Multivariate Calibration is Performative: A Perspective on Pitfalls and Progress

Sofian Zalouk [1] **Charles Marx** [2] **Syrine Belakaria** [2] **Christopher De Sa** [1] **Stefano Ermon** [2]

Abstract

Reliable structured prediction requires forecasts that are calibrated both marginally and jointly. We identify two fundamental challenges in post-hoc multivariate calibration: *performativity*, where recalibration alters the very statistics used to assess calibration, and *proportionality*, the need to preserve a model's learned dependence structure when correcting miscalibration. We show that two trivial forecasters can satisfy any notion of multivariate calibration, motivating a search for minimal corrections. By framing recalibration as a constrained optimization problem, we outline principled directions for feedback-aware algorithms that deliver reliable multivariate forecasts.

1. Introduction

Modern machine learning systems are increasingly deployed in domains that require not only accurate predictions but also reliable uncertainty quantification. This is particularly true in structured prediction settings—such as time-series forecasting, tabular prediction, or multivariate regression—where the model must output a distribution over a vector-valued target $Y \in \mathbb{R}^d$ given input features X.

In the univariate regression setting, the notion of *calibration* is well-understood: a forecaster is calibrated if its predicted quantiles (or prediction intervals) match empirical frequencies (Gneiting et al., 2007). For instance, a 90% prediction interval should contain the true outcome 90% of the time. This has led to a variety of reliable post-hoc calibration techniques—such as isotonic regression (Kuleshov et al., 2018) and histogram binning—that adjust predictive distributions to align with observed outcomes.

However, in structured prediction tasks, the calibration of individual output dimensions is not sufficient. Even if each marginal forecast is perfectly calibrated, the joint distribution may still be misspecified. Consider a weather forecasting system predicting the probability of rainfall at multiple locations within a region. If the marginal forecasts at each location are calibrated—but the joint dependence is not—then the system might frequently predict high probabilities of rain at all locations simultaneously, even when such joint events are rare in reality. This could lead to overestimated flood risk, inefficient resource allocation (e.g., unnecessary deployment of emergency crews), or poor performance in ensemble-based downstream tasks like trajectory sampling or hydrological simulation. In such settings, downstream applications rely not only on marginal correctness but also on the validity of the joint structure.

Several recent works have recognized the importance of modeling joint uncertainty and have proposed methods to enforce multivariate calibration using held-out calibration data. Classical techniques such as the Schaake Shuffle (Clark et al., 2004) and Ensemble Copula Coupling (ECC) (Schefzik et al., 2013) combine calibrated marginal forecasts with a copula, either from historical analogs or the raw ensemble, in an effort to produce realistic joint samples. More recent approaches, such as those by Kock et al. (2024) and Chung et al. (2024), aim to directly calibrate the joint distribution, either by replacing the model's dependence structure with one estimated from the calibration set, or by resampling from the forecaster to match empirical multivariate rank statistics.

Our goal is to clarify the conceptual landscape of post-hoc multivariate calibration and outline foundational issues that must be addressed. Our contributions are as follows:

- We identify an overlooked *performative prediction* challenge in multivariate calibration and demonstrate how this issue prevents existing recalibration methods from achieving reliable calibration.
- We define and motivate the concept of *proportionality* in multivariate calibration, and use it to demonstrate that existing recalibration approaches can erase useful information about the joint dependence structure of the original forecaster.
- We outline promising future directions—including op-

[1]Department of Computer Science, Cornell University [2]Department of Computer Science, Stanford University. Correspondence to: Sofian Zalouk <saz43@cornell.edu>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

timal transport, fixed-point iterative recalibration, and confidence set construction—and highlight open challenges in designing calibration methods.

We hope these contributions highlight the need for additional techniques in multivariate calibration and are a step toward resolving these challenges in structured prediction settings.

2. Background on Multivariate Calibration

We consider the problem of modeling the conditional distribution of a multivariate random variable $Y \in \mathbb{R}^d$ given covariates X. Let $H_{Y|X}$ and $h_{Y|X}$ denote the true conditional cumulative distribution function (CDF) and probability density function (PDF), respectively. Likewise, let $\hat{H}_{Y|X}$ and $\hat{h}_{Y|X}$ denote the forecaster's CDF and PDF, respectively.

Calibration in the Univariate Setting. In the univariate case ($d = 1$), a forecaster $\hat{H}_{Y|X}$ is said to be *probabilistically calibrated* (Gneiting et al., 2007) if:

$$P(\hat{H}_{Y|X}(Y) \leq p) = p, \quad \forall p \in [0, 1].$$

This is equivalent to requiring that the Probability Integral Transform (PIT) of the forecast, defined as $U = \hat{H}_{Y|X}(Y)$, is uniformly distributed: $U \sim \text{Uniform}[0, 1]$.

2.1. Multivariate Calibration via Projection Functions

Extending probabilistic calibration to multivariate targets introduces subtle challenges. While the PIT provides a natural tool in the univariate setting, its direct extension to higher dimensions is problematic. For one, the multivariate quantile function is not uniquely defined (Belloni & Winkler, 2011). Additionally, evaluating a multivariate CDF at its own samples does not yield a uniformly distributed variable (Genest & Rivest, 2001; Barbe et al., 1996). To address these limitations, prior work has proposed assessing calibration using scalar-valued projections—or *pre-rank functions*—of forecast–observation pairs.

This idea is well-established in the literature (Knüppel et al., 2022), and underlies a variety of diagnostic tools for multivariate forecasts. The key idea is to evaluate calibration via a real-valued summary statistic $g(\hat{H}, y)$, where $\hat{H}$ is a predictive distribution and y is the realized outcome. These projections reduce multivariate forecast–observation pairs to a scalar quantity that can be evaluated using univariate techniques, such as histogram uniformity or PIT-based diagnostics. Different choices of g emphasize different aspects of forecast quality, such as tail behavior, dependence structure, or spatial coherence.

Let $g : \mathcal{H} \times \mathcal{Y} \to \mathbb{R}$ be a pre-rank function that takes a predictive distribution and an outcome and returns a scalar. For each input x, let $Z = g(\hat{H}_{Y|x}, Y)$ where $Y \sim H_{Y|x}$, and let $\hat{Z} = g(\hat{H}_{Y|x}, \hat{Y})$, where $\hat{Y} \sim \hat{H}_{Y|x}$. Define $\hat{H}_{Z|x}$ to be the CDF of $\hat{Z}$ conditioned on x.

Definition 1 (g-calibration)**.** (Knüppel et al., 2022) A multivariate forecaster $\hat{H}_{Y|X}$ is said to be *g-calibrated* if, for all $p \in [0, 1]$,

$$P\left(\hat{H}_{Z|X}(Z) \leq p\right) = p,$$

where $Z = g(\hat{H}_{Y|X}, Y)$ and $\hat{H}_{Z|X}$ is the CDF of $\hat{Z} = g(\hat{H}_{Y|X}, \hat{Y})$ conditional on X.

This framework recovers many standard notions of calibration. For instance, choosing $g(\hat{H}, y) = \hat{H}(y)$ corresponds to *copula calibration* (Ziegel & Gneiting, 2014), while other choices recover rank-based or score-based notions.

To illustrate these concepts, we introduce a running example in Figure 1, to be used throughout the paper. The ground truth distribution is a bivariate Gaussian with negatively correlated components. The forecaster uses the correct marginal distributions but assumes independence across dimensions. As shown in Figure 1(a), this results in a forecaster that is marginally calibrated by design but fails the copula PIT diagnostic, revealing joint miscalibration. We will return to this example to evaluate the behavior of previous recalibration methods.

We further define an approximate notion of calibration, which allows for a margin of error and provides a measure of how close a forecaster is to satisfying g-calibration.

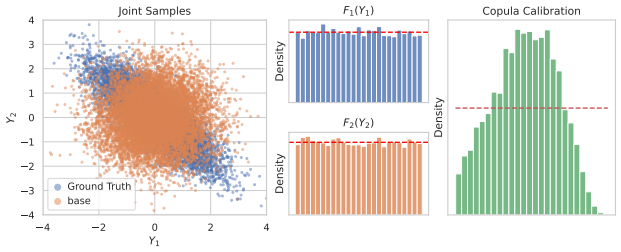
Definition 2. A multivariate forecaster $\hat{H}_{Y|X}$ is said to be (g, ε)-calibrated if, for all $p \in [0, 1]$,

$$\left| P\left(\hat{H}_{Z|X}(Z) \leq p\right) - p \right| \leq \varepsilon.$$

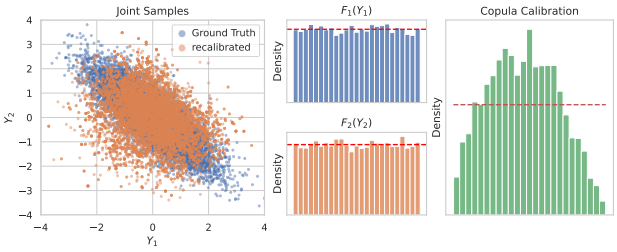
Performative prediction. In classical supervised learning, the data distribution is fixed and independent of the model. In contrast, the framework of *performative prediction* (Perdomo et al., 2020) considers settings in which the predictive model influences the distribution on which it is evaluated. Formally, deploying a forecaster $\hat{H}_{Y|X}$ induces a data distribution $\mathcal{D}(\hat{H})$ over covariate–response pairs (X, Y), reflecting structural feedback between prediction and outcome. In this setting, model training becomes inherently dynamic: updates aim to minimize risk over the distribution induced by the previous forecaster,

$$\hat{H}^{(t+1)} = \arg\min_{\hat{H}} \mathbb{E}_{(X,Y)\sim\mathcal{D}(\hat{H}^{(t)})} \left[\ell(X, Y; \hat{H})\right].$$

This motivates the study of *repeated risk minimization* that converges to a fixed point, and notions of *performative stability* and *optimality*, which formalize when such fixed points are well-defined and globally risk-minimizing.



(a) Base forecaster.



(b) Recalibrated forecaster.

Figure 1: Toy example illustrating marginal vs. joint (copula) calibration. The ground truth distribution is a bivariate Gaussian with nonzero correlation. The base forecaster (a) has the correct univariate marginals but an incorrect dependence structure, leading to uniform marginal histograms but a miscalibrated Copula PIT histogram. In (b), we apply the resampling method from Chung et al. (2024) to recalibrate the forecaster. While the marginals remain calibrated, the Copula PIT histogram remains non-uniform. This illustrates that ignoring the performative effect when enforcing multivariate calibration can result in forecasters that remain jointly miscalibrated.

3. Proportionality in Multivariate Recalibration

Trained forecasters learn meaningful joint dependence among target variables, and recalibration should preserve this structure to retain valuable predictive signal. We argue that recalibration should be *proportional*: it should intervene only to the extent necessary to achieve calibration, preserving the forecaster's learned dependence structure whenever possible. This section formalizes this principle and diagnoses where existing approaches fail to uphold it.

Let $\mathcal{H}$ denote the space of forecasters, and let $\mathcal{H}_{\text{cal}} \subset \mathcal{H}$ be the subset of g-calibrated forecasters. Given a miscalibrated forecaster $H \in \mathcal{H}$, a recalibration method $R : \mathcal{H} \to \mathcal{H}_{\text{cal}}$ satisfies α-*proportionality* if it stays close to the original model while ensuring calibration:

Definition 3 (α-Proportional Recalibration)**.** A recalibration map R is α-proportional if, for all $H \in \mathcal{H}$,

$$D(R(H), H) \leq \inf_{H' \in \mathcal{H}_{\text{cal}}} D(H', H) + \alpha,$$

for some divergence measure D over forecasters.

This minimality condition ensures that R behaves like a near-projection onto the calibrated set, applying only the smallest necessary adjustment. A natural special case is *invariance*: if a forecaster is already calibrated ($H \in \mathcal{H}_{\text{cal}}$), then a proportional method with $\alpha = 0$ should return it unchanged. Recalibration methods that fail to satisfy this property are over-correcting even in the absence of error.

Two Extremes: Perfect and Naive Calibration. Consider two extremes in the space of calibrated forecasters. At one end lies the uninformative forecaster H_Y, which ignores the input X and always outputs the marginal distribution of Y. While this forecaster is g-calibrated (Appendix A), it is clearly not useful—it produces constant predictions and offers no conditional signal. At the other end lies the Bayes-optimal forecaster $H_{Y|X}$, which is perfectly calibrated for all g but infeasible in practice. Our goal is to land somewhere between these extremes: to preserve as much of the original forecaster's signal as possible while intervening just enough to achieve calibration. The principle of proportionality provides a formal means of navigating this tradeoff.

Proportionality as a Diagnostic Tool. Proportionality also provides a useful lens for evaluating existing recalibration methods. Many approaches fall closer to the uninformative end of the spectrum by discarding the model's learned joint structure in favor of a fixed dependence estimated from historical data. For example, the Schaake Shuffle and the method of Kock et al. (2024) construct new joint distributions using empirical ranks or copulas fit directly on the calibration set, independent of the model's predictions. While these methods satisfy calibration by construction, proportionality reveals that they apply unnecessarily large interventions and fail to preserve the learned structure. In particular, they violate *invariance*, as they modify forecasts that may already be calibrated. Proportionality thus serves as a key desiderata and a diagnostic tool for identifying when recalibration methods overcorrect the model's outputs.

4. Performativity in Multivariate Recalibration

Up to this point, we have motivated the need for recalibration procedures that are proportionate and structure-preserving—ideally transforming a forecaster just enough to achieve calibration, without discarding its useful, learned dependence structure. Here we introduce what we believe to be a central and previously overlooked insight: *most widely used notions of multivariate calibration are inherently per-*

formative. That is, the act of recalibrating a forecaster alters the very pre-rank statistics used to assess whether calibration has been achieved—and more importantly, any method that does not account for this alteration is not calibrated with respect to the updated pre-rank statistics. We show that this issue is not incidental, but a general consequence of post-hoc g-calibration when g depends on the forecaster. As a concrete example, we demonstrate how the sampling-based method of Chung et al. (2024) falls into this pitfall.

Figure 1 illustrates this failure directly. Panel (a) shows the base forecaster, which has correctly calibrated marginals but an incorrect dependence structure. Panel (b) shows the result of applying the method of Chung et al. (2024), which resamples from the forecaster to achieve calibration. The post-processed model remains miscalibrated: a direct manifestation of performative prediction, where recalibration changes the distribution over pre-rank statistics.

Calibration as a performative learning objective. To formalize the recalibration problem, we consider finding a new forecaster $\hat{H}^{\text{new}}_{Y|X}$ that minimizes a chosen notion of calibration error. Let $g(\hat{H}, y)$ denote a pre-rank function, and let $Z = g(\hat{H}_{Y|X}, Y)$, $\hat{Z} = g(\hat{H}_{Y|X}, \hat{Y})$, with $Y \sim H_{Y|X}$, $\hat{Y} \sim \hat{H}_{Y|X}$. A common goal is to minimize a scalar-valued calibration error metric, such as

$$\mathrm{CE}_g(\hat{H}) = \mathbb{E}_{p\sim\mathcal{U}(0,1)} \left| P_X \left(\hat{H}_{Z|X}(Z) \le p \right) - p \right|,$$

where $\hat{H}_{Z|X}$ is the conditional CDF of $\hat{Z} \mid X$. This objective is often posed as: $\hat{H}^{\text{new}} \in \arg\min_{\hat{H}} \ \mathrm{CE}_g(\hat{H})$.

At first glance, this appears analogous to classical risk minimization. However, a key subtlety arises when g depends on the forecaster: the distributions of Z and $\hat{Z}$ used to compute $\mathrm{CE}_g(\hat{H})$ are themselves functions of $\hat{H}$. Thus, this is not an ordinary loss minimization problem, but a *performative prediction* objective (Perdomo et al., 2020).

More formally, we may write the distribution of the pre-rank statistics as $\mathcal{D}_g(\hat{H}) := \mathrm{Law}\left(g(\hat{H}, Y),\ g(\hat{H}, \hat{Y})\right)$, and reinterpret the calibration objective as

$$\hat{H}^{\text{new}} \in \arg\min_{\hat{H}} \ \mathrm{CE}_g(\mathcal{D}_g(\hat{H})).$$

This framing makes the *self-dependence* of the optimization problem explicit: the distribution over which calibration is assessed depends on the model being optimized. In the language of Perdomo et al. (2020), calibration is not evaluated on a fixed distribution, but rather on a *model-dependent* distribution $\mathcal{D}_g(\hat{H})$. Ignoring this feedback loop—as most existing approaches do—can lead to solutions that minimize calibration error with respect to an outdated diagnostic, while remaining miscalibrated under the updated one.

Impact on existing methods. This issue affects several prominent definitions of multivariate calibration, all of which define the pre-rank function $g(\hat{H}_{Y|x}, y)$ in a way that depends explicitly on the forecaster. In *copula calibration* (Ziegel & Gneiting, 2014), the pre-rank is defined as $g(\hat{H}_{Y|x}, y) = \hat{H}_{Y|x}(y)$, the model's own multivariate CDF evaluated at the true outcome. In *high-density region calibration* (Chung et al., 2024), it is taken to be the predictive density, $g(\hat{H}_{Y|x}, y) = \hat{h}_{Y|x}(y)$. In *score-based calibration* (Knüppel et al., 2022), the pre-rank is defined via a proper scoring function S as $g(\hat{H}_{Y|x}, y) = \mathbb{E}_{\hat{Y}\sim\hat{H}_{Y|x}}[S(\hat{Y}, y)]$. In each case, changing the forecaster induces a shift not only in the forecast distribution but also in the statistic used to evaluate calibration.

Fixed-point calibration. Our contribution is to clarify that many of the most statistically grounded notions of multivariate calibration are inherently performative. Without accounting for this, post-hoc methods may fail to achieve calibration. Rather than viewing this performativity as a flaw, we argue it is a defining feature of multivariate calibration: calibration should be understood as a fixed-point condition, where a forecaster remains calibrated even after transformation, relative to its own induced diagnostic.

5. Potential Approaches & Future Directions

Achieving reliable multivariate calibration requires satisfying two key principles: proportionality, to preserve the structure of the original forecaster, and performativity, to ensure that the forecaster is calibrated consistently after intervention. A promising direction is to pose recalibration as a constrained optimization problem, minimizing deviations from the original forecaster subject to calibration constraints. Optimal transport provides a natural foundation for this view, with recent work on vector copulas (Fan & Henry, 2023), vector quantiles (Carlier et al., 2016), and calibration in economics (Guo et al., 2021) and physics (Pollard & Windischhofer, 2022) suggesting cost-minimizing transport as a principled mechanism for enforcing calibration.

An alternative approach is to design iterative recalibration procedures that converge to fixed points, drawing on ideas from performative prediction (Perdomo et al., 2020). Resampling-based methods (Chung et al., 2024) offer a flexible starting point, but must account for the shifting pre-rank statistics induced by performativity.

Finally, calibrated multivariate forecasters may enable valid confidence sets that account for joint structure. Prior work in time-series conformal prediction (Sun & Yu, 2023; Angelopoulos et al., 2023) and multivariate quantiles (Garcin et al., 2023; Watts et al., 2024) offers tools for constructing such regions. Developing these methods is a promising direction for future work.

We believe these directions offer a foundation for building calibrated, interpretable, and practically useful multivariate forecasters.

References

Allen, S., Ziegel, J., and Ginsbourger, D. Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 150(760): 1315–1335, 2024.

Angelopoulos, A., Candes, E., and Tibshirani, R. J. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023.

Barbe, P., Genest, C., Ghoudi, K., and Rémillard, B. On kendall's process. *Journal of multivariate analysis*, 58 (2):197–229, 1996.

Belloni, A. and Winkler, R. L. On multivariate quantiles under partial orders. 2011.

Carlier, G., Chernozhukov, V., and Galichon, A. Vector quantile regression: an optimal transport approach. 2016.

Chung, Y., Char, I., and Schneider, J. Sampling-based multidimensional recalibration. In *Forty-first International Conference on Machine Learning*, 2024.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R. The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1): 243–262, 2004.

Fan, Y. and Henry, M. Vector copulas. *Journal of Econometrics*, 234(1):128–150, 2023.

Garcin, M., Guégan, D., and Hassani, B. A multivariate quantile based on kendall ordering. *Revstat-statistical journal*, 21(1):77–96, 2023.

Genest, C. and Rivest, L.-P. On the multivariate probability integral transformation. *Statistics & probability letters*, 53(4):391–399, 2001.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69 (2):243–268, 2007.

Guo, I., Loeper, G., Obloj, J., and Wang, S. Optimal transport for model calibration. *arXiv preprint arXiv:2107.01978*, 2021.

Knüppel, M., Krüger, F., and Pohle, M.-O. Score-based calibration testing for multivariate forecast distributions. *arXiv preprint arXiv:2211.16362*, 2022.

Kock, L., Rodrigues, G., Sisson, S. A., Klein, N., and Nott, D. J. Calibrated multivariate regression with localized pit mappings. *arXiv preprint arXiv:2409.10855*, 2024.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Pollard, C. and Windischhofer, P. Transport away your problems: Calibrating stochastic simulations with optimal transport. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1027:166119, 2022.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. Uncertainty quantification in complex simulation models using ensemble copula coupling. 2013.

Sun, S. H. and Yu, R. Copula conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Watts, A., Thompson, T., and Harvey, D. Confidence intervals on multivariate normal quantiles for environmental specification development in multi-axis shock and vibration testing. *arXiv preprint arXiv:2404.06565*, 2024.

Ziegel, J. F. and Gneiting, T. Copula calibration. 2014.

A. Appendix: Proofs

Lemma 4 (Approximate Calibration Implies Richness)**.** *Let $\mathcal{H}$ denote a space of continuous conditional forecasters $\hat{H}_{Y|X}$, and let $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a fixed projection function. For any $\varepsilon > 0$, the set of (g, ε)-calibrated forecasters contains uncountably many distinct distributions in $\mathcal{H}$.*

Proof. Let $Z = g(\hat{H}, Y)$ be the scalar pre-rank variable induced by the forecaster $\hat{H}$. (g, ε)-calibration only requires the PIT histogram of Z to be within ε of uniformity. Because g maps to a one-dimensional projection, many different forecasters $\hat{H}'$ can induce the same distribution over Z, even if their full joint distributions differ. In particular, we can perturb $\hat{H}$ along directions in the conditional distribution of $Y \mid X$ that leave the distribution of Z invariant or approximately invariant (within ε). Under mild continuity and support assumptions on $\hat{H}$, this defines an uncountable family of forecasters in $\mathcal{H}$ that all satisfy (g, ε)-calibration. □

Lemma 5 (The Uninformative Forecaster is g-Calibrated)**.** *Let $(X, Y) \sim H_{X,Y}$ be a joint distribution over covariates and outcomes. Define the* uninformative forecaster *as $\hat{H}_{Y|X} := H_Y$, i.e., a fixed forecast distribution equal to the marginal distribution of Y, independent of the input X. Then for any pre-rank function $g(\hat{H}, Y)$ that depends only on the forecaster $\hat{H}$ and the outcome Y, the forecaster $\hat{H}_{Y|X} = H_Y$ is g-calibrated.*

Proof. Let $g(\hat{H}, Y)$ be any scalar-valued pre-rank function that depends only on the forecaster and the outcome. Define:

$$Z := g(\hat{H}_{Y|X}, Y), \quad \hat{Z} := g(\hat{H}_{Y|X}, \hat{Y}) \quad \text{where } \hat{Y} \sim \hat{H}_{Y|X} = H_Y.$$

Since $\hat{H}_{Y|X} = H_Y$ is constant with respect to X, both Z and $\hat{Z}$ are functions of Y and $\hat{Y}$ respectively, but not of X. In particular, the distribution of $\hat{Z} \mid X$ is independent of X, and so is the conditional CDF:

$$\hat{H}_{Z|X}(z) = \mathbb{P}(g(H_Y, \hat{Y}) \le z \mid X) = \mathbb{P}(g(H_Y, \hat{Y}) \le z),$$

which is simply the marginal distribution of $\hat{Z}$.

Therefore, the PIT statistic $\hat{H}_{Z|X}(Z)$ is given by:

$$\hat{H}_{Z|X}(Z) = \mathbb{P}(g(H_Y, \hat{Y}) \le g(H_Y, Y)) = F_{g(H_Y, \hat{Y})}(g(H_Y, Y)),$$

where $F_{g(H_Y, \hat{Y})}$ denotes the CDF of $g(H_Y, \hat{Y})$.

Now note that both $Y \sim H_Y$ and $\hat{Y} \sim H_Y$, independently. So $g(H_Y, Y)$ and $g(H_Y, \hat{Y})$ are i.i.d. random variables with the same distribution. Hence, the rank of $g(H_Y, Y)$ among i.i.d. draws from that distribution is uniform:

$$\hat{H}_{Z|X}(Z) \sim \text{Uniform}[0, 1].$$

This proves that the uninformative forecaster is g-calibrated. □

B. Appendix B: Simple Pre-Rank Functions

While most definitions of multivariate calibration rely on pre-rank functions $g(\hat{H}, Y)$ that depend on the forecaster, a recent line of work introduces an alternative class of diagnostics that are *forecaster-independent* (Allen et al., 2024). These *simple pre-rank functions* are designed to depend only on the outcome Y, not on the covariates X or the forecast distribution $\hat{H}_{Y|X}$.

By construction, simple pre-rank functions are unaffected by recalibration. This avoids the core issue of performativity: since the diagnostic remains fixed as the forecaster is transformed, one-shot recalibration does not alter the calibration target. The PIT statistics based on these functions thus provide a stable diagnostic of multivariate forecast quality.

Common examples of simple pre-rank functions include:

- **Coordinate-wise projections:** $g(Y) = Y_i$ for some dimension i;
- **Norm-based scores:** $g(Y) = \|Y\|$, or $g(Y) = Y^\top A Y$ for a fixed positive semi-definite matrix A;

- **Mahalanobis distances:** $g(Y) = \sqrt{(Y-\mu)^\top \Sigma^{-1}(Y-\mu)}$, where μ, Σ are fixed.

Each simple pre-rank function captures a particular aspect of the multivariate distribution, such as location, spread, or marginal sharpness. To obtain a more complete picture of calibration behavior, it is often necessary to evaluate multiple simple pre-rank functions in parallel. This diagnostic strategy is used in Allen et al. (2024), where failure in any one function may indicate a specific mode of miscalibration.

While simple pre-rank functions offer a valuable tool for assessing multivariate calibration without inducing performativity, they are not expressive enough on their own to enforce or characterize full joint calibration. An interesting direction for future work is the design of recalibration procedures that can *simultaneously* enforce calibration across a collection of simple pre-rank functions. Doing so could provide a practical and robust alternative to fully forecaster-dependent diagnostics, while retaining the benefits of stability and interpretability.