Evaluating the Calibration of Multi-Step-Ahead Density Forecasts Using Raw Moments

Malte Knüppel

Deutsche Bundesbank

June 2012

Motivation

- Forecasts are increasingly often made in the form of densities (fan charts, forecasts of Bayesian models,...)
- Forecast evaluation is not restricted to point forecasts (commonly, the mean forecasts) in these cases
  - In the case of point forecasts, for example, one can investigate whether mean forecasts are biased
  - Analogously, in the case of density forecasts, one can ask whether the density forecasts coincide with the true densities (correct calibration). Example of incorrect calibration: Normal densities, mean forecasts are unbiased, but variance forecasts are too small.
- Aim: Design *simple* test for calibration of *multi-step*-ahead forecasts

Evaluating Density Forecasts - The PITs

- A realization x_t is transformed into a PIT_t (probability integral transform) of the forecast density according to

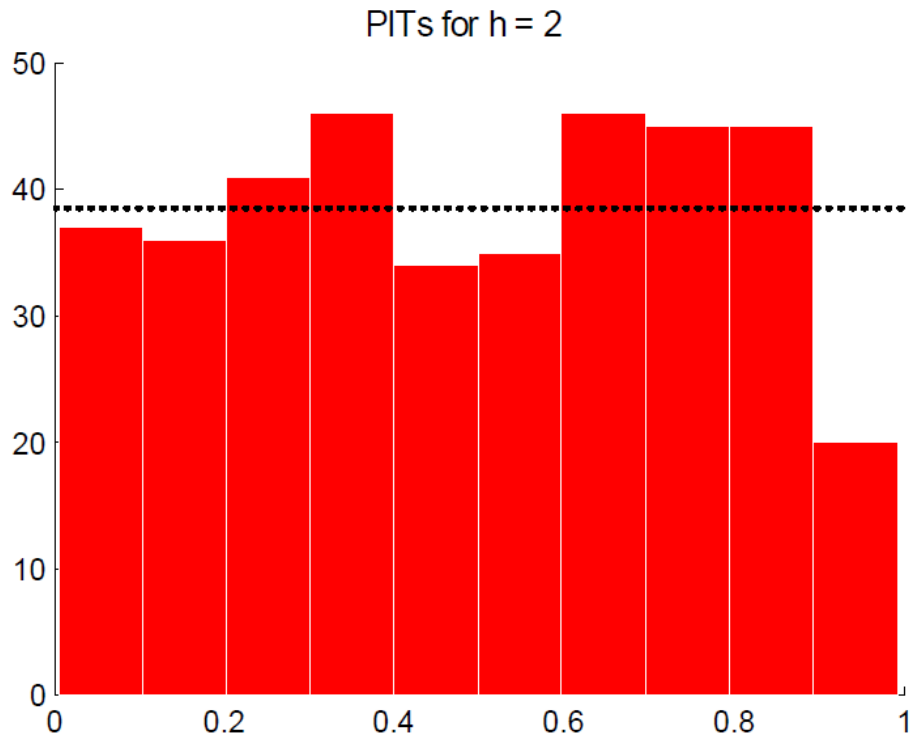
$$PIT_t = \int_{-\infty}^{x_t} \hat{f}_t(z)\, dz = \hat{F}_t(x_t)$$

  with $\hat{f}_t(\bullet)$ denoting the forecast density for period t
- If the density is calibrated correctly, PIT_t is uniformly distributed over interval $(0, 1)$, and tests can be based on this property
- Idea goes back to Rosenblatt (1952), appeared in Dawid (1984) and Smith (1985), was popularized by Diebold, Gunther & Tay (1998)

Evaluating Density Forecasts - The PITs

- Common way of presenting PIT_t: Histogram

PIT_t of 2-months-ahead CHF/USD exchange rate forecasts, T=385



- Histogram of PIT_t indicates, most notably, too few outcomes in upper decile - significant deviations?

Evaluating Density Forecasts - Existing Tests

- Several tests available to check if $PIT_t \sim U(0,1)$ under assumption that PIT_t is independent...
- ...but multi-step-ahead forecast errors are serially correlated, and so is PIT_t
- Tests used for multi-step-ahead density forecasts commonly rest on a second transformation

$$INT_t = \Phi^{-1}(PIT_t)$$

where $\Phi^{-1}(\bullet)$ is the standard normal inverse cumulative distribution function

- If PIT_t is uniformly distributed over $(0,1)$, INT_t (the inverse normal transform) is standard normally distributed
- Reason for this transformation: Serial correlation of normally distributed variables is easier to handle than serial correlation of uniformly distributed variables

Evaluating Density Forecasts - Existing Tests

- Three main approaches in the literature for serially correlated INT_t
  1. Use tests which require independence and issue a warning
  2. Based on Berkowitz (2001): Estimate

$$INT_t = c + \rho \cdot INT_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim N\left(0, \sigma^2\right)$ by maximum likelihood and use likelihood-ratio test of

$$H_0 : c = 0, \sigma^2 = 1 - \rho^2$$

  3. Based on normality tests for serially correlated data (Bai & Ng 2005, Bontemps & Meddahi 2005,...):
     Test for zero skewness and zero excess kurtosis of INT_t

- Corradi&Swanson (2005) proposed a test for multi-step-ahead forecasts accounting for parameter estimation uncertainty, but it is computationally burdensome and apparently never applied

Evaluating Density Forecasts - Existing Tests

- Drawbacks
  1. Tests which require independence: wrong (asymptotic) size
  2. Berkowitz test:
     Assumption concerning dynamics (AR(1)-process) can be incorrect
     $\Rightarrow$ wrong (asymptotic) size
     Only mean and variance used, skewness and kurtosis ignored
     $\Rightarrow$ power problems
  3. Normality tests:
     Only skewness and kurtosis used, mean and variance ignored
     $\Rightarrow$ power problems
- Latter approach could be extended to include lower moments, since mean and variance are known under H_0

Evaluating Density Forecasts - Raw-Moments Test

- One could test for for zero mean, unit variance, zero skewness, zero excess kurtosis
- But skewness and kurtosis are standardized moments, i.e. functions of mean and variance, which have to be estimated
  $\Rightarrow$ complicates tests
- Instead, one can use raw moments. Under H_0

$$
\begin{aligned}
E\left[INT_t\right] &= 0 \\
E\left[INT_t^2\right] &= 1 \\
E\left[INT_t^3\right] &= 0 \\
E\left[INT_t^4\right] &= 3 \\
&\vdots
\end{aligned}
$$

Evaluating Density Forecasts - Raw-Moments Test

- Testing raw moments is extremely simple
- Define vector

$$\mathbf{d}_t = \begin{bmatrix} INT_t \\ INT_t^2 - 1 \\ INT_t^3 \\ INT_t^4 - 3 \\ \vdots \end{bmatrix},$$

  and test whether $\frac{1}{T}\sum \mathbf{d}_t = \mathbf{0}$, using a *long-run* covariance matrix and the χ^2 distribution
- Instead of INT_t, one could just as well use PIT_t
- Testing is simplified by standardization of PIT_t

$$S\text{-}PIT_t = \sqrt{12}\,(PIT_t - 0.5)$$

  yielding uniformly distributed variables over $(-1.73, 1.73)$ with odd moments $= 0$ and variance $= 1$ under H_0

Evaluating Density Forecasts - Raw-Moments Test

- Testing based on $S\text{-}PIT_t$: Define vector

$$\mathbf{d}_t = \begin{bmatrix} S\text{-}PIT_t \\ S\text{-}PIT_t^2 - 1 \\ S\text{-}PIT_t^3 \\ S\text{-}PIT_t^4 - 1.8 \\ \vdots \end{bmatrix},$$

  and proceed as before.
- One could also consider other transformations of PIT_t. Only requirement is asymptotic normality of $\sum \mathbf{d}_t$
- Elements of long-run covariance matrix representing covariance between an even and an odd moment can be set to zero
  $\Rightarrow$ Better size and power properties
- In the following, quadratic spectral kernel is used

Simulations - Size

- Small sample performance: Consider MA(1)-process

$$x_t = \varepsilon_t + \theta \varepsilon_{t-1}$$

with $\varepsilon_t \sim N\left(0,\ 1/\left(1+\theta^2\right)\right)$, correctly calibrated density forecasts $\hat{f}_t = \phi\left(x_t\right)$, and sample size T.

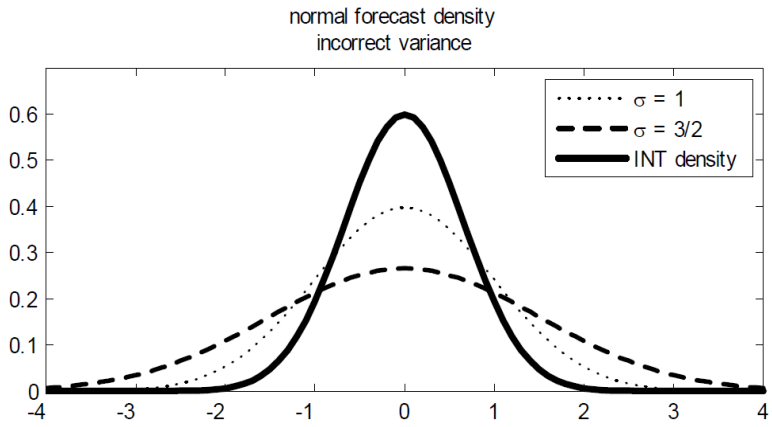
Actual size of tests if nominal size is 5%

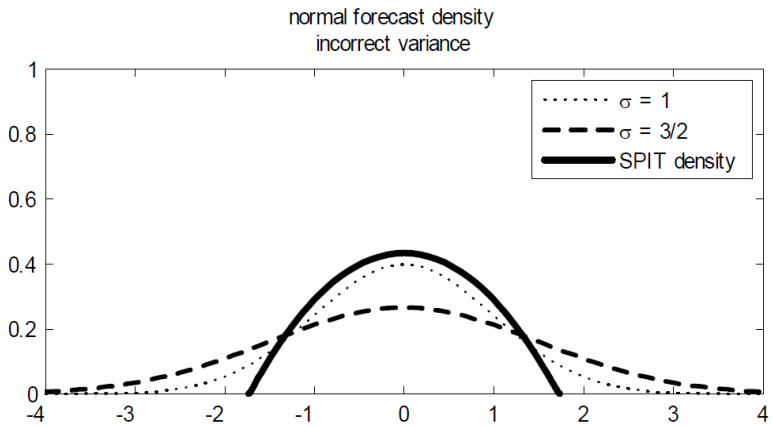
T	θ	Berkowitz	Bai&Ng	raw moments (1-4)	
		INT_t	INT_t	INT_t	$S\text{-}PIT_t$
50	0.0	0.051	0.023	0.169	0.034
50	0.9	0.024	0.013	0.147	0.026
200	0.0	0.050	0.090	0.128	0.046
200	0.9	0.023	0.064	0.147	0.044
1000	0.0	0.050	0.084	0.078	0.048
1000	0.9	0.023	0.085	0.087	0.050

- Size distortions of raw-moments test prohibitively large if test is based on INT_t, fairly contained if test is based on $S\text{-}PIT_t$

Simulations - Power

- Example:
  - $x_t \sim N(0,1)$, follows MA(1)-process like above
  - Forecast density $\hat{f}_t$ is $N\left(0, 1.5^2\right)$, i.e. too dispersed





- Clearly, INT_t not standard normal, and $S\text{-}PIT_t$ not uniformly distributed.
- How well do the tests discover these deviations?

Simulations - Size-adjusted Power

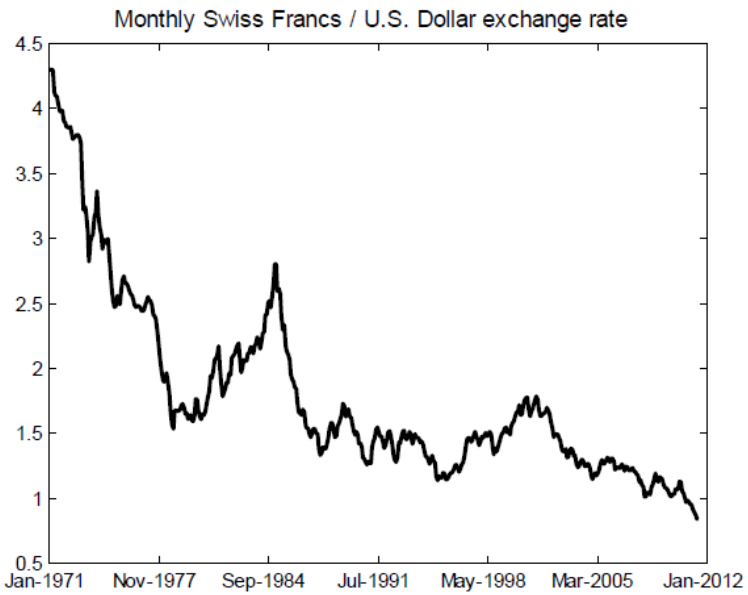
Power if forecast density is $N\left(0, 1.5^2\right)$ or Student's t (5 df, stand.)

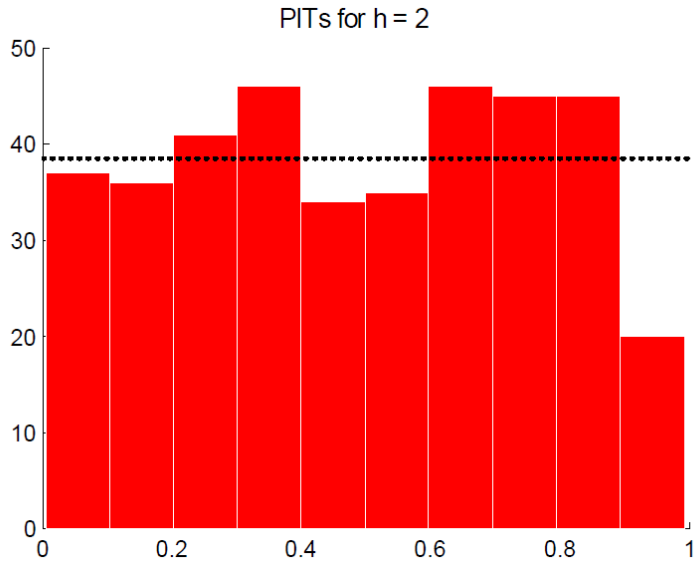
		$N\left(0, 1.5^2\right)$			standardized Student's t (5 df)		
T	θ	Berkowitz	Bai&Ng	raw moments	Berkowitz	Bai&Ng	raw moments
50	0.0	0.93	0.05	0.51	0.04	0.22	0.10
50	0.5	0.73	0.04	0.34	0.04	0.17	0.08
50	0.9	0.65	0.04	0.27	0.04	0.15	0.08
200	0.0	1.00	0.05	1.00	0.10	0.86	0.40
200	0.5	1.00	0.04	1.00	0.09	0.80	0.34
200	0.9	1.00	0.05	1.00	0.08	0.75	0.32

- Berkowitz and Bai&Ng tests can have very low power
- Raw-moments test here never has highest power, never lowest power, always at least moderate power in medium-sized samples

Application

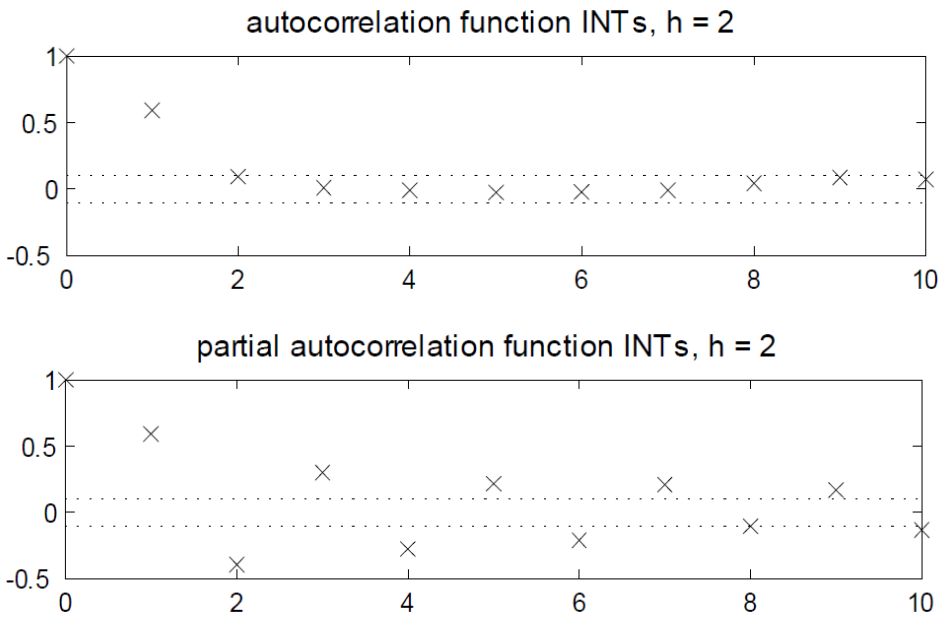
- Very simple model: Normal density forecasts for exchange rate h months ahead
  - mean = current value (random-walk assumption)
  - variance = MSFE of h-months-ahead mean forecasts during past 8 years (i.e. rolling window)





Application

- INT_t (and PIT_t) serially correlated



- INT_t (and PIT_t) appears to follow an MA($h-1$)-process

Application

Test results for $h = 2, 3, 4$

	moments						p-values	
	raw $S\text{-}PIT_t$				central INT_t			
	1st	2nd	3rd	4th	1st	2nd	Berkowitz	raw moments
$h = 2$	−0.04	0.88	−0.15	1.42	−0.06	0.78	0.085	0.027
$h = 3$	−0.07	0.88	−0.17	1.42	−0.07	0.77	0.191	0.039
$h = 4$	−0.09	0.89	−0.18	1.45	−0.09	0.77	0.286	0.197

- Evidence against correct calibration of 2- and 3-months-ahead density forecasts according to raw-moments test
- No such evidence according to Berkowitz test. Rejections probably caused by higher moments

Conclusions and Outlook

- Testing for correct calibration of multi-step-ahead density forecasts hardly addressed in the literature
- Existing approaches unsatisfactory due to neglected information or problematic assumptions
- Simple alternative given by testing raw moments and using (restricted) long-run covariance matrices
- Raw-moments tests should be based on $S\text{-}PIT_t$ (not on INT_t)
- Raw-moments tests have power against many misspecifications
- Berkowitz test appears more recommendable in small persistent samples due to mostly higher power
- Raw-moments tests can easily be extended to test for *complete* calibration. With m moments, use regression model

$$\mathbf{d}_t = \mathbf{c} + \boldsymbol{\rho} \circ \mathbf{d}_{t-h} + \boldsymbol{\varepsilon}_t$$

with $\mathbf{c}$ and $\boldsymbol{\rho}$ being $(m \times 1)$ vectors, and test $H_0 : \mathbf{c} = \boldsymbol{\rho} = \mathbf{0}$