Theoretical guidelines for a partially informed forecast examiner

Tsyplakov, Alexander

Department of Economics, Novosibirsk State University

2 April 2014

Theoretical Guidelines for a Partially Informed Forecast Examiner

Alexander Tsyplakov

Department of Economics, Novosibirsk State University

October 19, 2015

Abstract

The paper explores theoretical foundations behind evaluation of probabilistic forecasts. The emphasis is on a situation when the forecast examiner possesses only partially the information which was available and was used to produce a forecast. We argue that in such a situation forecasts should be judged by their 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 growing popularity of more complete—probabilistic—forecasts in econometrics.

Real-life forecasts are not perfect and we want to be able to diagnose imperfections in order to improve our forecasting methods. Several procedures were used for testing forecast calibration and efficiency in the literature on probabilistic 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) (this list includes closely related literature on interval/quantile forecasts). 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 we can call a forecast calibrated or efficient are not yet fully understood and formally stated. In this paper we want to make up for this omission.

Some papers on evaluation of probabilistic forecasts assume that there is a parametric model of the data which give the correct conditional distribution given information set for some true vector of parameter values (e.g. Corradi and Swanson, 2006a; Corradi and Swanson, 2006c; Chen, 2011). From this point of view forecast evaluation is a kind of model evaluation. The probabilistic properties of forecast evaluation statistics are governed by the model. However, this approach is not applicable when there is no formal parametric model behind the forecasts. Our view is that requirements to a good probabilistic forecast should not refer to a "model" or the "true parameters". It is more realistic to consider forecasting *methods* rather than forecasting *models* (cf. Giacomini and White, 2006). This permits us not to exclude forecasts which are not based on formal parametric models (e.g. survey forecasts, forecasts utilizing ad hoc exponential smoothing or neural networks).

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. A good principle of forecast evaluation, which the examiner could adhere, is to treat forecasts at their face value, in a black-box manner, without paying attention to their origin. The reason for this is simple. To test a bricklayer one can give him a trowel and measure the speed and quality of his work. Similarly, a direct and natural way to evaluate a forecasting technique is to make it being applied to some pertinent real-life problem and to compare the forecasts with the outcomes. Any extra requirements refer to something else rather than pure predictive performance. This does not mean that the examiner cannot take into account more complex considerations (for example, a would-be predictive performance in different circumstances), but such considerations require proper substantiation.

It is true that many forecasts are based on parametric models and the corresponding parameters have to be estimated from the data. If this is the case, then a forecasting method includes a model, some method of parameter estimation (MLE, GMM, etc.), some scheme of utilizing data (like recursive, rolling estimation, etc.) and other components. A common phenomenon is plain plug-in forecasting when estimated parameters are used as the true parameters and parameter uncertainty is neglected. Note that the black-box approach to forecast evaluation requires that if we want to take into account parameter uncertainty, then it should be embedded into the forecasts *before* forecast evaluation as it is not fair to handicap a forecasting method on the ground of unaccounted parameter uncertainty. For example, if the

model is a classical linear regression with Gaussian errors then the forecast can be stated as the corresponding scaled and shifted Student's distribution. For other models in general there is no unique natural way to deal with parameter uncertainty; however, forecasting literature provides many suggestions (e. g, Cooley and Parke, 1990; Barndorff-Nielsen and Cox, 1996). Of course, parameter uncertainty is an issue only in the frequentist framework; a typical Bayesian approach is also model-based, but it incorporates parameter uncertainty in a consistent way.

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).

This paper applies a similar approach to complete probabilistic forecasts (predictive distributions). Unlike most of the existing literature (Corradi and Swanson, 2006c is a vivid example) we emphasize the probability theory basis behind probabilistic forecasting, rather than statistical testing. 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.

We consider forecasting of some target outcome y which is a real-valued random variable. A complete probabilistic forecasts 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). To analyze the properties of a probabilistic forecast it is necessary to consider the joint distribution of the forecast $\hat{F}$ and the target outcome variable y specified on a common probability space. 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). Calibration means good conformity between a probabilistic forecast and the actual behavior of the target variable. However, this idea of conformity is vague and requires an accurate formulation. When evaluating a point forecast one compares one point to another, which is relatively simple. When evaluating a complete probabilistic forecast one has to compare a point to a CDF. Moreover, the CDF is random rather than fixed.

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.

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. There are many real-life situations in which Ψ^* and Ψ 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.

Therefore, 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 information set. Even in the case of academic forecasters, who rely on much clearer methods and data sources than augurs, an external examiner is not always able to ascertain the quality and integrity of these methods and sources. In general, an examiner needs procedures which can convincingly demonstrate imperfections of a miscalibrated forecast irrespectively of its origin.

Our 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 in-

formation 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 $\mathsf{E}[y|\Psi] = \hat{y}$.

For practical reasons it is convenient to express the implications of calibration in the form of moment conditions. 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 express them in the form of moment conditions.

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 auto-calibration.

The notion of auto-calibration can be further extended in two important directions. First, for a situation of sequential time-series forecast evaluation it generalizes to *sequential calibration* (which can be compared to the condition of uniformity and independence for PIT values). Second, in a situation, when several rival forecasts are available, it generalizes to *forecast encompassing*.

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* (also called optimality and rationality in different contexts). We call a forecast efficient if it is a maximizer of a pertinent performance measure (a score calculated according to some *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.

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, the last subsection of the section 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 and a technical counterexample 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 $\hat{F}_1, \ldots, \hat{F}_N$ (in the form of CDFs) and a sequence of points $y_1, \ldots, y_N$ 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 1). The forecasts are in the form of 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 was potentially accessible to the Riksbank's staff, such as any official statistics published before the date of forecast issue.

The current paper approaches to 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.[1] 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.

[1] 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).

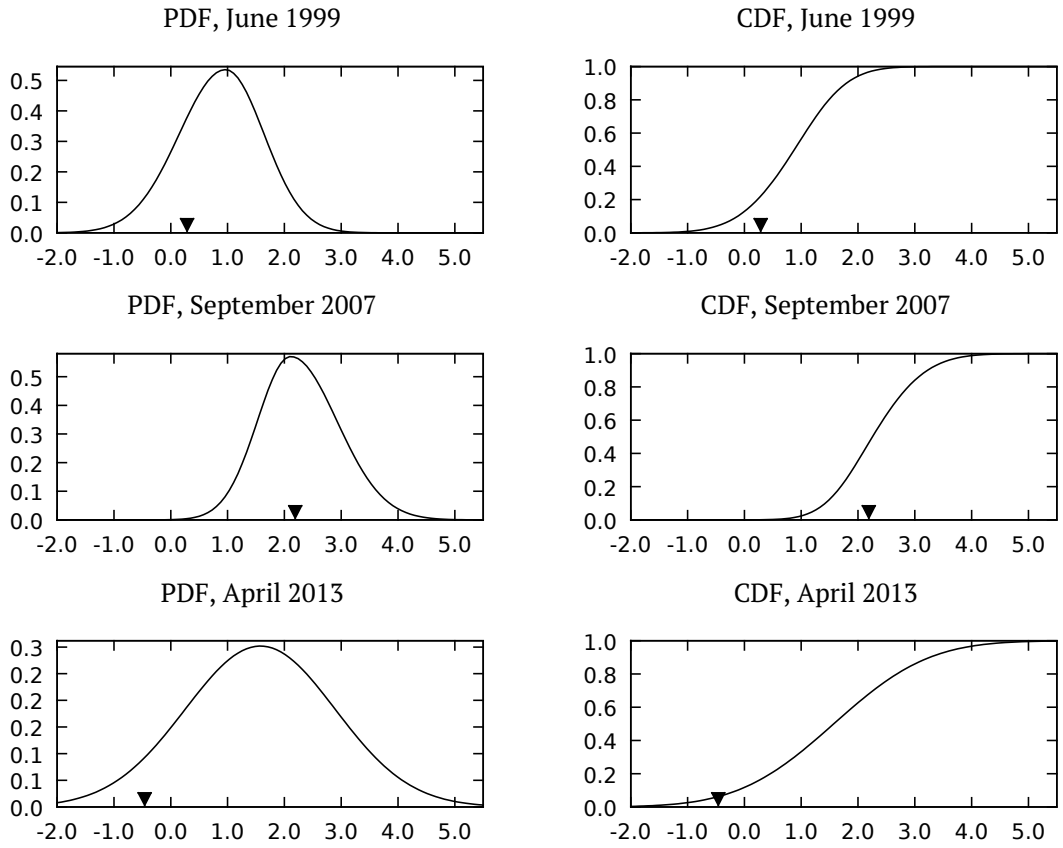


Figure 1: Riksbank's one-year-ahead CPI inflation forecasts. The forecasts are for the specified dates. Left graphs show the forecasts in the form of probability density function, right ones—the same forecasts in the form of cumulative distribution functions. Actual outcomes of the CPI inflation are marked by triangles.

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 $q \in \mathbb{R}$.

If the forecaster possesses some information which is not available to the examiner, then the examiner can potentially derive some new information from the forecast itself.[2] 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

[2]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".

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 $q \in \mathbb{R}$.

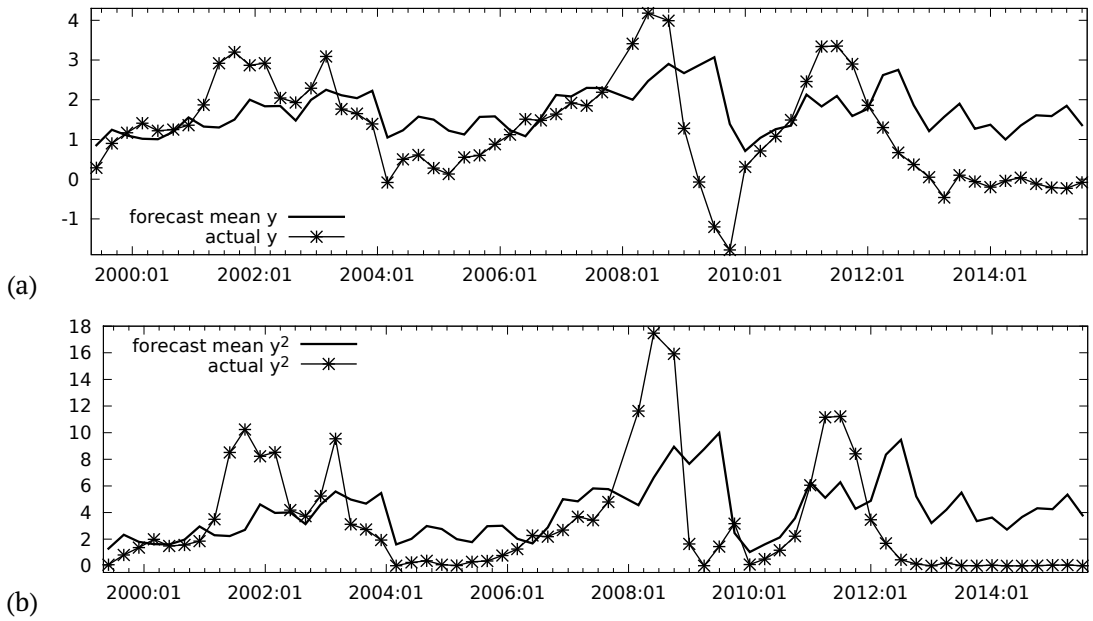
By definition any auto-calibrated forecast is ideally calibrated with respect to an appropriate information set. Moreover, a Ψ-measurable forecast, which is auto-calibrated given Ψ, must be ideally calibrated given Ψ. Conversely, it can be stated that any forecast, which is ideally calibrated with respect to some information set Ψ^* including Ψ, is auto-calibrated with respect to Ψ (Theorem 7). Therefore, 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.

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



(a)

(b)

Figure 2: Forecasts of the mean (a) and mean square (b) derived from the Riksbank's inflation forecasts compared to the actual outcomes.

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 1(a) plote 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, we can calculate the theoretical mean of y^2 implied by $\hat{F}$ and use it as a point forecast for y^2. See figure 1(b), which compares the forecasts of squared inflation to the actual squared inflation.

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 $\hat{F}$.[3] 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.

In fact, these 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 2.5). 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. Dropping $\hat{F}$ from g, that is, letting $g = n(y, w)$, produces a condition of conditional marginal calibration. In the 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. Next subsection discusses the two concepts of calibration, probabilistic and marginal calibration. Subsection 2.4 discusses orthogonality conditions, which are produced

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

by letting $g = r(y, \hat{F})a(w, \hat{F})$. In subsection 2.5 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 and marginal calibration

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 the required 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 $\hat{F}(y)$.[4] The PIT value $\hat{F}(y)$ is a random variable which, given the forecast, carries the same information as the target variable y. This 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 6 in Appendix.)

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).[5] 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).

[4] 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.

[5] 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.

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).[6] 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 condition 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: $w = I_A$.[7] Note that moment conditions (3) are weaker than the general moment conditions of conditional auto-calibration (2).

Further references and examples of moment conditions of probabilistic calibration can be found in subsection 2.4 and in Chen (2011).

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

[6] 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.

[7] 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 and thus by Definition 3 forecast $\hat{F}$ is probabilistically calibrated given Ψ.

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 9). 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: $w = I_A$.[8]

In particular, for $n = y$ we obtain the condition of mean unbiasedness of $\hat{F}$: $\mathsf{E}y = \mathrm{mean}(\hat{F})$. Gneiting, Balabdaoui, and Raftery (2007) propose a diagnostic diagram for unconditional marginal calibration based on binning and application of the condition $\mathsf{E}\mathrm{I}\{y \in (a,b]\} = \mathsf{E}[\hat{F}(b) - \hat{F}(a)]$ to the bins. Note also that figure 1 above can be considered as an illustration of a testing procedure for marginal calibration.

Theorem 8 states that both probabilistic and marginal calibration given Ψ are implied by auto-calibration given Ψ.[9] However, as we show below (subsection 2.5), even together probabilistic and marginal calibration are not sufficient for auto-calibration with respect to the same information set. Hence, calibration tests based on (3) and (4) can be incomplete as tests of conditional auto-calibration.

2.4 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.[10] 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:[11] if a forecast $\hat{F}$ is conditionally auto-calibrated with respect to

[8] 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.

[9] Gneiting and Ranjan (2013) observe that the ideally calibrated forecast is both (unconditionally) marginally calibrated and probabilistically calibrated.

[10] 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).

[11] 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.

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 $\hat{F}$. In practice one can represent a CDF F in function $a(w,F)$ by some characteristics of the corresponding distribution such as the mean, median or interquartile range.

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 as its argument and any Ψ-measurable w we must have

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

Similarly under conditional marginal calibration with respect to Ψ for any function $m(y)$ 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.

2.5 Sufficient conditions of 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.

We have already seen (Theorem 8) that 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[12] 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). See also Forecast C in Example 2 below.[13]

It can be seen that neither probabilistic, nor marginal calibration given Ψ is sufficient for auto-calibration given Ψ. Example 10 in the Appendix 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 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 11).

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:

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

[12] 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.

[13] 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).

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 12 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.[14]

2.6 Sequential auto-calibration

Consider a sequence $\hat{F}_t$ of h-step-ahead probabilistic forecasts of a univariate time series y_t, $t = 1, 2, \ldots$ in a recursive setting. The examiner's information set available for evaluating $\hat{F}_t$, which we denote Ψ_t, should include information on available previous values of the series and previously issued forecasts. We assume that all forecasts $\hat{F}_1, \ldots, \hat{F}_t$ are already known at time $t - h$ and thus

$$\sigma(y_1, , \ldots, y_{t-h}, \hat{F}_1 \ldots, , \hat{F}_{t-1}) \subseteq \Psi_t.$$

This suggests the following definition.

Definition 5. A sequence of forecasts $\hat{F}_t$, $t = 1, \ldots, T$ in a recursive h-step density forecasting situation is *sequentially auto-calibrated* if each forecast $\hat{F}_t$ is conditionally auto-calibrated with respect to $(y_1, , \ldots, y_{t-h}, \hat{F}_1 \ldots, , \hat{F}_{t-1})$.

If a sequence of one-step density forecasts of a time series y_t, $t = 1, \ldots., T$ is made from the full history of the same series, then calibration is frequently judged by analyzing the resulting series of PIT values

$$\hat{F}_t(y_t), \qquad t = 1, \ldots., T.$$

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

It is assumed that a sequence of such forecasts is calibrated if and only if the PIT values $\hat{F}_t(y_t)$ are independent and distributed as $U[0,1]$ (cf. Diebold, Gunther, and Tay, 1998). We will call this the uniformity and independence condition:

$$(\hat{F}_1(y_1),\dots,\hat{F}_T(y_T)) \sim U[0,1]^T. \tag{11}$$

The condition is very popular in the density forecast evaluation literature; e.g. Dawid (1984), Diebold, Gunther, and Tay (1998), Berkowitz (2001), Mitchell and Wallis (2011), Chen (2011). Mitchell and Wallis (2011) even call this "complete calibration". However, this is in fact not an independent mode of calibration. It can be a necessary condition of sequential auto-calibration or a sufficient condition of sequential ideal calibration or completely irrelevant depending on the situation.

Condition (11) should be primarily considered as a necessary condition of sequential auto-calibration in a specific setting. If in a recursive one-step-ahead density forecasting situation forecasts $\hat{F}_t$, $t = 1,\dots,T$ are sequentially auto-calibrated, then according to Theorem 13 (11) must hold. This is a generalization of Proposition in Diebold, Gunther, and Tay (1998), p. 867 in the spirit of partial information approach of the current paper.

For a recursive one-step-ahead density forecasting situation an interesting question is whether the uniformity and independence condition (11) is sufficient for sequential auto-calibration. In general the answer is negative. However, when we are sure that the forecaster uses only the previous history to produce forecasts, then conditioning on the previous history of y_t is equivalent to conditioning on the previous history of PIT values $\hat{F}_t(y_t)$ and, hence, the uniformity and independence condition is sufficient not only for sequential auto-calibration, but for sequentially *ideal* calibration. That is, according to Theorem 14 under (11) each $\hat{F}_t$ is ideally calibrated given $(y_1,,\dots,y_{t-1})$.

When forecasts are not measurable with respect to $\sigma(y_1,\dots,y_{t-1})$, uniformity and independence is not sufficient; it does not indicate a sequence of calibrated forecasts. Although for multistep forecasts, which are based only on $(y_1,\dots,y_{t-h})$, the uniformity and independence condition is sufficient for sequentially ideal calibration, it is not very useful since in general independence does not hold anyway. It can be additionally noted that for forecasts using real-time data subject to revisions the condition of independence of PIT values can be completely irrelevant.

The condition of serial independence of PIT values, which is a part of (11), can be expressed with the help of orthogonality conditions. For example, any function k of a the PIT value for moment t must be

uncorrelated with any function k_2 of lagged PIT values:

$$\mathsf{E}\left[\left(k(\hat{F}_t(y_t)) - \int_0^1 k(p)dp\right) k_2(\hat{F}_{t-s}(y_{t-s}))\right] = 0, \qquad s = 1,2,\ldots \tag{12}$$

Under uniformity and independence a series of transformed PIT values $k(\hat{F}_t(y_t))$, $t = 1,\ldots,T$ also must be i. i. d. and hence serially uncorrelated. Therefore, we can use autocorrelation functions of the PIT values and their transformations to test sequential auto-calibration of recursive forecasts as proposed in Diebold, Gunther, and Tay (1998). When the uniformity and independence condition is not applicable, we can still employ similar orthogonality conditions provided that we use only Ψ_t-measurable PIT variables for conditioning. For example, in the case of h-step-ahead forecasting we can use (12) for $s \geq h$. In the case of real-time forecasting if some preliminary estimates of y_{t-s}, say y^*_{t-s}, are observed at the time when the forecast is made, then we can use (12) with y_{t-s}, replaced by y^*_{t-s}.

In general there is no need to rely only on orthogonality conditions based on PIT values. Other conditions can be more suitable in many forecasting situations. It should be emphasized in particular that even though $\hat{F}_{t-h+1}(y_{t-h+1}),\ldots,\hat{F}_{t-1}(y_{t-1})$ cannot be used for conditioning purposes in evaluation of h-step-ahead forecast of y_t, various functions of forecasts $\hat{F}_{t-h+1},\ldots,\hat{F}_t$ can.

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.

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

$$\mathsf{E}g(y,w,\hat{F}_2) = \mathsf{E}\int_{-\infty}^{\infty} g(t,w,\hat{F}_2)d\hat{F}_1(t). \tag{13}$$

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 full probabilistic forecasts 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

$$\mathsf{E}[g(y, w, \hat{F}_2) - g(y, w, \hat{F}_1)] = \mathsf{E}\left[\int_{-\infty}^{\infty} g(t, w, \hat{F}_2)\, d\hat{F}_1(t) - \int_{-\infty}^{\infty} g(t, w, \hat{F}_1)\, d\hat{F}_1(t)\right]. \tag{14}$$

The two forms of forecast encompassing conditions, (13) and (14), 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, (20) 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 (11). 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 tests of calibration/efficiency 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.

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. 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 $\bar{d} / se(\bar{d})$.[15] The most subtle aspect here is adequate calculation of the standard error $se(\bar{d})$ for dependent d_i. Figure 2 illustrates that for multistep forecasts the d_i series can be strongly autocorrelated. In Example 1 below the usual heteroskedasticity and autocorrelation consistent (HAC) 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 of 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).

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. When forecasts are in the form of distribution functions it is natural to base discussion of forecast efficiency on the notion of a proper scoring rule. A *scoring rule* is a function $S(F, y)$ of a CDF F and an outcome value y used to judge the accuracy or success of full probabilistic forecasts. If $\hat{F}_1, \ldots, \hat{F}_N$ is a series of realizations of predictive distribution functions, and $y_1, \ldots, y_N$ is a series of realized outcomes, then the average score is given by $\sum_{i=1}^{N} S(\hat{F}_i, y_i) / N$. It is assumed that a more accurate/successful forecast has a higher average score.

Not any arbitrary scoring rule is suitable for forecast evaluation. The general requirement is that scoring rules used for forecast evaluation must be *proper*. One can define the expected score function as

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

the expected score of F_2 given that y is distributed as F_1:

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

(Note the overloaded notation used.) By definition, if the scoring rule S is proper, then the expected score is maximized with respect to F_2, when F_2 coincides with F_1:

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

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.) Proper scoring rules are known to encourage truthful forecast statement: if a forecast is assessed according to a proper scoring rule, then the forecaster cannot expect to benefit by cheating and reporting forecast distributions which he believes to be incorrect.

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).

The notion of a proper scoring rule is closely related to maximization of expected utility or minimization of expected loss by a forecast user. Indeed, define a scoring rule S as the utility of an outcome y under the best action

$$S(F, y) = u(y, a(F)),$$

where the best action $a(F)$ is given by (e.g. Pesaran and Skouras, 2002)

$$a(F) \in \arg\max_a \int_{-\infty}^{\infty} u(t, a) dF(t).$$

Such a utility-based scoring rule is proper since

$$S(F_1, F_1) = \int_{-\infty}^{\infty} u(t, a(F_1)) dF_1(t) \geq \int_{-\infty}^{\infty} u(t, a(F_2)) dF_1(t) = S(F_2, F_1)$$

(cf. Diebold, Gunther, and Tay, 1998, Gneiting and Raftery, 2007). Maximization of expected utility provides economic foundation for the theory of evaluation of probabilistic forecasts, but one can abstract from this and focus instead on proper scoring rules.

In our analysis we can allow the scoring rule to depend on some additional variable w, which represents the environment conditions at the moment, when the forecast is made, and which can effect the efficiency of the forecasts: $S = S(F, y; w)$. If Ψ is the relevant information set, then one can use a Ψ-measurable random variable here. However, we prefer to economize on notation and hide the additional variables.

An important property of an ideally calibrated forecast is that it achieves the maximum expected score if the scoring rule used is proper. Diebold, Gunther, and Tay (1998), p. 866: "...If a forecast coincides with the true data generating process, then it will be preferred by all forecast users, regardless of loss function". See also Granger and Pesaran (2000). Formally, for any proper scoring rule S the forecast, which is ideally calibrated with respect to Ψ, attains the highest expected score among the Ψ-measurable forecasts. Indeed, if $\mathring{F} = \mathbb{F}(\cdot|\Psi)$ is the ideal forecast and $\hat{F}$ is another Ψ-measurable forecast

$$\mathsf{E}(S(\mathring{F}, y)|\Psi) = \int_{-\infty}^{\infty} S(\mathring{F}, t) d\mathring{F}(t) = S(\mathring{F}, \mathring{F}) \geq S(\hat{F}, \mathring{F}) = \int_{-\infty}^{\infty} S(\hat{F}, t) d\mathring{F}(t) = \mathsf{E}(S(\hat{F}, y)|\Psi)$$

and (by the law of iterated expectations)

$$\mathsf{E}S(\mathring{F}, y) \geq \mathsf{E}S(\hat{F}, y).$$

Thus, when a forecast is ideal with respect to Ψ, it can be called efficient or optimal. Under appropriate additional conditions the inequality here is strict if the scoring rule S is strictly proper and the alternative forecast is not ideal (Holzmann and Eulert, 2014).

Therefore, 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 and $\Psi \subseteq \Psi^*$, then an efficient forecast based on Ψ^* must be ideally calibrated given Ψ^* and thus auto-calibrated given Ψ 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

We can use the first order conditions of score maximization to derive moment conditions of efficiency. Consider a CDF-to-CDF transformation $T(F, w, \delta)$ depending on a real vector of parameters δ and an additional variable w. We require that $F = T(F, w, 0)$. Suppose that $\hat{F}$ is an efficient forecast and Ψ is the relevant information set. The transformation T can produce 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 $\mathsf{E}S(\hat{F}_\delta, y)$ is differentiable as a function of δ, then we must have

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

Under appropriate regularity conditions the differentiation and expectation operations are interchangeable and we obtain the following moment conditions:

$$\mathsf{E}\frac{d}{d\delta}S(\hat{F}_\delta, y)\Big|_{\delta=0} = 0. \tag{15}$$

By the same logic assuming that S is proper we must have for an arbitrary non-random CDF F

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

where

$$F_\delta = T(F, w, \delta),$$

since the maximum of $\int_{-\infty}^{\infty} S(F_\delta, t)dF(t) = 0$ is achieved at $\delta = 0$.

It can be seen that efficiency conditions (15) can be considered as auto-calibration conditions of general type (2) with

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

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):

$$F_\delta(y) = F(y - w'\delta). \tag{16}$$

For example, consider a density forecast with log-density

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

and the logarithmic scoring rule

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

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 $c = \text{mean}(F)$:

$$F_\delta(y) = F((y - c(F)) \exp(-w'\delta) + c(F)). \tag{17}$$

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 give 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 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 reasonable model modification. If an efficiency test diagnoses miscalibration, we could estimate the corresponding parameters by maximizing the average score and replace the rejected forecast by the recalibrated one.

3.3 Calibration, efficiency and sharpness

Another link between calibration and efficiency is provided by the sharpness principle of probabilistic forecasting.

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). Users can 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 choice between forecasts solely on the basis of their sharpness.

In Gneiting, Balabdaoui, and Raftery (2007) a conjecture called "the sharpness principle" was put forward, which states that the problem of finding a good forecast can be viewed as the problem of maximizing sharpness subject to calibration. It can be shown that the conjecture is actually true provided that a vague "calibration" notion is replaced by (conditional or unconditional) auto-calibration.[16]

First, for a proper scoring rule $S(F,F)$ can be viewed as a measure of sharpness of a distribution F. For a proper scoring rule $-S(F,F)$ is a concave[17] function of F and thus, according to DeGroot (1962), can

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

[17] Function $S(F_1, F_2)$ is linear in the second argument. Therefore $S(F_\alpha, F_\alpha) = \alpha S(F_\alpha, F_1) + (1-\alpha) S(F_\alpha, F_2) \le \alpha S(F_1, F_1) + (1-\alpha) S(F_2, F_2)$ for $F_\alpha = \alpha F_1 + (1-\alpha) F_2$ and $\alpha \in [0,1]$.

be viewed as a measure of uncertainty of a probability distribution with CDF F; see also Bröcker (2009). For the logarithmic scoring rule $-S(F,F)$ is the familiar Shannon's entropy measure.

Second, for a forecast which is auto-calibrated we have $\mathsf{E}S(\hat{F},y) = \mathsf{E}S(\hat{F},\hat{F})$, i. e. the expected score of such a forecast equals its expected sharpness. The fact follows from (2) for $g = S(F,y)$. If, as suggested above, the score can depend on an additional Ψ-measurable variable w, then we have to replace unconditional auto-calibration by auto-calibration with respect to w or Ψ to obtain $\mathsf{E}S(\hat{F},y;w) = \mathsf{E}S(\hat{F},\hat{F};w)$.

It follows that auto-calibrated forecasts can be compared on the basis of the levels of their expected sharpness. Sharpness is no more a deceptive characteristic when only auto-calibrated forecasts are considered. 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.

Another intuitively expected property of well-calibrated forecasts is that the more complete information has the forecaster, the sharper is the forecast which he can potentially produce. Let $\mathring{F}_1 = \mathbb{F}(\cdot|\Psi_1)$ be the ideal forecast based on Ψ_1 and $\mathring{F}_2 = \mathbb{F}(\cdot|\Psi_2)$ the ideal forecast based on Ψ_2, where Ψ_1 is a "richer" information set than Ψ_2 ($\Psi_2 \subseteq \Psi_1$). Then

$$\mathsf{E}S(\mathring{F}_1,y) = \mathsf{E}S(\mathring{F}_1,\mathring{F}_1) \geq \mathsf{E}S(\mathring{F}_2,y) = \mathsf{E}S(\mathring{F}_2,\mathring{F}_2)$$

with strict inequality if $\mathring{F}_1 \neq \mathring{F}_2$ almost surely and S is strictly proper. See Holzmann and Eulert (2014) for a proof. Similar results for the discrete outcome case can be found in DeGroot and Fienberg (1983) and Bröcker (2009). A corollary is that a forecast, which is auto-calibrated given Ψ, can never be less efficient than the ideally calibrated forecast given Ψ.

We can further study the relationship between the expected score and the expected sharpness for forecasts lacking conditional auto-calibration. Let d denote a divergence indicator (generalized distance) between two non-random CDFs F_1 and F_2 defined as

$$d(F_2,F_1) = S(F_1,F_1) - S(F_2,F_1).$$

The divergence $d(F_2,F_1)$ is non-negative, if the rule S is proper. It is zero, when the two distributions coincide. For the logarithmic scoring rule d is the Kullback–Leibler distance. Let $\mathring{F} = \mathbb{F}(\cdot|\Psi,\hat{F})$ be the conditional CDF of y given Ψ and $\hat{F}$, which can be considered as a "fully recalibrated" version of forecast $\hat{F}$ given Ψ. Since $\mathsf{E}S(\hat{F},y) = \mathsf{E}S(\hat{F},\mathring{F})$, the expected score of a (possibly miscalibrated) forecast $\hat{F}$ can in

general be decomposed as follows:

$$\mathsf{E}S(\hat{F}, y) = \mathsf{E}S(\mathring{F}, \hat{F}) - \mathsf{E}d(\hat{F}, \mathring{F}). \tag{18}$$

The first term can be interpreted as the expected sharpness of the fully recalibrated version of $\hat{F}$, while the second term relates to the divergence between $\hat{F}$ and $\mathring{F}$, i. e. it is a measure of miscalibration of forecast $\hat{F}$ with respect to the information contained in itself and Ψ. An unconditional version of this partitioning for the dichotomous outcomes and the Brier score was developed in Sanders (1963). Bröcker (2009) extended it to the case of an arbitrary finite-support discrete distribution and arbitrary proper scoring rules.

The principle of maximizing sharpness subject to calibration which was considered here is difficult to apply in practice, because achieving perfect (unconditional or conditional) auto-calibration of a forecast may prove too challenging. However, this principle provides a useful insight into the essence of probabilistic forecasting. In particular, it is clear that the advantage of using proper scoring rules for forecast comparison is that they provide the right balance of sharpness and calibration. If other—not proper—scoring rules were used for forecast evaluation, then the forecaster would have an incentive to report miscalibrated (for example, too sharp) forecasts.

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 above, the expected sharpness of an auto-calibrated forecast equals its expected score. In general if $\hat{F}$ is auto-calibrated given Ψ, then from (2) with $g = S(F, y)a(w, F)$ we have that for any function $a(w, F)$ and any Ψ-measurable w

$$\mathsf{E}[(S(\hat{F}, y) - S(\hat{F}, \hat{F}))a(w, \hat{F})] = 0. \tag{19}$$

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, from (2) with $g = S(F_2, y)b(w, F_1, F_2)$ we have for any function b and any Ψ-measurable w that

$$\mathsf{E}[(S(\hat{F}_2, y) - S(\hat{F}_2, \hat{F}_1))b(w, \hat{F}_1, \hat{F}_2)] = 0. \tag{20}$$

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, evaluation of the Swedish Riksbank's inflation forecasts

Our first example illustrates the 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 were already used 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 using the forecasts in the unmodified form. Additional details about the Riksbank's forecasts can be found in Appendix C.

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). These scoring rules modify the ordinary logarithmic score by a weighting function $\gamma(y)$ $(0 \le \gamma(y) \le 1)$, which is chosen in such a way, that it 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

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

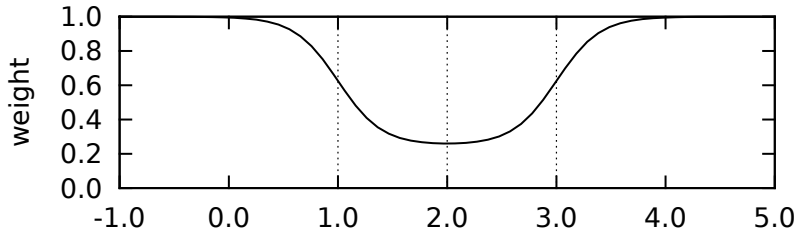


Figure 3: The weighting function $\gamma(y)$ for evaluation of the Riksbank's inflation forecasts.

The formulas for the location and scale moment function 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 3):

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

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 (16) with $w = 1$. It gives $\chi^2_1 = 1.55$ with a p-value of 21%.

Location test 2 The next 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 (16) 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 (17) 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 (17) with $c = \mu$ and $w = (1, \ln(\hat{\sigma}^2))$, where $\hat{\sigma}^2$ is the variance of the predictive distribution $\hat{F}$. It gives $\chi^2_1 = 1.89$ 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.

4.2 Example 2, pitfalls of uniform and independent PIT values

Consider the following artificial example, which highlights problems with the uniformity and independence condition. Starting from $y_0 \sim N(0,1)$ define y_t for $t \geq 1$ recursively:

$$y_t = \mu_t + \epsilon_t,$$
$$\mu_t = \Phi^{-1}(\{K\Phi(y_{t-1})\})\sqrt{1-\lambda}.$$

where $\Phi(\cdot)$ is the standard normal CDF, $\{\cdot\}$ is the fractional part, K is a large integer, $\epsilon_t \sim N(0,\lambda)$ is an independent Gaussian white noise series, and $\lambda \in (0,1)$.

We assume that the relevant information set for the forecast at time $t \geq 1$ is $\Psi_t = \sigma(y_0, y_1, \ldots, y_{t-1})$. Three different forecasts are considered.

Forecast A: $N(\mu_t, \lambda)$. This forecast reflects the data-generating process and is ideally calibrated with respect to Ψ_t.

Forecast B: $N(0,1)$. This forecast corresponds to the unconditional distribution of y_t and is unconditionally auto-calibrated. PIT values of this forecast are dependent. However, the data-generating process incorporates a highly non-linear transformation, which disguises the dependence. For large K it is impossible to find any traces of serial dependence in the series of the PIT values for Forecast B by the means of PIT-based tests ordinarily used for forecast evaluation.

Forecast C: $N(\mu_t + \xi_t, \lambda + \beta)$, where ξ_t is an independent Gaussian white noise $\xi_t \sim N(0,\beta)$. Forecast C is based on Forecast A, but contains additional noise. It has PIT values $\Phi((\epsilon_t - \xi_t)/\sqrt{\lambda+\beta})$, which are distributed as $U[0,1]$ and independent. Moreover, PIT values are distributed as $U[0,1]$ conditionally on Ψ_t (that is, probabilistically calibrated given Ψ_t). However, Forecast C is not auto-calibrated with respect to Ψ_t, since it is not marginally calibrated given Ψ_t.

Here we have two forecasts with uniform and independent PIT values and one forecast with uniform PIT values and a non-obvious dependence in PIT values. We can run a battery of tests for PIT uniformity

and independence such as those listed in Chen (2011) and Mitchell and Wallis (2011). However, the result of such exercise is foreseeable so we skip it.

The example is artificial and is not directly related to real forecasting problems, but it is suggestive. Although from the point of view of the usual tests of uniformity and independence all the three forecasts look indistinguishably perfect, they are different in terms of efficiency.

For example, the expected logarithmic score of a single forecast of y_t is given by $\alpha - \log(\lambda)/2$ for Forecast A, α for Forecast B and $\alpha - \log(\lambda+\beta)/2$ for Forecast C, where $\alpha = -(\log(2\pi)+1)/2$. Thus, forecast B is dramatically worse than A for $\lambda \ll 1$ and C is dramatically worse than A for $\beta \gg \lambda$. This observation allows to make three important conclusions.

First, this example in general shows that the uniformity and independence condition is unreliable. A forecast, which seems perfect when evaluated by uniformity and independence tests can in fact be very poor in terms of efficiency.

Second, Forecasts B demonstrates that it can be hard to detect miscalibration when there is a non-obvious non-linearity. Theoretically one could devise powerful calibration tests for Forecast B, but this would require understanding of the nonlinear structure of the data-generating process.

Third, Forecasts C demonstrates that the concept of conditional probabilistic calibration is unreliable when the forecast can include extraneous noise.

4.3 Example 3, combined forecasts, simulation

Our second example relates to calibration testing of combined forecasts. 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. The first is the equal-weight linear pool[18] of two partial conditional CDFs: $\hat{F}_c = \frac{1}{2}\hat{F}_x + \frac{1}{2}\hat{F}_z$. The second and third forecasts are recalibrated versions of $\hat{F}_c$. The recalibration (improving in order to achieve better calibration) is implemented via an INT-based model:

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

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

Table 1: 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
Expected sharpness	−1.761	−1.616	−1.517	−1.485
Expected miscalibration	0.127	0.111	0.040	0
% best	0	0	1.41	98.59
Test 1, $\mathrm{INT} \perp 1, x$	99.68	99.92	4.30	5.11
Test 2, $y - \hat{y} \perp 1, x$	99.71	99.88	4.99	5.22
Test 3, $\mathrm{INT}^2 - 1 \perp 1$	71.35	5.07	4.59	5.06
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. Expected sharpness is $\mathsf{E}S(\hat{F}, \hat{F})$. Expected miscalibration is $\mathsf{E}d(\hat{F}, \hat{F}_{xz})$ from the decomposition (18). 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.

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 conditionally auto-calibrated with respect to $\sigma(x)$ and can be regarded as a perfectly recalibrated variant of $\hat{F}_c$ (note that $\sigma(\hat{F}_c) = \sigma(x, z)$.).

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

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

Test 3 $\mathsf{E}[\mathrm{INT}^2 - 1] = 0$.

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

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 $S_x = S(\hat{F}_x, y)$ and $\hat{S}_x = S(\hat{F}_x, \hat{F})$.

Test 1 is a conditional location test based on INT values. Test 2 is a conditional location test based directly on the target variable. Test 3 is an unconditional scale test based on INT values. 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 (see conditions (20)). Tests 1 and 3 relate to probabilistic calibration given x, Test 2 relates to marginal calibration given x, while Tests 4 and 5 are more general.

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.[19] Table 1 shows the results on the expected logarithmic score, expected sharpness, expected miscalibration, comparison of average logarithmic scores and rejection rates for 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 model had the highest average logarithmic score. It can be seen that better recalibration increases sharpness and reduces miscalibration.

The basic combined forecast $\hat{F}_c$ is underdispersed 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 3, 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. 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. Even though z is unobservable to the examiner directly, he can use the characteristics of the forecast itself to detect miscalibration. 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.

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

Thus, the concept of conditional auto-calibration can be important in practice, not only in theory.

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 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.

The paper highlights the difference between conditional auto-calibration and the less general concepts of conditional probabilistic calibration (PIT uniformity) and conditional marginal calibration. Importantly, our results suggests that a great caution is required when using the condition of uniformity and independence of PIT values as a definition of ideal calibration or efficiency. The moment conditions described here can be used to extend forecast evaluation techniques to situations where this uniformity and independence condition is not necessary or inappropriate.

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 variables. This observation is closely related to the problem of choosing variables and functions for calibration testing. There is no guarantee that necessary and sufficient conditions of auto-calibration (8), (9) or (10) can provide 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. (Forecast B in Example 2 is an illustration to this problem.)

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. In general 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, less specific tests, like tests based on predicted efficiency conditions, 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

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.

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. (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.

DeGroot, M. H., and S. E. Fienberg (1983): "The Comparison and Evaluation of Forecasters," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2), 12–22.

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 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.

Pesaran, M. H., and S. Skouras (2002): "Decision-Based Methods for Forecast Evaluation," in *A Companion to Economic Forecasting*, ed. by M. P. Clements, and D. F. Hendry, chap. 11, pp. 241–267. Blackwell.

Sanders, F. (1963): "On Subjective Probability Forecasting," *Journal of Applied Meteorology*, 2, 191–201.

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 6. *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 7. *If a forecast $\hat{F}$ is ideally calibrated with respect to Ψ^*, then it is conditionally auto-calibrated with respect to Ψ for any $\Psi \subseteq \Psi^*$.*

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 8. *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 9. *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

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

□

Example 10. The actual distribution of y is described by its conditional distribution given w: $\mathbb{F}(q|w) =$

$F^\circ_w(q)$, where $w = 1$ or $w = 2$ with equal probabilities and

$$F^\circ_1(q) = \begin{cases} \frac{3}{2}q, & q \in [0, \frac{1}{4}], \\ \frac{1}{2}q + \frac{1}{4}, & q \in [\frac{1}{4}, \frac{3}{4}], \\ \frac{3}{2}q - \frac{1}{2}, & q \in [\frac{3}{4}, 1], \end{cases} \qquad F^\circ_2(q) = \begin{cases} \frac{1}{2}q, & q \in [0, \frac{1}{4}], \\ \frac{3}{2}q - \frac{1}{4}, & q \in [\frac{1}{4}, \frac{3}{4}], \\ \frac{1}{2}q + \frac{1}{2}, & q \in [\frac{3}{4}, 1]. \end{cases}$$

The 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 $\frac{1}{2}F^\circ_1(q) + \frac{1}{2}F^\circ_2(q) = \frac{1}{2}F_1(q) + \frac{1}{2}F_2(q)$ ($= q$ for $q \in [0,1]$) and $\frac{1}{2}F^\circ_1(F_1^{-1}(\alpha)) + \frac{1}{2}F^\circ_2(F_2^{-1}(\alpha)) = \alpha$, 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 $\hat{F} = F^\circ_w$ which is not the case here.

Theorem 11. *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 12. *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 $\hat{F}(q) = \mathbb{F}(q|\Psi, \hat{F})$. □

Theorem 13. *Suppose that in a recursive one-step density forecasting situation each forecast $\hat{F}_t$, $t = 1, \ldots, T$ is auto-calibrated with respect to $\Psi_t = \sigma(y_1, \ldots, y_{t-1}, \hat{F}_{1,}, \ldots, \hat{F}_{t-1})$.*[20] *Then $(p_1, \ldots, p_T) \sim U[0,1]^T$, where $p_t = \hat{F}_t(y_t)$.*

Proof. By Theorem 8 we have $p_t|\Psi_t \sim U[0,1]$. Since $\sigma(p_1, \ldots, p_{t-1}) \subseteq \Psi_t$, it follows that $p_t|p_1, \ldots, p_{t-1} \sim U[0,1]$. Using this fact and starting induction from $p_1 \sim U[0,1]$ we obtain $(p_1, \ldots, p_T) \sim U[0,1]^T$. □

Theorem 14. *Suppose that in a recursive one-step density forecasting situation each forecast $\hat{F}_t$, $t = 1, \ldots, T$ is Ψ_t-measurable, where $\Psi_t = \sigma(y_1, \ldots, y_{t-1})$, and that $(p_1, \ldots, p_T) \sim U[0,1]^T$, where $p_t = \hat{F}_t(y_t)$. Then each forecast $\hat{F}_t$ is ideally calibrated with respect to Ψ_t.*

Proof. From $(p_1, \ldots, p_T) \sim U[0,1]^T$ it follows that $p_t|p_1, \ldots, p_{t-1} \sim U[0,1]$. In a recursive one-step density forecasting situation if each $\hat{F}_t$ is Ψ_t-measurable we have by inductive application of Theorem 6

$$\sigma(p_1, \ldots, p_{t-1}) = \sigma(y_1, \ldots, y_{t-1}) = \Psi_t$$

and thus $p_t|\Psi_t \sim U[0,1]$. By Theorem 11 this means that $\hat{F}_t$ is ideally calibrated with respect to Ψ_t. □

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

The GCsL scoring rule is defined as

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

[20] Here Ψ_1 is assumed to be the trivial σ-algebra.

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 (16) 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

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

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

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

we obtain

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

Similarly a scale efficiency test corresponding to the transformation (17) 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 19.

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`.