Measuring forecast calibration

Chris Ferro, Keith Mitchell, Heba Bashaykh

Department of Mathematics

University of Exeter, UK

VALPRED workshop (Aussois, 12 March 2020)

What is calibration?

A set of forecasts is calibrated if the probabilistic meaning of the forecasts matches empirical reality.

Example. It rains on 40% of the days that were given a 40% chance of rain.

What is calibration?

A set of forecasts is calibrated if the probabilistic meaning of the forecasts matches empirical reality.

Example. It rains on 40% of the days that were given a 40% chance of rain.

This is an example of auto-calibration…

Auto-calibration

$X =$ outcome

$F =$ cumulative distribution function of probability forecast for X

Consider F and X as random variables taking values in a population of forecasts and outcomes.

Definition. F is **auto-calibrated** if $X \mid F \sim F$. The X that occur when the forecast is F have a frequency distribution equal to F.

Example. Let $X \in \{0, 1\}$, $f =$ forecast probability on $\{X = 1\}$ and $p_f = \Pr(X = 1 \mid f)$. Forecasts are auto-calibrated if $f = p_f$. A calibration (or reliability) diagram plots the points (f, p_f).

Decomposition of proper scores

X = outcome

F = cumulative distribution function of probability forecast for X

s = negatively oriented scoring rule: awards score $s(F, X)$

Write $s(F, G)$ for $\mathrm{E}_X\{s(F, X)\}$ when $X \sim G$.

Definition. s is **proper** if $s(F, G) \geq s(G, G)$ for all F and G.

Decomposition of proper scores

X = outcome

F = cumulative distribution function of probability forecast for X

s = negatively oriented scoring rule: awards score $s(F, X)$

Write $s(F, G)$ for $\mathrm{E}_X\{s(F, X)\}$ when $X \sim G$.

Definition. s is **proper** if $s(F, G) \geq s(G, G)$ for all F and G.

Decompose expected scores to obtain measures of calibration.
Write $X \sim P$ and $X \mid F \sim P_F$ and $d(F, G) = s(F, G) - s(G, G)$.

$$\mathrm{E}\{s(F, X)\} = \underbrace{\mathrm{E}_F\{d(F, P_F)\}}_{\text{Miscalibration}} - \underbrace{\mathrm{E}_F\{d(P, P_F)\}}_{\text{Resolution}} + \underbrace{s(P, P)}_{\text{Uncertainty}}$$

Example. $\mathrm{E}\{(f - X)^2\} = \mathrm{E}_f\{(f - p_f)^2\} - \mathrm{E}_f\{(p_f - p)^2\} + p(1-p)$

Beyond auto-calibration

$$\mathrm{E}\{s(F,X)\} = \underbrace{\mathrm{E}_F\{d(F,P_F)\}}_{\text{Miscalibration}} - \underbrace{\mathrm{E}_F\{d(P,P_F)\}}_{\text{Resolution}} + \underbrace{s(P,P)}_{\text{Uncertainty}}$$

Auto-calibration concerns the biases of specific forecasts, F.

We might want to know about biases in other circumstances too, for example in certain seasons or atmospheric regimes.

Beyond auto-calibration

$$\mathrm{E}\{s(F,X)\} = \underbrace{\mathrm{E}_F\{d(F,P_F)\}}_{\text{Miscalibration}} - \underbrace{\mathrm{E}_F\{d(P,P_F)\}}_{\text{Resolution}} + \underbrace{s(P,P)}_{\text{Uncertainty}}$$

Auto-calibration concerns the biases of specific forecasts, F.

We might want to know about biases in other circumstances too, for example in certain seasons or atmospheric regimes.

Estimating P_F is difficult when forecasts are unique.

This is addressed by binning but that needs large samples unless X is binary, particularly if we stratify by season etc.

Other modes of calibration can help us to learn about biases in any circumstances of interest and when data are limited.

What is calibration?

Marginal and probabilistic calibration

Score decompositions

Deterministic forecasts

Conclusion

Marginal and probabilistic calibration

$F =$ cumulative distribution function of probability forecast for X

$P =$ unconditional (climatological) distribution of X

$\bar{F} =$ unconditional (climatological) mixture distribution, $\mathrm{E}(F)$

Definition. F is **marginally calibrated** if $\bar{F} = P$.

Forecast probabilities for $\{X \le x\}$ are correct on average: $\mathrm{E}\{F(x)\} = \Pr(X \le x)$. Assessed by plotting $\bar{F}$ and P.

Marginal and probabilistic calibration

F = cumulative distribution function of probability forecast for X

P = unconditional (climatological) distribution of X

$\bar{F}$ = unconditional (climatological) mixture distribution, $\mathrm{E}(F)$

Definition. F is **marginally calibrated** if $\bar{F} = P$.

Forecast probabilities for $\{X \leq x\}$ are correct on average: $\mathrm{E}\{F(x)\} = \Pr(X \leq x)$. Assessed by plotting $\bar{F}$ and P.

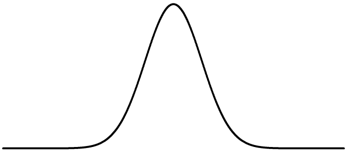
Definition. F is **probabilistically calibrated** if $F(X) \sim \mathrm{U}(0,1)$.

Forecast quantiles exceed X with the correct frequency: $\Pr\{X \leq F^{-1}(p)\} = p$. Assessed with PIT histogram and tests.

Forecasts can satisfy none, one or both of these definitions...

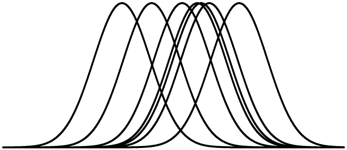
Marginal and probabilistic calibration

$X \sim \mathrm{N}(\mu, \sigma^2)$ where $\mu \sim \mathrm{N}(0, 1)$



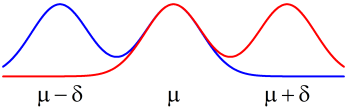
Marginal but not probabilistic:

$F = \mathrm{N}(\mu', \sigma^2)$ where $\mu' \sim \mathrm{N}(0, 1)$



Probabilistic but not marginal:

$F = \frac{1}{2}\mathrm{N}(\mu, \sigma^2) + \frac{1}{2}\mathrm{N}(\mu + \Delta, \sigma^2)$
where $\Delta \sim \mathrm{U}\{-\delta, \delta\}$



Conditional marginal and probabilistic calibration

Consider stratifying (F, X) by the value of A (e.g. by season).

$P_A =$ conditional (climatological) distribution of X given A

$\bar{F}_A =$ conditional (climatological) mixture distribution, $\mathrm{E}(F \mid A)$

Definition. F is **marginally calibrated w.r.t. A** if $\bar{F}_A = P_A$.

Forecast probabilities for $\{X \le x\}$ are correct on average within the strata defined by A: $\mathrm{E}\{F(x) \mid A\} = \Pr(X \le x \mid A)$.

Conditional marginal and probabilistic calibration

Consider stratifying (F, X) by the value of A (e.g. by season).

$P_A =$ conditional (climatological) distribution of X given A

$\bar{F}_A =$ conditional (climatological) mixture distribution, $\mathrm{E}(F \mid A)$

Definition. F is **marginally calibrated w.r.t. $\boldsymbol{A}$** if $\bar{F}_A = P_A$.

Forecast probabilities for $\{X \leq x\}$ are correct on average within the strata defined by A: $\mathrm{E}\{F(x) \mid A\} = \Pr(X \leq x \mid A)$.

Definition. F is **prob. calibrated w.r.t. $\boldsymbol{A}$** if $F(X) \mid A \sim \mathrm{U}(0{,}1)$.

Forecast quantiles exceed X with the correct frequency within the strata defined by A: $\Pr\{X \leq F^{-1}(p) \mid A\} = p$.

A should be knowable when the forecast is made.

Conditional marginal and probabilistic calibration

$P_A =$ conditional (climatological) distribution of X given A

$\bar{F}_A =$ conditional (climatological) mixture distribution, $\mathrm{E}(F \mid A)$

Definition. F is **marginally calibrated w.r.t. A** if $\bar{F}_A = P_A$.

Definition. F is **prob. calibrated w.r.t. A** if $F(X) \mid A \sim \mathrm{U}(0,1)$.

Conditional marginal calibration ⇒ marginal calibration.

Conditional probabilistic calibration ⇒ probabilistic calibration.

If $A = F$ then both definitions are equivalent to auto-calibration.

If $A = (F, B)$ then both are equivalent and forecasts are 'ideal'.

What is calibration?

Marginal and probabilistic calibration

Score decompositions

Deterministic forecasts

Conclusion

Score decompositions: marginal calibration

$$
\begin{aligned}
P &= \text{unconditional distribution of } X \\
P_A &= \text{conditional distribution of } X \text{ given } A \\
\bar{F}_A &= \text{conditional mixture, } \mathrm{E}(F \mid A)\text{, of } F \text{ given } A
\end{aligned}
$$

Marginal calibration corresponds to $\bar{F}_A = P_A$.

Score decompositions: marginal calibration

P = unconditional distribution of X

P_A = conditional distribution of X given A

$\bar{F}_A$ = conditional mixture, $\mathrm{E}(F \mid A)$, of F given A

Marginal calibration corresponds to $\bar{F}_A = P_A$.

$$\mathrm{E}\{s(F,X)\} = \underbrace{\mathrm{E}_A[\mathrm{E}\{s(F,X) \mid A\} - s(\bar{F}_A, P_A)]}_{\mathrm{MEX}_A}$$

$$+ \underbrace{\mathrm{E}_A\{d(\bar{F}_A, P_A)\}}_{\mathrm{MMC}_A} - \underbrace{\mathrm{E}_A\{d(P, P_A)\}}_{\mathrm{RES}_A} + \underbrace{s(P,P)}_{\mathrm{UNC}}$$

RES_A = resolution from stratifying X by A

MMC_A = marginal miscalibration within strata

MEX_A = excess variation of F about $\bar{F}_A$ within strata

Score decompositions: probabilistic calibration

$P_A =$ conditional distribution of X given A

$C_A =$ conditional distribution of $F(X)$ given A

$Q_A = C_A^{-1} P_A$ is the (fixed) forecast for which $Q_A(X) \mid A \sim C_A$

Probabilistic calibration corresponds to $Q_A = P_A$.

Score decompositions: probabilistic calibration

$P_A =$ conditional distribution of X given A

$C_A =$ conditional distribution of $F(X)$ given A

$Q_A = C_A^{-1} P_A$ is the (fixed) forecast for which $Q_A(X) \mid A \sim C_A$

Probabilistic calibration corresponds to $Q_A = P_A$.

$$\mathrm{E}\{s(F,X)\} = \underbrace{\mathrm{E}_A[\mathrm{E}\{s(F,X) \mid A\} - s(Q_A, P_A)]}_{\mathrm{PEX}_A} + \underbrace{\mathrm{E}_A\{d(Q_A, P_A)\}}_{\mathrm{PMC}_A} - \underbrace{\mathrm{E}_A\{d(P, P_A)\}}_{\mathrm{RES}_A} + \underbrace{s(P, P)}_{\mathrm{UNC}}$$

$\mathrm{PMC}_A =$ probabilistic miscalibration within strata

$\mathrm{PEX}_A =$ excess variation of F about Q_A within strata

Score decompositions

Marginal:
$$\mathrm{E}_A[\mathrm{E}\{s(F,X)\mid A\} - s(\bar{F}_A, P_A)] + \mathrm{E}_A\{d(\bar{F}_A, P_A)\} - \mathrm{E}_A\{d(P, P_A)\} + s(P,P)$$

Probabilistic:
$$\mathrm{E}_A[\mathrm{E}\{s(F,X)\mid A\} - s(Q_A, P_A)] + \mathrm{E}_A\{d(Q_A, P_A)\} - \mathrm{E}_A\{d(P, P_A)\} + s(P,P)$$

If $A = (F, B)$ then $\bar{F}_A = Q_A = F$ and both decompositions are

$$\mathrm{E}\{s(F,X)\} = \mathrm{E}_A\{d(F, P_A)\} - \mathrm{E}_A\{d(P, P_A)\} + s(P,P).$$

If $A = F$ then we retrieve the decomposition for auto-calibration.

If A is constant then we obtain decompositions for unconditional marginal and probabilistic calibration.

If A identifies which bin a forecast belongs to then the marginal decomposition corresponds to other published decompositions.

What is calibration?

Marginal and probabilistic calibration

Score decompositions

Deterministic forecasts

Conclusion

Deterministic forecasts

$\hat{X} = T(F)$, a functional, T, of F issued as a point forecast for X

$s =$ negatively oriented scoring function: awards score $s(\hat{X}, X)$

Write $s(\hat{X}, G)$ for $\mathrm{E}_X\{s(\hat{X}, X)\}$ when $X \sim G$.

Definition. s is **consistent for *T*** if $s(\hat{X}, G) \geq s(T(G), G)$ for all $\hat{X}$ and G.

Deterministic forecasts

$\hat{X} = T(F)$, a functional, T, of F issued as a point forecast for X

$s =$ negatively oriented scoring function: awards score $s(\hat{X}, X)$

Write $s(\hat{X}, G)$ for $\mathrm{E}_X\{s(\hat{X}, X)\}$ when $X \sim G$.

Definition. s is **consistent for *T*** if $s(\hat{X}, G) \geq s(T(G), G)$ for all $\hat{X}$ and G.

Write $X \sim P$, $X \mid F \sim P_F$ and $d(\hat{X}, G) = s(\hat{X}, G) - s(T(G), G)$.

Definition. F is **auto-calibrated for *T*** if $T(P_F) = \hat{X}$. The X that occur when the forecast is F have functional T equal to $\hat{X}$.

$$\mathrm{E}\{s(\hat{X}, X)\} = \underbrace{\mathrm{E}_F\{d(\hat{X}, P_F)\}}_{\text{mc}} - \underbrace{\mathrm{E}_F\{d(T(P), P_F)\}}_{\text{res}} + \underbrace{s(T(P), P)}_{\text{unc}}$$

Functional calibration

Definition. $V(\hat{X}, X)$ is an **identifying function** for T if it is increasing in $\hat{X}$ and $\mathrm{E}_X\{V(T(G), X)\} = 0$ when $X \sim G$.

Example. If $T(G) = G^{-1}(p)$ then $V(\hat{X}, X) = \mathbb{1}(X \leq \hat{X}) - p$.

Example. If $T(G) = \int h(x)\,\mathrm{d}G(x)$ then $V(\hat{X}, X) = \hat{X} - h(X)$.

Functional calibration

Definition. $V(\hat{X}, X)$ is an **identifying function** for T if it is increasing in $\hat{X}$ and $\mathrm{E}_X\{V(T(G), X)\} = 0$ when $X \sim G$.

Example. If $T(G) = G^{-1}(p)$ then $V(\hat{X}, X) = \mathbb{1}(X \leq \hat{X}) - p$.

Example. If $T(G) = \int h(x)\,\mathrm{d}G(x)$ then $V(\hat{X}, X) = \hat{X} - h(X)$.

Definition. $\hat{X}$ is **calibrated for *T*** if $\mathrm{E}\{V(\hat{X}, X)\} = 0$.

Definition. $\hat{X}$ is **calibrated for *T* w.r.t. *A*** if $\mathrm{E}\{V(\hat{X}, X) \mid A\} = 0$.

Properties defined by V are correct on average (within strata): moments ≈ marginal calibration; quantiles ≈ prob. calibration.

Conditional calibration for $T \Rightarrow$ calibration for T.

If $A = F$, the definition is equivalent to auto-calibration for T.

Score decompositions: functional calibration

$P_A =$ conditional distribution of X given A

$\bar{V}_A =$ conditional expectation, $\mathrm{E}\{V(\hat{X}, X) \mid A\}$, of V given A

$T_A =$ the fixed forecast for which $\mathrm{E}_X\{V(T_A, X) \mid A\} = \bar{V}_A$

Calibration for T corresponds to $T_A = T(P_A)$.

Score decompositions: functional calibration

$P_A =$ conditional distribution of X given A

$\bar{V}_A =$ conditional expectation, $\mathrm{E}\{V(\hat{X}, X) \mid A\}$, of V given A

$T_A =$ the fixed forecast for which $\mathrm{E}_X\{V(T_A, X) \mid A\} = \bar{V}_A$

Calibration for T corresponds to $T_A = T(P_A)$.

$$\mathrm{E}\{s(\hat{X}, X)\} = \underbrace{\mathrm{E}_A[\mathrm{E}\{s(\hat{X}, X) \mid A\} - s(T_A, P_A)]}_{\mathrm{FEX}_A} + \underbrace{\mathrm{E}_A\{d(T_A, P_A)\}}_{\mathrm{FMC}_A} - \underbrace{\mathrm{E}_A\{d(T(P), P_A)\}}_{\mathrm{res}_A} + \underbrace{s(T(P), P)}_{\mathrm{unc}}$$

$\mathrm{FMC}_A =$ functional miscalibration within strata

$\mathrm{FEX}_A =$ excess variation of $\hat{X}$ about T_A within strata

Summary

1. Auto-calibration tells us about the biases of specific forecasts only and can be impractical to measure.
2. Other modes of calibration can tell us about biases in any circumstances of interest and when data are limited.
3. Conditional marginal calibration: climatological forecast distributions are correct within strata.

   Conditional probabilistic calibration: coverages of prediction intervals are correct within strata.

   Conditional functional calibration: relevant properties of point forecasts are correct within strata.
4. Scores can be decomposed to measure these modes of calibration, generalizing existing decompositions.
5. These decompositions also measure the skill of forecasts relative to a reference forecast within strata.

References

Bröcker (2009) Reliability, sufficiency, and the decomposition of proper scores. Quarterly Journal of the Royal Meteorological Society, 135, 1512–1519.

Diebold, Gunther, Tay (1998) Evaluating density forecasts with applications to financial risk management. International Economic Review, 39, 863–882.

Ehm, Ovcharov (2017) Bias-corrected score decomposition for generalized quantiles. Biometrika, 104, 473–480.

Ferro, Mitchell, Bashaykh (2020) Measures of calibration for probabilistic and deterministic forecasts. In preparation.

Gneiting, Balabdaoui, Raftery (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society B, 69, 243–268.

Gneiting, Ranjan (2013) Combining predictive distributions. Electronic Journal of Statistics, 7, 1747–1782.

Mitchell, Wallis (2011) Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. Journal of Applied Econometrics, 26, 1023–1040.

Stephenson, Coelho, Jolliffe (2008) Two extra components in the Brier score decomposition. Weather and Forecasting, 23, 752–757.

Tsyplakov (2011) Evaluating density forecasts: a comment. MPRA paper no. 32728, http://mpra.ub.uni-muenchen.de/32728/