

Оценка качества вероятностных прогнозов: корректные скоринговые правила и моменты

В статье дается обзор вероятностного прогнозирования и обсуждается теоретический подход к оценке качества плотностных прогнозов, основанный на корректных скоринговых правилах и моментах. Данный подход опробован на условном примере прогнозирования в модели авторегрессии второго порядка, а также на примере прогнозирования фондового индекса РТС.

Ключевые слова: вероятностный прогноз; калибровка прогноза; вероятностное интегральное преобразование; скоринговое правило.

JEL classification: C53; C18.

1. Введение

В последние годы все больший акцент в литературе по прогнозированию делается на различные вероятностные прогнозы. Первоначальный толчок к развитию вероятностного прогнозирования дали исследования в области метеорологии. Привычный детерминистический или категорический прогноз (в форме «завтра будет дождь») имеет много недостатков, см., например, (Murphy, Winkler, 1984; Дмитриев, 2000). Желательно, чтобы прогнозист сообщал не только некоторый правдоподобный уровень прогнозируемого показателя (т.е. точечный прогноз), но и неопределенность, которая его сопровождает. Важность этого в метеорологии объясняется тем, что знание неопределенности помогает пользователям прогноза принимать более эффективные и информированные решения, продумывать все возможные варианты развития событий (Гилл, 2008). Это касается как простейших повседневных дел (например, как сегодня одеться), так и стихийных бедствий с катастрофическими последствиями. «Использование вероятностей является обычным способом выражения неопределенности и широко распространено на практике. Важно то, что вероятности основаны на объективных научных методах, что позволяет им быть надежными, заслуживающими доверия и хорошо откалиброванными по отношению к истинному вероятностному распределению изучаемых явлений» (Гилл, 2008). Подробный обзор развития вероятностного прогнозирования в метеорологии можно найти в статье (Murphy, Winkler, 1984).

Очевидно, что указанные соображения о том, что пользователям бывает важно владеть информацией о степени неопределенности прогноза и вероятностях различных вариантов развития событий, в полной мере относятся и к прогнозированию экономических показателей. Именно поэтому вероятностные прогнозы приобретают все большую популярность среди экономистов.

Пожалуй, самым известным плотностным прогнозом (т.е. в виде функции плотности) является прогноз инфляции Банка Англии (Britton et al., 1998; Clements, 2004). Такие прог-

нозы публикуются Банком Англии начиная с 1996 года. Прогнозы ежеквартальные, с горизонтом от одного квартала до двух лет. Плотностной прогноз имеет вид двухчастного нормального распределения, «склеенного» из двух половин обычного нормального распределения; подобный вид прогнозного распределения позволяет отразить асимметрию. Графически прогнозы Банка Англии публикуются в виде так называемых *веерных диаграмм* (*fan charts*), составленных из интервальных прогнозов. Подробное описание соответствующих процедур можно найти в (Britton et al., 1998).

Другой известный вероятностный прогноз — это «Опрос профессиональных прогнозистов» (SPF), проводимый Федеральным резервным банком Филадельфии (Engelberg et al., 2009; Diebold et al., 1999b). В рамках опроса эксперты в числе прочего оценивают и вероятности того, что прогнозируемый макроэкономический показатель попадет в указанные интервалы. Эта процедура не позволяет экспертам полностью указать функцию распределения и приводит к дискретизации прогнозов, т. е. прогнозы имеют вид гистограмм.

Практика использования точечных прогнозов без указания вероятностей и вида функции распределения пока доминирует, но постепенно приходит осознание того, что интерпретация и использование точечных прогнозов могут быть связаны с большими практическими сложностями. Особенно много проблем возникает в том случае, если перед прогнозистом не ставится конкретная задача, т. е. не указывается, на какую именно характеристику прогнозного распределения следует ориентироваться (математическое ожидание, моду, медиану) или, альтернативно, по какой скоринговой функции будет оцениваться успешность прогноза (например, основанной на абсолютной величине ошибки прогноза или квадрате ошибки). Некоторые из этих проблем обсуждаются в (Engelberg et al., 2009; Gneiting, 2011).

В данной статье рассматривается вопрос о том, каким образом можно сравнивать между собой разные вероятностные прогнозы и судить об их качестве.

Простейшую схему принятия решений на основе вероятностных прогнозов можно описать следующим образом. Пользователь прогноза выбирает некоторые действия a , последствия которых зависят от реализации y прогнозируемой случайной величины Y . Если предпочтения пользователя прогноза описываются функцией полезности $u(a, y)$, и имеется вероятностный прогноз в форме функции распределения F , то оптимальные действия $a(F)$ задаются следующим образом:

$$a(F) \in \arg \max_a Eu(Y, a) \quad \text{при } Y \sim F,$$

см., например, (Pesaran, Skouras, 2002). Здесь и далее E — оператор математического ожидания, $Y \sim F$ означает, что случайная величина Y имеет распределение F .

Можно утверждать, что прогноз F_1 лучше, чем F_2 , если он обеспечивает более высокий уровень ожидаемой полезности:

$$Eu(Y, a(F_1)) > Eu(Y, a(F_2)).$$

Это дает экономическую основу для теории оценки качества вероятностных прогнозов. Заметим, что в этой формулировке прогнозы F_1 и F_2 имеют случайный характер. Часто в теории прогнозирования вместо максимизации ожидаемой полезности рассматривается минимизация ожидаемых потерь.

2. Основные понятия, определения и свойства

2.1. Скоринговые правила

Под *скоринговым правилом* (*scoring rule*) понимается функция $S(F, y)$, зависящая от функции распределения F и значения прогнозируемой величины y . Предполагается, что согласно этому правилу начисляются условные баллы, отражающие качество или успешность прогноза. Если $F_1, \dots, F_N$ — серия реализаций прогнозных функций распределения, а $y_1, \dots, y_N$ — серия фактически наблюдаемых значений прогнозируемой величины, то прогноз тем успешнее, чем выше средний балл, рассчитываемый как

$$\frac{1}{N} \sum_{i=1}^N S(F_i, y_i).$$

Обозначим через $\bar{S}(F_2, F_1)$ математическое ожидание балла от прогнозного распределения F_2 по скоринговому правилу S , если прогнозируемая величина распределена как F_1 , т. е.

$$\bar{S}(F_2, F_1) = E S(F_2, Y) \text{ при } Y \sim F_1.$$

Согласно определению, скоринговое правило S является *корректным* (*proper*), если

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

и *строго корректным* (*strictly proper*), если неравенство здесь строгое при $F_2 \neq F_1$. Если прогноз оценивается по корректному скоринговому правилу, то прогнозисту невыгодно обманывать и сообщать прогнозное распределение, которое он сам считает неверным. Подробный обзор этой темы можно найти в (Gneiting, Raftery, 2007; Bröcker, Smith, 2007).

Как объясняется в работе (Tsyplakov, 2011), за основу для оценки качества прогнозов в форме функции распределения логично взять понятие корректного скорингового правила, поскольку оно тесно связано с максимизацией ожидаемой пользователем прогноза полезности. Действительно, определим скоринговое правило S как полезность результата y при действии $a(F)$:

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

Такое основанное на полезности скоринговое правило будет корректным, поскольку

$$\bar{S}(F_1, F_1) = Eu(Y, a(F_1)) \geq Eu(Y, a(F_2)) = \bar{S}(F_2, F_1) \text{ при } Y \sim F_1,$$

см., например, (Gneiting, Raftery, 2007). Поэтому, анализируя успешность вероятностных прогнозов, можно сосредоточиться на корректных скоринговых правилах и абстрагироваться от максимизации ожидаемой полезности.

Первое корректное скоринговое правило было предложено Г. Брайером для дискретных прогнозируемых величин (Brier, 1950). Пусть переменная Y может принимать k значений $(1, \dots, k)$, $\pi_j = P_F\{Y = j\}$ — вероятность, которую прогнозное распределение F приписывает событию $Y = j$, и $I\{A\}$ — индикаторная функция события (условия) A . Тогда *скоринговое правило Брайера*, называемое также квадратичным скоринговым правилом, задается формулой

$$S(F, y) = \sum_{j=1}^k (\pi_j - I\{y = j\})^2.$$

В эконометрике наибольшей популярностью пользуется логарифмическое скоринговое правило для плотностных прогнозов. Если прогнозное распределение F является абсолютно непрерывным, то $F'(y)$ — соответствующая функция плотности, и логарифмическое скоринговое правило имеет вид

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

Для дискретного распределения

$$S(F, y) = \ln \pi_y.$$

Очевидна близкая связь логарифмического скорингового правила с логарифмической функцией правдоподобия.

Не столь популярное, но довольно естественное скоринговое правило — это *правило непрерывной ранжированной вероятности* (*continuous ranked probability score, CRPS*). Его исходное определение следующее:

$$S(F, y) = \int_{-\infty}^{\infty} (F(t) - I\{t \leq y\})^2 dt.$$

По виду это квадрат расстояния в метрике L^2 между прогнозной функцией распределения и эмпирической функцией распределения для одиночного наблюдения y . Gneiting и Raftery (2007) предлагают альтернативное представление данного правила:

$$S(F, y) = E |Y - y| - \frac{1}{2} E |Y - Y'| \text{ при независимых } Y \sim F \text{ и } Y' \sim F.$$

Это скоринговое правило привлекательно тем, что его можно рассматривать как обобщение функции потерь в виде абсолютного значения ошибки, ведь такая функция потерь является популярным критерием оценки качества точечных прогнозов.

2.2. Калибровка

Безусловная калибровка

Для вероятностного прогноза важно выполнение свойства калибровки (Diebold et al., 1999a; Gneiting et al., 2007). Под калибровкой понимается соответствие между вероятностными прогнозами и фактическим поведением прогнозируемой величины. На практике далеко не всегда вероятностные прогнозы хорошо калиброваны. В частности, обычным явлением является чрезмерная уверенность (Lichtenstein et al., 1982). Например, рассмотрим ситуацию, когда указаны центральные 90%-ные прогнозные интервалы. По замыслу, для таких интервалов в 5% случаев прогнозируемая величина по факту должна оказываться ниже нижней границы интервала, в 5% случаев — выше верхней границы, и в 90% случаев — в границах интервала. На практике вполне может оказаться, что только в 50% случаев прогнозируемая величина попадает в интервал. Может также наблюдаться смещение прогнозного распределения вверх или вниз. Например, смещение вверх (в случае центрального 90%-ного интервала) может проявляться в том, что в 9% случаев прогнозируемая величина оказывается выше верхней границы и только в 1% случаев — ниже нижней границы.

Причиной таких несоответствий может быть то, что об успешности прогнозов судят не по корректным скоринговым правилам, а по каким-то другим критериям. Например, чтобы

казаться более умелым прогнозистом, человек может сообщать слишком узкий прогнозный интервал, преувеличивая степень своей уверенности. Также бывает, что прогнозисту более выгодно сообщать хорошие новости, чем плохие: скажем, при прогнозировании безработицы давать заниженные прогнозы, а при прогнозировании темпов роста ВВП — завышенные. Некоторые люди, наоборот, склонны завышать вероятность неблагоприятного развития событий; возможно, при этом они осознанно или неосознанно стремятся оказаться в ситуации, когда можно будет сказать: «Ну вот, я же говорил, что нас ждет кризис».

Даже если субъективный фактор не играет большой роли, как при прогнозировании с помощью эконометрических моделей, неправильная калибровка прогнозов все равно встречается очень часто, поскольку все модели в той или иной степени являются несовершенными. Для совершенствования используемых в прогнозировании методов и моделей требуется уметь диагностировать неправильную калибровку, чтобы затем исправлять прогнозы, делая их «более калиброванными».

Для диагностирования калибровки прогнозов непрерывно распределенной прогнозируемой величины можно использовать так называемые значения вероятностного интегрального преобразования (ВИП-значения, *probability integral transform*, *PIT-values*). Это показатель, который наиболее часто используется для диагностики калибровки эконометрических прогнозов (например (Diebold et al., 1998; Mitchell, Wallis, 2011)). Если указано прогнозное распределение F , а фактически наблюдаемое значение прогнозируемой величины равно y , то соответствующее ВИП-значение по определению равно $F(y)$. ВИП-значения $P = F(Y)$ хорошо калиброванного прогноза величины Y должны быть равномерно распределены на отрезке $[0, 1]$, т. е. $P \sim U[0, 1]$. Например, если $F(y) = \frac{1}{2}$, то y — медиана прогнозного распределения, и при хорошей калибровке вероятность того, что Y не превышает y , равна 50%. Как следствие, вероятность того, что P не превышает $F(y) = \frac{1}{2}$, тоже должна быть равна 50%. Так же и для других квантилей, т. е. вероятность события $P \leq \alpha$ должна быть равна α . В данной статье будем называть такую калибровку безусловной¹. Оценить правильность безусловной калибровки можно, в частности, с помощью гистограммы ВИП-значений на отрезке $[0, 1]$: она должна быть почти плоской (Diebold et al., 1998; Gneiting et al., 2007; Mitchell, Wallis, 2011).

Калибровка относительно информационного множества

Предыдущие рассуждения не уделяют должного внимания условному характеру калибровки и полноте использования имеющейся у прогнозиста полезной информации².

Предположим, что прогнозирование основано на некоторой имеющейся информации Ω . Тогда F зависит от Ω : $F = F(\cdot; \Omega)$. Использование условного относительно Ω анализа позволяет усилить определение калибровки. Прогноз $F = F(\cdot; \Omega)$ калиброван относительно Ω , если

$$F(y) = G(y | \Omega),$$

где $G(y | \Omega)$ — функция распределения Y , условная относительно Ω ³. Такой прогноз можно назвать *идеальным* (среди всех прогнозов, основанных на Ω).

¹ В (Gneiting et al., 2007) данный аспект калибровки назван вероятностной калибровкой.

² «Оценка качества вероятностных прогнозов по их калибровке игнорирует условный аспект» (Clements, Taylor, 2003, p. 446).

³ Понятия калибровки относительно информационного множества и автокалибровки введены в работе (Tsyplakov, 2011). Математически строгий вариант первого из этих определений был независимо предложен

Если Y имеет непрерывное распределение, признаком идеальности вероятностного прогноза F относительно Ω является то, что соответствующее ВИП-значение $P = F(Y | \Omega)$ не просто равномерно распределено на отрезке $[0, 1]$, а распределено так условно относительно Ω :

$$P | \Omega \sim U[0, 1].$$

Важное свойство идеально калиброванного прогноза состоит в том, что он характеризуется максимальным ожидаемым баллом при использовании любого корректного скорингового правила⁴:

$$E S(G, Y) \geq E S(F, Y).$$

Более того, при соответствующих дополнительных условиях неравенство здесь будет строгим, если скоринговое правило S является строго корректным. Таким образом, *если прогноз неправильно калиброван, то имеется потенциальная возможность повышения его качества, измеряемого средним баллом по корректному скоринговому правилу*. В определенном смысле в основе понятия калибровки лежит максимизация баллов по корректному скоринговому правилу.

Независимость и равномерность ВИП-значений

В случае, когда последовательно строятся одношаговые прогнозы некоторого временного ряда Y_t по полной предыстории того же ряда, калибровку обычно проверяют с помощью анализа соответствующего ряда ВИП-значений. Рассмотрим последовательность F_t вероятностных прогнозов одномерного временного ряда Y_t , основанных на его собственной предыстории $\Omega_{t-1} = \sigma(Y_1, \dots, Y_{t-1})$, $t = 1, 2, \dots$ (для $t = 1$ прогноз безусловный). Определим соответствующие ВИП-значения следующим образом:

$$P_t = F_t(Y_t), \quad t = 1, 2, \dots$$

Последовательность прогнозов является правильно калиброванной тогда и только тогда, когда ВИП-значения P_t независимы и распределены как $U[0, 1]$.

О независимости ВИП-значений можно судить, например, по автокорреляционной функции самих ВИП-значений, ОНП-значений, о которых речь пойдет ниже (в конце п. 3.2), а также по автокорреляционной функции различных функций от них, например, модулей ОНП-значений, степеней (Diebold et al., 1998). Из независимости следует отсутствие серийной корреляции.

В общем случае равномерность и независимость ВИП-значений не является признаком идеального прогноза, поскольку прогнозист может использовать и дополнительную, в том числе ненужную, информацию. Кроме того, это свойство не выполняется для прогнозов на несколько шагов вперед⁵.

в (Gneiting, Ranjan, 2011). Оно также похоже на определение эффективности интервальных прогнозов относительно информационного множества в (Christoffersen, 1998).

⁴ «...Если прогноз совпадает с истинным порождающим данные процессом, то этот прогноз будет предпочтительным для всех пользователей, каковы бы ни были их функции потерь» (Diebold et al., 1998, p. 866), см. также (Granger, Pesaran, 2000).

⁵ См. обсуждение этого в (Tsyplakov, 2011).

Автокалибровка

Промежуточный между безусловной и идеальной калибровками вариант — это автокалибровка. Прогноз F называется автокалиброванным, если

$$F(y) = G_F(y),$$

где $G_F(y) = G(y|F)$ — функция распределения Y , условная относительно F . Здесь информацией для оценки калибровки служит сам прогноз. Если прогноз автокалиброван, то он безусловно калиброван, а если прогноз идеальный, то он автокалиброван. Проще всего это свойство понять на примере прогнозирования дихотомической величины, принимающей значения 0 и 1. Среди случаев, в которых прогноз указывает вероятность π для события $Y=1$, это событие должно наблюдаться с той же самой вероятностью: $P(Y=1|\pi) = \pi^6$.

2.3. Калибровка и острота прогноза

Под остротой (*sharpness*) прогноза понимается характеристика, отражающая степень конкретности прогноза или концентрации прогнозного распределения (Gneiting et al., 2007). Один из способов наглядно представить эту характеристику — рассмотреть прогнозный интервал, соответствующий прогнозируемому распределению. Например, если эксперт говорит, что к концу года курс рубля наверняка будет в промежутке от 31 до 32 рублей за доллар, то он делает довольно острый прогноз. Если же называется промежуток от 10 до 100 рублей за доллар, то такой прогноз очень расплывчатый. С одной стороны, пользователю прогноза хотелось бы иметь как можно более острый прогноз, но с другой стороны не следует забывать о калибровке. Делая слишком острый прогноз, эксперт рискует преувеличить степень своей уверенности, о чем уже говорилось выше, что для пользователя прогноза чревато неожиданностями (зачастую не очень приятными).

В (Gneiting et al., 2007) выдвинут в качестве гипотезы тезис, что задачу поиска хорошего прогноза можно рассматривать как задачу максимизации остроты при условии калибровки. Как показано в (Tsyplakov, 2011), этот тезис оказывается верным, если под калибровкой понимать автокалибровку.

Во-первых, для корректного скорингового правила величина $\bar{S}(F, F)$ может рассматриваться как мера остроты прогноза F (поскольку для корректного скорингового правила $-\bar{S}(F, F)$ — это вогнутая функция⁷ и, согласно (DeGroot, 1962), может рассматриваться как мера неопределенности распределения вероятностей F). Так, для логарифмического скорингового правила $\bar{S}(F, F)$ будет энтропией Шеннона со знаком минус.

Во-вторых, для автокалиброванного прогноза $ES(F, Y) = E\bar{S}(F, F)$, т. е. ожидаемый балл равен его ожидаемой остроте.

⁶ «Если, например, мы прогнозируем, что вероятность того, что в следующем квартале начнет спад, равна 20%, и если среди всех случаев, когда мы делаем такой прогноз, доля таких, что спад начнется, равна 20%, и если это соответствие имеет место для всех других возможных предсказанных вероятностей, то прогнозы правильно калиброваны» (Galbraith, van Norden, 2011).

⁷ Вогнутость функции $-\bar{S}(F, F)$ следует из линейности по второму аргументу и свойства корректности.

Следовательно, автокалиброванные прогнозы можно сравнивать на основе уровня их ожидаемой остроты. Идеальный прогноз является самым острым из всех автокалиброванных прогнозов, поскольку он характеризуется наибольшим ожидаемым баллом.

Еще одно интуитивно ожидаемое свойство правильно калиброванных прогнозов состоит в том, что чем более полной информацией владеет прогнозист, тем более острый прогноз он потенциально может сделать. Допустим, Ω_1 — более «богатое» информационное множество, чем Ω_2 , т. е. Ω_1 содержит всю информацию из Ω_2 и, возможно, еще какую-то полезную информацию. Пусть $G_1 = G(\cdot | \Omega_1)$ — идеальный прогноз на основе Ω_1 , а $G_2 = G(\cdot | \Omega_2)$ — идеальный прогноз на основе Ω_2 . Тогда

$$E S(G_1, Y) = E \bar{S}(G_1, G_1) \geq E S(G_2, Y) = E \bar{S}(G_2, G_2).$$

Если прогноз не автокалиброван, то его острота может быть обманчива. В общем случае имеет место следующее разложение ожидаемого балла для прогноза F (возможно, неправильно калиброванного):

$$E S(F, Y) = E \bar{S}(G_F, G_F) - E d(F, G_F).$$

Первое слагаемое интерпретируется как ожидаемая острота прогноза $G_F = G(\cdot | F)$, который как бы является «перекалиброванным» вариантом прогноза F , а второе слагаемое связано с показателем расхождения между F и G_F , т. е. является измерителем неправильной калибровки прогноза F относительно информации, содержащейся в нем самом. Здесь через d обозначен показатель расхождения (обобщенное расстояние) между распределениями F_1 и F_2 , определяемый как

$$d(F_2, F_1) = \bar{S}(F_1, F_1) - \bar{S}(F_2, F_1).$$

Если правило S корректное, то соответствующий показатель расхождения $d(F_2, F_1)$ неотрицателен, и равен нулю, когда два распределения совпадают. Для логарифмического скорингового правила d является расстоянием Кульбака–Лейблера.

Рассмотренный принцип максимизации остроты при условии калибровки трудно применять на практике, потому что достижение идеальной автокалибровки прогноза может оказаться слишком сложной задачей. Однако данный принцип способствует пониманию сущности вероятностного прогнозирования. В частности, становится понятным, что преимущество использования корректных скоринговых правил состоит в том, что они обеспечивают правильный баланс остроты и калибровки при сравнении прогнозов. Использование для оценки прогнозов других скоринговых правил, не отвечающих этому требованию, может подтолкнуть прогнозиста к предоставлению неправильно калиброванных, например слишком острых, прогнозов.

3. Тестирование калибровки на основе моментов

3.1. Общая идея тестирования калибровки на основе моментов

На практике удобно выразить условия калибровки в виде некоторых условий на моменты. Теоретические моменты можно заменить выборочными, основанными на серии прогнозов, и посмотреть, насколько результат далек от того, что должно быть в теории. Предположим, теория гласит, что при правильной калибровке математическое ожидание величины

g должно быть нулевым: $Eg = 0$. Рассчитаем эту величину для серии реализаций прогнозных функций $F_1, \dots, F_N$ и серии фактически наблюдаемых значений прогнозируемой величины $y_1, \dots, y_N$ и найдем соответствующий выборочный момент

$$\bar{g} = \frac{1}{N} \sum_{i=1}^N g_i.$$

Если $\bar{g}$ сильно отличается от нуля, то можно сделать вывод, что прогноз плохо калиброван.

Для тестирования условий на моменты можно использовать обычные t -отношения вида

$$t = \frac{\bar{g}}{se(\bar{g})}.$$

Самый тонкий аспект здесь — корректное вычисление стандартной ошибки $se(\bar{g})$. Если все сделано правильно, и прогноз калиброван, то асимптотически такая статистика распределена как $N(0, 1)$. Ниже в примерах использованы устойчивые к гетероскедастичности и автокорреляции (*heteroskedasticity and autocorrelation consistent, HAC*) стандартные ошибки.

3.2. Безусловная калибровка

Когда Y имеет непрерывное распределение, ВИП-значения безусловно калиброванного прогноза F равномерно распределены на отрезке $[0, 1]$: $F(Y) \sim U[0, 1]$. Проверять безусловную калибровку можно, сравнивая выборочные моменты для ВИП-значений с соответствующими (безусловными) моментами распределения $U[0, 1]$. Рассмотрим функцию $k = k(p)$, зависящую от вероятности $p \in [0, 1]$, и определим

$$\kappa = E k(P) \quad \text{при } P \sim U[0, 1].$$

Если прогноз F — безусловно калиброванный, то

$$E k(F(Y)) - \kappa = 0.$$

В частности, в этих обозначениях можно записать условие, что вероятность события $F(Y) \leq \alpha$ равна α , о чем говорилось выше при обсуждении безусловной калибровки. Для этого следует взять

$$k = I\{p \leq \alpha\}, \quad \kappa = \alpha. \quad (1)$$

Ранее также обсуждалась калибровка интервальных прогнозов. Рассмотрим центральный прогнозный интервал уровня β . Если F — прогнозная функция распределения, то интервал будет иметь вид

$$[F^{-1}(1/2 - \beta/2), F^{-1}(1/2 + \beta/2)].$$

При безусловной калибровке вероятность попадания Y в этот интервал равна β . Попадание Y в данный интервал равносильно попаданию ВИП-значения $F(Y)$ в интервал $I_\beta = [0.5 - 0.5\beta, 0.5 + 0.5\beta]$. Поэтому здесь можно взять

$$k = I\{p \in I_\beta\}, \quad \kappa = \beta. \quad (2)$$

Если $F(Y)$ имеет распределение $U[0, 1]$, то обратное нормальное преобразование (ОНП) этой величины будет распределено в соответствии со стандартным нормальным распределением:

$$\Phi^{-1}(F(Y)) \sim N(0, 1),$$

где $\Phi^{-1}(\cdot)$ — функция, обратная к стандартной нормальной функции распределения. Вместо моментов ВИП-значений можно использовать моменты ОНП-значений. Например, ОНП-значения должны иметь нулевое среднее:

$$k = \Phi^{-1}(F(Y)), \quad \kappa = 0.$$

3.3. Автокалибровка

Из теории точечного прогнозирования (см., например, обзор (Цыплаков, 2006)) известно, что оптимальным в среднеквадратическом смысле точечным прогнозом является математическое ожидание, условное относительно информационного множества Ω , на основе которого делается прогноз. Для такого прогноза выполнено условие ортогональности: ошибка прогноза не коррелирует с любой величиной, полученной на основе Ω . В статье (Mitchell, Wallis, 2011) была выдвинута идея о том, что контролировать калибровку вероятностных прогнозов тоже можно за счет проверки аналогичных условий ортогональности. Как показано в (Tsyplakov, 2011), эта идея поддается дальнейшему обобщению, что позволяет разрабатывать различные виды диагностических тестов для калибровки прогноза. Многие тесты и критерии калибровки/эффективности, предложенные в литературе, подпадают под этот подход.

Пусть k и κ имеют тот же смысл, что и выше, и $l(F)$ — некоторая функция от функции распределения F . Если прогноз F является автокалиброванным, то имеет место условное ограничение на моменты

$$E[k(F(Y)) - \kappa | F] = 0,$$

откуда следует условие ортогональности

$$E([k(F(Y)) - \kappa]l(F)) = 0 \quad (3)$$

для любой функции l . Это означает, что любая функция k от ВИП-значения не коррелирует с любой функцией l от прогнозной функции распределения. В качестве l можно взять какую-нибудь характеристику прогнозного распределения, такую как математическое ожидание или медиану.

Для проверки автокалибровки прогноза можно использовать и более общие ограничения на моменты. Для функции $r = r(y, F)$, зависящей от исхода y и функции распределения F , определим

$$\rho = \rho(F) = E r(Y, F) \quad \text{при } Y \sim F.$$

Автокалибровка вероятностного прогноза F эквивалентна условию

$$E[r(Y, F) - \rho(F)] = 0$$

для любой r , т. е. величина ρ должна быть несмещенным точечным прогнозом r .

Можно рассмотреть следующий пример. Если F — функция распределения случайной величины Y , то $F(y_0)$ — вероятность события $Y \leq y_0$. Выполнение аналогичного условия на моменты можно проверить на основе

$$r = I\{y \leq y_0\}, \quad \rho = F(y_0),$$

что соответствует маргинальной калибровке в (Gneiting et al., 2007).

3.4. Калибровка относительно информационного множества

Идеальная калибровка F относительно Ω является более сильным свойством и требует условной несмещенности ρ в качестве прогноза r :

$$E[r(Y, F) - \rho(F) | \Omega] = 0.$$

Эта условная несмещенность эквивалентна следующей безусловной ортогональности:

$$E([r(Y, F) - \rho(F)]q) = 0$$

для любой q , зависящей от Ω . Тем самым ρ должна быть несмещенным точечным прогнозом r , и ошибки прогноза не должны коррелировать с любой величиной q , построенной на основе доступной информации Ω .

В частности, имеем следующее условие ортогональности для ВВП-значений:

$$E([k(F(Y)) - \kappa]q) = 0.$$

Например, можно взять

$$k = I\{p \leq \alpha\}, \quad \kappa = \alpha$$

и тестировать идеальность калибровки относительно Ω , проверяя отсутствие корреляции между индикаторной функцией $I\{F(Y) \leq \alpha\}$ и любой функцией q , зависящей только от имеющейся информации Ω , для каждого $\alpha \in [0, 1]$. В (Christoffersen, 1998) аналогичное условие предложено для тестирования условного покрытия прогнозного интервала.

Чтобы обобщить этот подход, рассмотрим функцию $m = m(y, \Omega)$, зависящую от исхода y и прогнозной информации Ω . Пусть

$$\mu = \mu(\Omega) = E[m(Y, \Omega)] \quad \text{при } Y \sim F = F(\cdot; \Omega).$$

Калибровка F относительно Ω эквивалентна следующему условию на моменты:

$$E[(m(Y, \Omega) - \mu(\Omega))] = 0$$

для любого m .

Рассмотрим пример применения этого более общего типа условий. Идея заключается в тестировании калибровки одного метода прогнозирования относительно другого⁸. Предположим, требуется проверить, что $F_1 = F_1(\cdot; \Omega)$ правильно калиброван, и $F_2 = F_2(\cdot; \Omega)$ —

⁸ В этом примере можно найти параллели с тестом на равную точность прогноза, основанным на разнице баллов по логарифмическому скоринговому правилу (или на информационном критерии Кульбака–Лейблера, KLIC), см. (Amisano, Giacomini, 2007; Mitchell, Wallis, 2011).

альтернативный прогноз. Пусть $m = S(F_2, Y) - S(F_1, Y)$ для некоторого корректного скорингового правила S . В этом случае

$$\mu = \bar{S}(F_2, F_1) - \bar{S}(F_1, F_1).$$

Соответственно, для прогноза F_1 , калиброванного относительно Ω , имеем

$$E([S(F_2, Y) - S(F_1, Y)] - [\bar{S}(F_2, F_1) - \bar{S}(F_1, F_1)]) = 0. \quad (4)$$

Далее будем называть это условие на моменты *относительной калибровкой прогноза* (ОКП). Основанный на нем тест будет обладать мощностью против альтернативы, что F_2 калиброван относительно Ω , потому что тогда

$$E(m - \mu) = E[\bar{S}(F_2, F_2) - \bar{S}(F_1, F_2)] + E[\bar{S}(F_1, F_1) - \bar{S}(F_2, F_1)] > 0,$$

если скоринговое правило является строго корректным.

Пара взаимных ОКП-тестов « F_1 против F_2 » и « F_2 против F_1 » может помочь в оценке возможных выгод от объединения двух прогнозов.

4. Пример: прогнозирование процесса авторегрессии

Чтобы проиллюстрировать изложенные выше идеи, рассмотрим сначала искусственно сконструированный имитационный пример, основанный на примере из статьи (Mitchell, Wallis, 2011). Рассматривается прогнозирование процесса AR(2): $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varepsilon_t$ с независимыми гауссовскими возмущениями $\varepsilon_t \sim N(0, 1)$. В таблице 1 приведены формулы для шести конкурирующих между собой прогнозов:

- Ideal, идеальный прогноз, учитывающий всю доступную информацию, т.е. Y_{t-1} и Y_{t-2} ;
- Climt, климатологический прогноз, который не использует Y_{t-1} и Y_{t-2} , а представляет собой безусловное распределение;
- AR1, использующий только Y_{t-1} ;
- AR2, использующий только Y_{t-2} ;
- Combo, комбинированный прогноз, являющийся выпуклой комбинацией AR1 и AR2 с одинаковыми весами;
- Unfocus, нефокусированный прогноз, содержащий посторонний шум τ_t .

Фактический порождающий данные процесс представлен идеальным прогнозом. В этом примере нет необходимости оценивать параметры, поскольку они известны. Рассмотрим здесь только случай $\varphi_1 = 0.15$, $\varphi_2 = 0.2^9$. Если не указано иное, длина ряда берется равной $T = 150$.

Приводимая ниже табл. 2 основана на 5000 экспериментов. Цифры, которые указаны для тестов — это частоты отклонения для 5%-го асимптотического уровня значимости при использовании стандартных нормальных квантилей. Тестовые статистики — это t -отношения с робастными стандартными ошибками Ньюи–Уэста и лагом усечения 4. Функции $S(F_2, F_1)$ для смесей нормальных распределений, требуемые для расчета ОКП-статистик, вычислялись методом Монте-Карло по 100 экспериментам.

⁹ Этот случай рассматривался в (Mitchell, Wallis, 2011).

Таблица 1. Определения шести прогнозов для AR(2)

Ideal	$N(\varphi_1 Y_{t-1} + \varphi_2 Y_{t-2}, 1)$
Climt	$N(0, \sigma_Y^2)$
AR1	$N(\rho_1 Y_{t-1}, \sigma_1^2)$
AR2	$N(\rho_2 Y_{t-2}, \sigma_2^2)$
Combo	$0.5N(\rho_1 Y_{t-1}, \sigma_1^2) + 0.5N(\rho_2 Y_{t-2}, \sigma_2^2)$
Unfocus	$0.5N(\varphi_1 Y_{t-1} + \varphi_2 Y_{t-2}, 1) + 0.5N(\varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \tau_t, 1),$ где $\tau_t = -1$ или 1 с равными вероятностями

Примечание. $\rho_1 = \varphi_1 / (1 - \varphi_2)$, $\rho_2 = \varphi_1 \rho_1 + \varphi_2$, $\sigma_Y^2 = 1 / (1 - \varphi_1 \rho_1 - \varphi_2 \rho_2)$, $\sigma_1^2 = (1 - \rho_1^2) \sigma_Y^2$, $\sigma_2^2 = (1 - \rho_2^2) \sigma_Y^2$.

Таблица 2. Статистики для шести прогнозов AR(2)

	Ideal	Climt	AR1	AR2	Combo	Unfocus
Ожидаемый балл	-1.418	-1.456	-1.438	-1.430	-1.425	-1.529
% лучший	66.5	1.2	6.8	12.2	13.2	0.1
% лучший, 1500	98.6	0.0	0.0	0.1	1.3	0.0
Тест $z \times \text{mean}$	4.5	—	4.5	4.3	17.4	99.8
Тест ΔL , Ideal	—	45.7	32.5	25.4	16.9	90.9
Тест ОКП, Ideal	—	92.6	78.3	62.4	25.8	100.0
Тест ОКП, Climt	5.5	—	4.9	5.0	0.7	100.0
Тест ОКП, AR1	4.7	62.7	—	40.2	3.3	100.0
Тест ОКП, AR2	3.4	86.6	58.6	—	9.4	100.0
Тест ОКП, Combo	4.2	86.8	48.6	31.3	—	100.0
Тест ОКП, Unfocus	3.9	37.0	17.8	9.9	6.0	—

В таблице 2 строка «Ожидаемый балл» показывает приближенное значение математического ожидания балла по логарифмическому скоринговому правилу, полученное методом Монте-Карло. Как и следовало ожидать, идеальный прогноз имеет наивысший ожидаемый балл. Несфокусированный прогноз Unfocus демонстрирует наихудший результат, за ним идет климатологический прогноз Climt. Прогнозы AR1 и AR2 в среднем хуже своей комбинации Combo.

Ожидаемый балл показывает асимптотический потенциал прогноза, проявляющийся при стремлении количества наблюдений к бесконечности, и не может быть доступен в практической ситуации прогнозирования. Если серия прогнозов не очень длинная, то неидеальные прогнозы могут получить более высокий средний балл, чем идеальный.

Строка «% лучший» показывает долю экспериментов, в которых соответствующая модель имела наивысший средний балл по логарифмическому правилу при использовании $T = 150$ наблюдений. Идеальный прогноз был лучшим с вероятностью около 2/3. Это «умеренно большое» значение объясняется малыми значениями двух коэффициентов авторегрессии ($\varphi_1 = 0.15$, $\varphi_2 = 0.2$), что делает истинный порождающий данные процесс довольно близким к вариантам AR1, AR2 и Combo с точки зрения ожидаемого балла по логарифми-

ческому правилу. При $T = 1500$ (строка «% лучший, 1500») идеальный прогноз стабильно опережает конкурирующие прогнозы. Видно, что средний балл по корректному скоринговому правилу является неплохим критерием выбора вероятностного прогноза, и что он ведет себя предсказуемым образом.

Следует обратить особое внимание на несфокусированный прогноз Unfocus. В безусловном смысле он является калиброванным, и поэтому ВИП-значения имеют равномерное распределение. Кроме того, ряды ВИП-значений серийно независимы. Здесь не приведены результаты соответствующих тестов, см. табл. II, табл. III и рис. 2 в (Mitchell, Wallis, 2011). В то же время данный прогноз не автокалиброван, что несложно обнаружить с помощью теста на ортогональность между ОНП-значениями $z_t = \Phi^{-1}(F_t(y_t))$ и средним прогнозного распределения $\text{mean}(F_t)$. В таблице 2 он помечен как «Тест $z \times \text{mean}$ ». Для автокалиброванного прогноза имеем $E(\text{mean}(F_t)z_t) = 0$ — условие ортогональности вида (3) для автокалибровки. Таблица показывает, что данное ограничение отклонялось почти в 100% экспериментов в случае несфокусированного прогноза.

Комбинированный прогноз Combo, который представляет собой взвешенную комбинацию прогнозов AR1 и AR2, не обладает свойством безусловной калибровки (как часто бывает для составляемых подобным образом комбинированных прогнозов (Gneiting, Ranjan, 2011)) и, как следствие, не автокалиброван. Результаты теста $z \times \text{mean}$ подтверждают, что этот прогноз не автокалиброван (хотя частота отклонения нулевой гипотезы не очень велика). В то же время для других прогнозов свидетельств неправильной калибровки нет. Действительно, прогнозы Ideal, AR1 и AR2 являются автокалиброванными.

Для проверки калибровки прогнозных моделей друг относительно друга используем тест на относительную калибровку прогноза, основанный на свойстве ОКП (4). Используем логарифмическое скоринговое правило. Для увеличения мощности теста он сделан односторонним, поскольку ожидается, что тестовая статистика будет положительной в ситуациях, когда альтернативные прогнозы потенциально могут быть использованы для улучшения прогноза, калибровку которого требуется проверить. Примечательны результаты сравнения AR1 с AR2 и AR2 с AR1. Тесты часто будут указывать на полезность объединения этих двух моделей. Видно, что тест против идеального прогноза имеет высокую мощность. Можем сравнить его с тестом на равенство предсказательной силы в духе (Amisano, Giacomini, 2007), основанном на разности полученных средних баллов $\Delta L_t = S(G_t, y_t) - S(F_t, y_t)$ (он помечен как «Тест ΔL , Ideal»). Последний тест также реализован как односторонний для увеличения его мощности и для сопоставимости с ОКП-тестом. Сравнение однозначно отдает предпочтение ОКП-тесту как инструменту выявления неидеальных прогнозов.

5. Пример: прогнозирование фондового индекса

Второй пример призван показать, как изложенные выше идеи могут использоваться для оценки качества прогнозов реальных временных рядов. Данные — ежедневные значения индекса PTC (RTSI) на момент закрытия. Они охватывают период с 01.09.1995 по 13.07.2012 гг. Ряд значений индекса приводится к стационарности путем перехода к темпам прироста в процентах: $R_t = (\ln RTSI_t - \ln RTSI_{t-1}) \times 100$. Получается ряд из 4209 наблюдений. Прогнозируемая переменная — темп прироста на 10 периодов вперед. Таким образом, в мо-

мент t строится прогноз величины $y_{t+10} = R_{t+1} + \dots + R_{t+10} = (\ln RTSI_{t+10} - \ln RTSI_t) \times 100$ на основе информации о $R_1, \dots, R_t$ (горизонт в десять периодов соответствует примерно двум неделям физического времени). Рассматриваются следующие прогнозы¹⁰.

1) Исторический прогноз на основе полной истории Ист(t) использует все прошлые наблюдения темпов прироста за 10 периодов: $y_{10}, \dots, y_t$. Исторические наблюдения ресемплированы с целью получения ансамбля размера 1000.

2) Ист(200) — тоже исторический прогноз в виде ансамбля точек, но использует только сдвигающийся отрезок из 200 последних наблюдений $y_{t-199}, \dots, y_t$ и не использует ресемплирование.

3) Прогноз ЭС основан на экспоненциальном сглаживании волатильности $\sigma_{t+1}^2 = (1 - \lambda)R_t^2 + \lambda\sigma_t^2$ с множителем затухания $\lambda = 0.95$, см. (RiskMetrics, 1996). Прогнозное распределение имеет вид $N(0, 10\sigma_{t+1}^2)$. Рекуррентное вычисление волатильности начинается с выборочной дисперсии первых 200 наблюдений.

4) Прогноз GARCH основан на стандартной модели GARCH с распределением Стюдента GARCH(1,1)- t с ненулевым средним (Росси, 2010). Модель оценивается рекурсивно¹¹ методом максимального правдоподобия. Прогнозное распределение представлено ансамблем из 1000 будущих траекторий, полученных методом Монте-Карло.

Все прогнозы строятся последовательно для прогнозного периода, начинающегося с 200-го наблюдения. Они сравниваются по наблюдаемым средним баллам CRPS. Данный показатель можно рассчитать за $O(S \ln S)$ операций, необходимых для осуществления сортировки, если F_t представлено выборкой объема S (Gneiting, Raftery, 2007). Одно из преимуществ CRPS по сравнению с логарифмическим скоринговым правилом заключается в том, что его можно применить к прогнозу непрерывной случайной величины, имеющему вид дискретного распределения, что как раз и имеется в данном случае.

В таблице 3 приведены следующие статистики.

1) CRPS — средняя величина CRPS.

2) Тест α — статистика для безусловной калибровки, основанная на (1) с $\alpha = 0.5$. Данная статистика связана с местоположением, характеризуемым медианой прогнозного распределения.

3) Тест β — статистика для безусловной калибровки, отражающая покрытие центрального 50%-го интервала (2).

4) Тест ОКП, [метод] — односторонний ОКП-тест против указанного метода, такой же, как в предыдущем примере, но основанный на CRPS.

Выбор тестов на калибровку здесь достаточно произволен. Описанная выше схема позволяет разработать множество различных тестов, включая другие разновидности тестов на безусловную калибровку. Такие тесты являются формальными аналогами визуального анализа ВВП-гистограмм, метода, который часто используется для оценки качества плотностных прогнозов¹².

¹⁰ Темпы прироста индекса РТС характеризуются значимой серийной корреляцией первого порядка, но ее влияние на 10-периодном горизонте прогноза несущественно, поэтому прогнозы ее не учитывают.

¹¹ Используются все наблюдения, известные к тому моменту времени, когда производится прогноз, поэтому длина ряда с каждым шагом увеличивается на единицу.

¹² Проблема с гистограммами заключается в том, что потенциально они могут вводить в заблуждение, если не снабжены доверительной полосой, робастной к гетероскедастичности и автокорреляции.

Таблица 3. Прогнозы индекса РТС

	Ист(t)	Ист(200)	ЭС	GARCH
CRPS	-5.164	-5.168	-5.047	-5.038
Тест α	-0.006 (0.019)	-0.003 (0.019)	-0.084*** (0.020)	0.005 (0.019)
Тест β	0.131*** (0.018)	0.039** (0.017)	-0.026 (0.015)	0.005 (0.016)
Тест ОКП, Ист(t)	—	0.366*** (0.084)	0.261*** (0.054)	0.123*** (0.037)
Тест ОКП, Ист(200)	0.357*** (0.077)	—	0.324*** (0.083)	0.264*** (0.077)
Тест ОКП, ЭС	0.493*** (0.059)	0.566*** (0.096)	—	0.189*** (0.048)
Тест ОКП, GARCH	0.374*** (0.046)	0.525*** (0.093)	0.207*** (0.047)	—

Примечание. В скобках приведены робастные стандартные ошибки Ньюи–Уэста с лагом усечения 10. ***, **, * — значимость на 1, 5, 10%-ном уровне соответственно. ОКП-тесты основаны на CRPS.

Пример относится к типичной для практики ситуации, когда ни один из сравниваемых прогнозов нельзя назвать идеальным. Все прогнозы в той или иной степени неправильно калиброваны (см. табл. 3). Например, Ист(t), имея (безусловно) корректное местоположение, о чем говорит тест α , демонстрирует заметный недостаток остроты, о чем говорит тест β : фактические значения слишком редко попадают в хвосты распределения. ЭС предполагает фиксированное нулевое среднее, а это не подтверждается данными наблюдений, и поэтому тест α указывает на отрицательное смещение; заметна также скошенность гистограммы на рис. 1а.

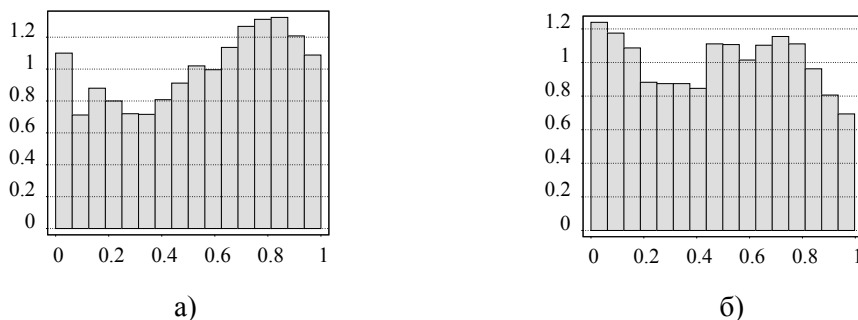


Рис. 1. Гистограммы VIP-значений: а) экспоненциальное сглаживание, б) GARCH

В целом GARCH выглядит почти как правильно калиброванный прогноз, если судить о калибровке, исходя из обычных критериев на основе VIP. Гистограмма VIP-значений на рис. 1б не является идеально ровной, но ее неровности не очень серьезны, что подтверждается двумя тестами на безусловную калибровку из табл. 3. Кроме того, нет заметных признаков автокорреляции после лага 10 как в $z_t = \Phi^{-1}(F_t(y_t))$, так и в $|z_t|$. Например, 11-й коэффициент автокорреляции равен 0.078 для z_t и -0.039 для $|z_t|$, с асимптотическими стандартными ошибками 0.044 и 0.034 соответственно¹³.

¹³ Здесь используется приближение Бартлетта для дисперсии r_{11} , которое исходит из отсутствия автокорреляции после 10-го лага и задается формулой $(1 + 2r_1^2 + \dots + 2r_{10}^2)/T$, где r_k — k -й коэффициент автокорреляции.

Модель GARCH лидирует по среднему уровню CRPS, близко за ней следует экспоненциальное сглаживание. Однако все методы в парных сравнениях по ОКП-тестам показывают значимую неправильность калибровки. Например, и это примечательно, GARCH не в состоянии объяснить показатели экспоненциального сглаживания, которое можно рассматривать как его «более дешевый» заменитель. Результаты указывают на возможность усовершенствования прогнозов путем их объединения.

6. Заключение

Вероятностные прогнозы предоставляют гораздо больше информации для принятия экономических решений, чем обычные точечные прогнозы, поэтому они имеют хорошие перспективы. Следует более широко внедрять вероятностные прогнозы в практику экономического прогнозирования. При использовании вероятностного прогнозирования желательно опираться на основные понятия и теоретические свойства. Некоторые из таких понятий и свойств рассмотрены в данной статье.

С помощью приведенных примеров:

- подтверждены известные из литературы полезные характеристики корректных скоринговых правил;
- проведена проверка работоспособности подхода к тестированию калибровки вероятностных прогнозов на основе теоретических ограничений на моменты;
- проверена работоспособность введенного автором понятия автокалибровки, продемонстрирована важность проверки этого свойства в случае, когда прогноз может содержать посторонний шум;
- проведена проверка разработанного автором теста на относительную калибровку прогноза, обнаружена высокая мощность этого теста при взаимной проверке калибровки пары вероятностных прогнозов.

Список литературы

- Гилл Дж. (2008). Уведомление поставщиков услуг о неопределенностях в прогнозах. *Бюллетень Всемирной метеорологической организации*, 57 (4), 237–243.
- Дмитриев В. Г. (2000). О возможности представления категорических прогнозов в вероятностной форме. *Метеорология и гидрология*, 11, 31–38.
- Росси Э. (2010). Одномерные GARCH-модели: обзор. *Квантиль*, 8, 1–67.
- Цыплаков А. (2006). Введение в прогнозирование в классических моделях временных рядов. *Квантиль*, 1, 3–19.
- Amisano G., Giacomini R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, 25 (2), 177–190.
- Brier G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1), 1–3.
- Britton E., Fisher P., Whitley P. (1998). The inflation report projections: Understanding the fan chart. *Bank of England Quarterly Bulletin*, 30–37.

Bröcker J., Smith L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22, 382–388.

Christoffersen P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39 (4), 841–862.

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

Clements M. P., Taylor N. (2003). Evaluating interval forecasts of high-frequency financial data. *Journal of Applied Econometrics*, 18 (4), 445–456.

DeGroot M. H. (1962). Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33 (2), 404–419.

Diebold F. X., Gunther T. A., Tay A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39 (4), 863–883.

Diebold F. X., Hahn J., Tay A. S. (1999a). 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., Tay A. S., Wallis K. F. (1999b). Evaluating density forecasts of inflation: The survey of professional forecasters. In: Engle R. F., White H. (eds.). *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*. Oxford: Oxford University Press, 76–90.

Engelberg J., Manski C. F., Williams J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, 27 (1), 30–41.

Galbraith J. W., van Norden S. (2011). Kernel-based calibration diagnostics for recession and inflation probability forecasts. *International Journal of Forecasting*, 27 (4), 1041–1057.

Gneiting T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106 (494), 746–762.

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

Gneiting T., Raftery A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Gneiting T., Ranjan R. (2011). Combining predictive distributions. *ArXiv preprint*, arXiv: 1106.1638v1 [math. ST].

Granger C. W. J., Pesaran M. H. (2000). A decision-theoretic approach to forecast evaluation. In: Chan W.-S., Li W. K., Tong H. (eds.). *Statistics and Finance: An Interface*. Imperial College Press.

Lichtenstein S., Fischhoff B., Phillips L. D. (1982). Calibration of probabilities: The state of the art to 1980. In: Kahneman D., Slovic P., Tversky A. (eds.). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press, 306–334.

Mitchell J., Wallis K. F. (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, 26 (6), 1023–1040.

Murphy A. H., Winkler R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79 (387), 489–500.

Pesaran M. H., Skouras S. (2002). Decision-based methods for forecast evaluation. In: Clements M. P., Hendry D. F. (eds.). *A Companion to Economic Forecasting*. Blackwell Publishing, 241–267.

RiskMetrics (1996). RiskMetrics (TM) — technical document (4th ed.). J. P. Morgan / Reuters.

Tsyplakov A. (2011). Evaluating density forecasts: A comment. University Library of Munich, Germany. <http://mpira.ub.uni-muenchen.de/32728/>.