of $\hat{F}_c$ for both location and scale produces forecast $\hat{F}_{r2}$, which does not show obvious signs of either probabilistic or marginal conditional miscalibration with 200 observations. However, since it does not coincide with the ideally recalibrated forecast $\hat{F}_{xz}$, it cannot be auto-calibrated. Indeed, Test 4 and 5 often signal miscalibration.

Here partial miscalibration criteria (like tests for conditional probabilistic calibration) are not able to signal miscalibration in $\hat{F}_{r2}$ given x. The corresponding tests have low power. Even though z is not observable to the examiner directly, he can use the characteristics of the forecast itself to detect mis-

[20] That is, $\int_{-\infty}^{\infty} g(t)dF(t) \approx \int_{y_{\min}}^{y_{\max}} g(t)dF(t) \approx \sum_{i=1}^{M} g(y_i - h/2)(F(y_i) - F(y_i - h))$, where $h = (y_{\max} - y_{\min})/M$, $y_i = y_{\min} + ih$. We used $y_{\min} = -15$, $y_{\max} = 15$, $M = 150$ throughout.

calibration. It can be also seen from this example that an encompassing predicted efficiency test can be useful in situations where the direction of miscalibration is not obvious. If there exists some baseline forecast, then we can use it for miscalibration diagnostics without additional analysis. Tests 4 and 5 use general moment conditions (2) of forecast calibration, and they cannot be reduced to testing conditional probabilistic or marginal calibration given x. These tests have nontrivial power even in the difficult case of forecast $\hat{F}_{r2}$. Thus, the concept of conditional auto-calibration can be important in practice, not only in theory. Artificially narrowing the family of testable conditions can lead to weak tests unable to reject at a given sample size and significance level.

5 Conclusion

In evaluation of probabilistic forecasts it is desirable to rely on fundamental concepts and theoretical properties. Some of such concepts and properties were considered in this paper. In particular, before testing forecast efficiency and calibration it is wise to understand what exactly is tested. Conditional auto-calibration given an information set is an important fundamental concept which can be used. The paper highlights the difference between conditional auto-calibration and the less general concepts of conditional probabilistic calibration (PIT uniformity) and conditional marginal calibration.

The notion of conditional auto-calibration is closely connected to the notion of forecast efficiency, which in this paper is defined via expected score maximization based on a proper scoring rule. While score maximization can be viewed as the implicit basis of forecast evaluation, conditional auto-calibration is an empirically relevant substitute condition.

An interesting aspect of connections between calibration and efficiency, which can provide helpful intuition to a forecast examiner, is the principle of maximizing sharpness subject to calibration. The concept of auto-calibration helps to derive this principle from expected score maximization.

Forecast efficiency, conditional marginal and probabilistic calibration, conditional auto-calibration all can be expressed by various moment conditions, including orthogonality conditions. These conditions lead to a general framework for evaluation of probabilistic forecasts. The framework can facilitate construction of various new tests. This is exemplified by general forecast encompassing tests introduced in this paper, including tests based on predicted efficiency condition.

One can never be sure that a forecast is conditionally auto-calibrated (probabilistically calibrated, marginally calibrated). All of the theoretical results on sufficient moment conditions of calibration require the corresponding identities to be satisfied for arbitrary functions and arbitrary conditioning vari-

ables. This observation is closely related to the problem of choosing variables and functions for calibration testing. There is no guarantee that conditions (8), (9) or (10) can produce tests with good power. One can only state that an examiner utilizing such narrow conditions would not be fundamentally non-comprehensive. Perhaps, some more general tests based on conditions (2) can be more powerful. However, the choice of test functions can be non-obvious.

Our view is that the problem is a fundamental one and there is no universal solution. Forecast evaluation is an art in the same sense that forecasting itself is an art. However, one can suggest a possible broad strategy for a forecast examiner: try to build as good forecast as you yourself can or find some other good forecast and use this baseline forecast to test forecast encompassing. Encompassing can relate to a specific aspect like location or scale or be more general, e. g. based on some predicted efficiency conditions. It is reasonable to start testing miscalibration in several obvious directions, but to discover non-obvious miscalibration one has to be creative.

Tests for location and scale are constructive and suggest a direction of improvement. However, there is a problem with less specific calibration tests, like tests based on predicted efficiency conditions. After rejection of the null of calibration we do not always know how to improve our forecasting method. This is not something peculiar to forecast evaluation. Ordinary goodness-of-fit and diagnostic tests for statistical models have the same problem. For example, if a Pearson's chi-square test rejects the fitted distribution, then in general we do not know how to change the distribution. Similar to goodness-of-fit and diagnostic tests, non-specific calibration tests are useful as they can signal non-obvious miscalibration and can stimulate search in less obvious directions. Even if we do not know immediately what to do with such knowledge of miscalibration, this is not a good reason to be blind about it.

References

Bao, Y., T.-H. Lee, and B. Y. Saltoğlu (2007): "Comparing Density Forecast Models," *Journal of Forecasting*, 26, 203–225.

Barndorff-Nielsen, O. E., and D. R. Cox (1996): "Prediction and Asymptotics," *Bernoulli*, 2(4), 319–340.

Berkowitz, J. (2001): "Testing Density Forecasts, With Applications to Risk Management," *Journal of Business & Economic Statistics*, 19(4), 465–474.

Bierens, H. J. (2004): *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge University Press.

Boero, G., J. Smith, and K. F. Wallis (2011): "Scoring Rules and Survey Density Forecasts," *International Journal of Forecasting*, 27(2), 379–393.

Britton, E., P. Fisher, and J. Whitley (1998): "The Inflation Report Projections: Understanding the Fan Chart," *Bank of England Quarterly Bulletin*, 38, 30–37.

Bröcker, J. (2009): "Reliability, Sufficiency, and the Decomposition of Proper Scores," *Quarterly Journal of the Royal Meteorological Society*, 135(643), 1512–1519.

Bröcker, J., and L. A. Smith (2007): "Scoring Probabilistic Forecasts: The Importance of Being Proper," *Weather and Forecasting*, 22, 382–388.

Brockwell, A. E. (2007): "Universal Residuals: A Multivariate Transformation," *Statistics & Probability Letters*, 77, 1473–1478.

Chen, Y.-T. (2011): "Moment Tests for Density Forecast Evaluation in the Presence of Parameter Estimation Uncertainty," *Journal of Forecasting*, 30(4), 409–450.

Chong, Y. Y., and D. F. Hendry (1986): "Econometric Evaluation of Linear Macro-Economic Models," *Review of Economic Studies*, 53(4), 671–690.

Christoffersen, P. F. (1998): "Evaluating Interval Forecasts," *International Economic Review*, 39(4), 841–862.

Clements, M. P. (2004): "Evaluating the Bank of England Density Forecasts of Inflation," *The Economic Journal*, 114, 844–866.

——— (2006): "Evaluating the Survey of Professional Forecasters Probability Distributions of Expected Inflation Based on Derived Event Probability Forecasts," *Empirical Economics*, 31(1), 49–64.

Clements, M. P., and D. I. Harvey (2010): "Forecast Encompassing Tests and Probability Forecasts," *Journal of Applied Econometrics*, 25(6), 1028–1062.

Clements, M. P., and N. Taylor (2003): "Evaluating Interval Forecasts of High-Frequency Financial Data," *Journal of Applied Econometrics*, 18(4), 445–456.

Cooley, T. F., and W. R. Parke (1990): "Asymptotic Likelihood-Based Prediction Functions," *Econometrica*, 58(5), 1215–1234.

Corradi, V., and N. R. Swanson (2006a): "Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification," *Journal of Econometrics*, 133(2), 779–806.

——— (2006b): "Predictive Density and Conditional Confidence Interval Accuracy Tests," *Journal of Econometrics*, 135(1-2), 187–228.

——— (2006c): "Predictive Density Evaluation," in *Handbook of Economic Forecasting*, ed. by C. W. J. Granger, G. Elliott, and A. Timmermann, vol. 1, chap. 5, pp. 197–286. North-Holland, Amsterdam.

Cox, D. R. (1961): "Tests of Separate Families of Hypotheses," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 105–123, Berkeley. University of California Press.

Cox, D. R. (1962): "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2), 406–424.

Dawid, A. P. (1984): "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society. Series A (General)*, 147(2), 278–292.

DeGroot, M. H. (1962): "Uncertainty, Information, and Sequential Experiments," *The Annals of Mathematical Statistics*, 33(2), 404–419.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39(4), 863–883.

Diebold, F. X., J. Hahn, and A. S. Tay (1999): "Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange," *Review of Economics and Statistics*, 81(4), 661–673.

Diebold, F. X., and G. D. Rudebusch (1989): "Scoring the Leading Indicators," *The Journal of Business*, 62(3), 369–391.

Diebold, F. X., A. S. Tay, and K. F. Wallis (1999): "Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters," in *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*, ed. by R. F. Engle, and H. White, pp. 76–90. Oxford University Press, Oxford.

Diks, C., V. Panchenko, and D. van Dijk (2011): "Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails," *Journal of Econometrics*, 163(2), 215–230.

Engelberg, J., C. F. Manski, and J. Williams (2009): "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters," *Journal of Business and Economic Statistics*, 27(1), 30–41.

Engle, R. F., and S. Manganelli (2004): "CAViaR," *Journal of Business & Economic Statistics*, 22(4), 367–381.

Ferguson, T. S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Galbraith, J. W., and S. van Norden (2011): "Kernel-Based Calibration Diagnostics for Recession and Inflation Probability Forecasts," *International Journal of Forecasting*, 27(4), 1041–1057.

Geweke, J., and G. Amisano (2011): "Optimal Prediction Pools," *Journal of Econometrics*, 164(1), 130–141.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74(6), 1545–1578.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007): "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society: Series B*, 69, 243–268.

Gneiting, T., and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., and R. Ranjan (2013): "Combining Predictive Distributions," *Electronic Journal of Statistics*, 7, 1747–1782.

Granger, C. W. J. (1999): "Outline of Forecast Theory Using Generalized Cost Functions," *Spanish Economic Review*, 1, 161–173.

Granger, C. W. J., and M. H. Pesaran (2000): "A Decision-Theoretic Approach to Forecast Evaluation," in *Statistics and Finance: An Interface*, ed. by W.-S. Chan, W. K. Li, and H. Tong. Imperial College Press.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

Holzmann, H., and M. Eulert (2014): "The Role of the Information Set for Forecasting — With Applications to Risk Management," *The Annals of Applied Statistics*, 8(1), 595–621.

Kallenberg, O. (2002): *Foundations of Modern Probability*. Springer, 2 edn.

Knüppel, M. (2015): "Evaluating the Calibration of Multi-Step-Ahead Density Forecasts Using Raw Moments," *Journal of Business and Economic Statistics*, 33(2), 270–281.

Kupiec, P. H. (1995): "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, 3(2), 73–84.

Lahiri, K., and L. Yang (2015): "A Further Analysis of the Conference Board's New Leading Economic Index," *International Journal of Forecasting*, 31(2), 446–453.

Lichtenstein, S., B. Fischhoff, and L. D. Phillips (1982): "Calibration of Probabilities: The State of the Art to 1980," in *Judgment under Uncertainty: Heuristics and Biases*, ed. by D. Kahneman, P. Slovic, and A. Tversky, pp. 306–334. Cambridge University Press, Cambridge, UK.

Lopez, J. A. (1998): "Methods for Evaluating Value-at-Risk Estimates," *Economic Policy Review*, (October), 119–124.

Mincer, J. A., and V. Zarnowitz (1969): "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, ed. by J. A. Mincer, pp. 3–46. National Bureau of Economic Research.

Mitchell, J., and S. G. Hall (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.

Mitchell, J., and K. F. Wallis (2011): "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness," *Journal of Applied Econometrics,*, 26(6), 1023–1040.

Murphy, A. H., and R. L. Winkler (1987): "A General Framework for Forecast Verification," *Monthly Weather Review*, 115, 1330–1338.

Pascual, L., J. Romo, and E. Ruiz (2001): "Effects of parameter estimation on prediction densities: a bootstrap approach," *International Journal of Forecasting*, 17, 83–103.

Shiller, R. J. (1978): "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review," *Journal of Monetary Economics*, 4(1), 1–44.

The National Institute of Economic Research (2013): "The Riksbank Has Systematically Overestimated Inflation," *The Swedish Economy*, pp. 29–33.

Tsyplakov, A. (2011): "Evaluating Density Forecasts: A Comment," MPRA Paper 32728, University Library of Munich, Germany.

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(2), 165–175.

Appendix A. Theorems and proofs

Theorem 1. *For any forecast $\hat{F}$ in a density forecasting situation PIT value $\hat{F}(y)$ is a random variable, which is measurable with respect to $\sigma(\hat{F}, y)$. Moreover, $\sigma(\hat{F}, y) = \sigma(\hat{F}, \hat{F}(y))$.*

Proof. Fix some real $\alpha \in [0,1]$. Note that for any real q

$$(\hat{F}(y) \le \alpha) \subseteq (\hat{F}(q) \le \alpha) \cup (y \le q),$$

since values of $\hat{F}$ are non-decreasing functions. Thus we have $(\hat{F}(y) \le \alpha) \subseteq C_\alpha$, where $C_\alpha = \cap_{q \in \mathbb{Q}}((\hat{F}(q) \le \alpha) \cup (y \le q))$. Note that $C_\alpha \in \sigma(\hat{F}, y)$, because $(\hat{F}(q) \le \alpha) \cup (y \le q) \in \sigma(\hat{F}, y)$ for each q. If $\hat{F}(y) > \alpha$, then by continuity of values of $\hat{F}$ there exists $q \in \mathbb{Q}$ such that $\alpha < \hat{F}(q) < \hat{F}(y)$. Consequently $(\hat{F}(y) > \alpha) \cap C_\alpha = \emptyset$ and $C_\alpha = (\hat{F}(y) \le \alpha)$. Thus, $(\hat{F}(y) \le \alpha) \in \sigma(\hat{F}, y)$ for any real α, which proves that $\hat{F}(y)$ is measurable with respect to $\sigma(\hat{F}, y)$.

We thereby proved that $\sigma(\hat{F}, \hat{F}(y)) \subseteq \sigma(\hat{F}, y)$.

By the same logic in order to prove that $\sigma(\hat{F}, y) \subseteq \sigma(\hat{F}, \hat{F}(y))$ we prove that $(y \le q) \in \sigma(\hat{F}, \hat{F}(y))$ for any $q \in [a, b)$, where $[a, b]$ is the region of strict monotonicity of values of $\hat{F}$. We similarly note that $(y \le q) \subseteq (\hat{F}(q) \ge \alpha) \cup (\hat{F}(y) \le \alpha)$ for any real α since values of $\hat{F}$ are non-decreasing. Thus $(y \le q) \subseteq D_q$, where $D_q = \cap_{\alpha \in \mathbb{Q}}((\hat{F}(q) \ge \alpha) \cup (\hat{F}(y) \le \alpha))$, $D_q \in \sigma(\hat{F}, \hat{F}(y))$. If $y > q$ for $q \in [a, b)$, then $\hat{F}(q) < \hat{F}(y)$ and there exists $\alpha \in \mathbb{Q}$ such that $\hat{F}(q) < \alpha < \hat{F}(y)$. This proves that $D_q = (y \le q)$ and $(y \le q) \in \sigma(\hat{F}, \hat{F}(y))$. □

Theorem 2. *If a forecast $\hat{F}$ is ideally calibrated with respect to Ψ^*, then it is conditionally auto-calibrated with respect to any information set Ψ such that $\Psi \subseteq \Psi^*$.*

Proof. Since $\sigma(\Psi, \hat{F}) \subseteq \Psi^*$ we have $\mathbb{F}(q|\Psi, \hat{F}) = \mathsf{E}[\mathbb{F}(q|\Psi^*)|\Psi, \hat{F}] = \mathsf{E}[\hat{F}(q)|\Psi, \hat{F}] = \hat{F}(q)$. □

Theorem 3. *If a forecast $\hat{F}$ is conditionally auto-calibrated with respect to Ψ, then it is marginally calibrated given Ψ and (in a density forecasting situation) probabilistically calibrated given Ψ.*

Proof. Marginal calibration follows from $\mathsf{E}(\mathbb{F}(q|\Psi, \hat{F})|\Psi) = \mathbb{F}(q|\Psi)$ for $q \in \mathbb{R}$. From (1) for $g = \mathrm{I}\{F(y) \le p\}$ and $p \in \mathbb{R}$ it follows that $\hat{F}(y)|\Psi, \hat{F} \sim U[0,1]$, because in a density forecasting situation the corresponding $\hat{g}$ is the value of the $U[0,1]$ CDF at p. This entails $\hat{F}(y)|\Psi \sim U[0,1]$ (probabilistic calibration). □

Theorem 4. *If a forecast $\hat{F}$ is marginally calibrated given Ψ, then for any n and any Ψ-measurable w it satisfies*

$$\mathsf{E}n(y,w) = \mathsf{E}\int_{-\infty}^{\infty} n(t,w)d\hat{F}(t).$$

Proof. Apply the law of iterated expectations to

$$\mathsf{E}(n(y,w)|\Psi) = \int_{-\infty}^{\infty} n(t,w)d\mathbb{F}(t|\Psi) = \int_{-\infty}^{\infty} n(t,w)d\mathsf{E}(\hat{F}(t)|\Psi) = \mathsf{E}\left[\int_{-\infty}^{\infty} n(t,w)d\hat{F}(t)\,\middle|\,\Psi\right].$$

□

Theorem 5. *In a density forecasting situation a forecast, which is both probabilistically and marginally calibrated given Ψ, can be not auto-calibrated given Ψ.*

Proof. It suffices to find an example with a trivial Ψ. Suppose that the actual distribution of y is described by its conditional distribution given w: $\mathbb{F}(q|w) = F_w^\circ(q)$, where $w = 1$ or $w = 2$ with equal probabilities and

$$F_1^\circ(q) = \begin{cases} \frac{3}{2}q, & q \in [0, \frac{1}{4}], \\ \frac{1}{2}q + \frac{1}{4}, & q \in [\frac{1}{4}, \frac{3}{4}], \\ \frac{3}{2}q - \frac{1}{2}, & q \in [\frac{3}{4}, 1], \end{cases} \qquad F_2^\circ(q) = \begin{cases} \frac{1}{2}q, & q \in [0, \frac{1}{4}], \\ \frac{3}{2}q - \frac{1}{4}, & q \in [\frac{1}{4}, \frac{3}{4}], \\ \frac{1}{2}q + \frac{1}{2}, & q \in [\frac{3}{4}, 1]. \end{cases}$$

Forecast $\hat{F}$ is also based on w: $\hat{F}(q) = F_w(q)$, where

$$F_1(q) = \begin{cases} q, & q \in [0, \frac{1}{2}], \\ \frac{1}{2}q + \frac{1}{4}, & q \in [\frac{1}{2}, \frac{3}{4}], \\ \frac{3}{2}q - \frac{1}{2}, & q \in [\frac{3}{4}, 1], \end{cases} \qquad F_2(q) = \begin{cases} q, & q \in [0, \frac{1}{2}], \\ \frac{3}{2}q - \frac{1}{4}, & q \in [\frac{1}{2}, \frac{3}{4}], \\ \frac{1}{2}q + \frac{1}{2}, & q \in [\frac{3}{4}, 1]. \end{cases}$$

Since $\frac{1}{2}F_1^\circ(q) + \frac{1}{2}F_2^\circ(q) = \frac{1}{2}F_1(q) + \frac{1}{2}F_2(q)$ ($= q$ for $q \in [0,1]$) and $\frac{1}{2}F_1^\circ(F_1^{-1}(\alpha)) + \frac{1}{2}F_2^\circ(F_2^{-1}(\alpha)) = \alpha$, it can be seen that $\hat{F}$ is both marginally and probabilistically calibrated. However, $\sigma(\hat{F}) = \sigma(w)$ and thus for $\hat{F}$ to be auto-calibrated we must have $\hat{F} = F_w^\circ$ which is not the case here. □

Theorem 6. *If forecast $\hat{F}$ is Ψ-measurable and is either marginally calibrated or (in a density forecasting situation) probabilistically calibrated with respect to Ψ, then $\hat{F}$ is ideally calibrated with respect to Ψ.*

Proof. For marginal calibration obviously $\mathbb{F}(q|\Psi) = \mathsf{E}(\hat{F}(q)|\Psi) = \hat{F}(q)$. For probabilistic calibration $\mathbb{F}(q|\Psi) = \mathsf{E}(\mathrm{I}\{y \le q\}|\Psi) = \mathsf{E}(\mathrm{I}\{\hat{F}(y) \le \hat{F}(q)\}|\Psi) = \hat{F}(q)$. □

Theorem 7. *If a forecast $\hat{F}$ satisfies condition*

$$\mathsf{E}(r(y,\hat{F})|\Psi) = \mathsf{E}\left[\int_{-\infty}^{\infty} r(t,\hat{F})d\hat{F}(t) \;\middle|\; \Psi\right]$$

for any r, then it is conditionally auto-calibrated with respect to Ψ.

Proof. Let $r(y,F) = \mathrm{I}\{y \le q\}a(w,F)$, where a is some function with additional variable w playing the role of a placeholder. For arbitrary q, a and w we must have

$$\mathsf{E}[(\mathrm{I}\{y \le q\} - \hat{F}(q))a(w,\hat{F})|\Psi] = 0.$$

By the substitution property of conditional expectation fixed w here can be replaced by an arbitrary Ψ-measurable random variable w. Then by the law of iterated expectations

$$\mathsf{E}[(\mathrm{I}\{y \le q\} - \hat{F}(q))a(w,\hat{F})] = 0.$$

Since a here is arbitrary, it follows that

$$\mathsf{E}[\mathrm{I}\{y \le q\} - \hat{F}(q)|\Psi,\hat{F}] = 0,$$

for any $q \in \mathbb{R}$ which is equivalent to $\hat{F}(q) = \mathbb{F}(q|\Psi,\hat{F})$. □

Appendix B. Location and scale efficiency tests based on the GCsL score

The GCsL scoring rule is defined as

$$S(y,F) = \gamma(y)\log F'(y) + (1-\gamma(y))\log(I_1(F))$$

for a weighting function $\gamma(y)$, where

$$I_1(F) = \int_{-\infty}^{\infty}(1-\gamma(t))dF(t).$$

Following the logic of subsection 3.2, it can be derived that the location efficiency test corresponding to the transformation (14) is a test of orthogonality between

$$-\gamma(y)\hat{\ell}'(y) - (1-\gamma(y))I_2(\hat{F})/I_1(\hat{F})$$

and some conditioning variable w, where $\hat{\ell}(y)$ is the log-density corresponding to $\hat{F}$ and

$$I_2(F) = \int_{-\infty}^{\infty} (1-\gamma(t))\ell'(t)dF(t).$$

Indeed, for $F_\delta(y) = F(y - w'\delta)$ and $\ell(y) = \log(F'(y))$ we have

$$S(y, F_\delta) = \gamma(y)\ell(y - w'\delta) + (1-\gamma(y))\log(I_1(F_\delta)),$$

where

$$I_1(F_\delta) = \int_{-\infty}^{\infty} (1-\gamma(t))dF(t - w'\delta) = \int_{-\infty}^{\infty} (1-\gamma(t))\exp(\ell(t - w'\delta))dt.$$

Since

$$\frac{d}{d\delta} I_1(F_\delta)\mid_{\delta=0} = \int_{-\infty}^{\infty} (1-\gamma(t))\ell'(t)\exp(\ell(t))dt \cdot (-w) = I_2(F)\cdot(-w),$$

we obtain

$$\frac{d}{d\delta} S(y, F_\delta)\mid_{\delta=0} = \gamma(y)\ell'(y)\cdot(-w) + (1-\gamma(y))I_2(F)/I_1(F)\cdot(-w).$$

Similarly a scale efficiency test corresponding to the transformation (15) is a test of orthogonality between

$$-\gamma(y)\hat{\ell}'(y)(y - c(\hat{F})) - (1-\gamma(y))I_3(\hat{F})/I_1(\hat{F}) - 1$$

and some variable w, where $c(F)$ is a central point of F and

$$I_3(F) = \int_{-\infty}^{\infty} (1-\gamma(t))\ell'(t)(t - c(F))dF(t).$$

For a numerical integration method needed to calculate I_1, I_2 and I_3 see footnote 20.

Appendix C. Riksbank's and NIER's forecasts of Swedish inflation

The Riksbank's density forecasts of CPI inflation were prepared four times a year (at approximately quarterly frequency) and were published in the bank's *Inflation Report* until 2006. Since 2007 they were published in *Monetary Policy Report* three times a year with additional forecasts in *Monetary Policy Updates*. We used the April *Monetary policy update* data to complement the *Monetary Policy Report*. There was a one-time shift in publication dates. Also there was a gap due to missing 2006:4 issue of *Inflation Report*. In 2004 the definition of the Swedish Consumer price index has changed. We used the old index for the

outcomes of the target variable up to December of 2004 (the data are from the 2004:4 issue of *Inflation Report*). The so called *shadow* version of CPI was used.

The forecasts were in the form of a two-piece normal distribution with the CDF given by

$$\begin{cases} \frac{2\sigma_1}{\sigma_1+\sigma_2}\Phi\left(\frac{y-\mu}{\sigma_1}\right), & y \le \mu, \\ 1-\frac{2\sigma_2}{\sigma_1+\sigma_2}\Phi\left(-\frac{y-\mu}{\sigma_2}\right), & y > \mu, \end{cases}$$

where $\Phi(\cdot)$ is the standard normal CDF, μ is the mode of the distribution, σ_1 as σ_2 are the left and right scaling parameters. Since 2007 the forecasts were in the form of a symmetric normal distribution (which is a special case of a two-piece normal with $\sigma_1 = \sigma_2$). The numerical data for the boundaries of central predictive intervals and the mode μ can be found on the Riksbank's web site.[21] The parameters σ_1, σ_2 were recovered from the boundaries and the mode by the nonlinear least squares.

The point forecasts of inflation by the National Institute of Economic Research (used for Riksbank's forecasts evaluation) were issued 4 times a year. With each Riksbank's forecast we associate the nearest potentially available NIER's forecast, which correspond to the same target date. The data for the forecasts issued from 2001 to 2010 are the same as used in NIER (2013) publication. More recent forecasts are available from the NIER's website.[22]

[21] `http://www.riksbank.se`.
[22] `http://www.konj.se/757.html`.