Calibrated Multivariate Distributional Regression with Pre-Rank Regularization

Aya Laajil **Elnura Zhalieva** **Naomi Desobry** **Souhaib Ben Taieb**

Department of Statistics and Data Science, Mohamed Bin Zayed University of Artificial Intelligence

Abstract

The goal of probabilistic prediction is to issue predictive distributions that are as informative as possible, subject to being *calibrated*. Whereas univariate calibration is well understood, achieving multivariate calibration remains challenging. Recent work has introduced pre-rank functions, scalar projections of multivariate predictions and observations, as diagnostics for multivariate calibration, but they are mainly used for post-hoc evaluation. We propose a regularization-based calibration method that enforces multivariate calibration during training of multivariate distributional regression models using pre-rank functions. We further introduce a novel PCA-based pre-rank that projects predictions onto principal directions of the predictive distribution. Through simulations and experiments on 18 real-world datasets, the proposed approach improves multivariate pre-rank calibration without compromising predictive accuracy, and our PCA pre-rank reveals dependence-structure misspecifications that are not detected by existing pre-ranks.

1 INTRODUCTION

Probabilistic models aim to quantify uncertainty by issuing full predictive distributions rather than point estimates. For such models to be reliable in downstream decision-making, their predictive distributions must be *calibrated*, meaning statistically consistent with observed outcomes (Gneiting et al., 2007; Gneiting, 2014). While univariate calibration is well understood, achieving and enforcing calibration in multivariate settings remains substantially more challenging due to the need to capture joint dependence structures. In practice, probabilistic models are typically trained by minimizing strictly proper scoring rules. However, minimizing a proper scoring rule does not guarantee calibration, particularly under model misspecification or limited data (Guo et al., 2017; Buchweitz et al., 2025; Wessel et al., 2025).

Although multivariate distributional regression models are widely used (Klein, 2024), calibration is still most often assessed only marginally, which does not ensure that the joint dependence structure is reliably calibrated, even when each marginal appears well calibrated (Gneiting et al., 2008) (see Figure 1).

Existing work has developed post-hoc diagnostics for multivariate calibration, most notably multivariate rank histograms and pre-rank functions (Allen et al., 2023). A pre-rank function maps a multivariate forecast-observation pair to a univariate summary that captures a specific aspect of forecast behavior. They may be chosen to assess different mis-specifications of the predictive distribution. However, these approaches are designed primarily for evaluation.

In this paper, we go beyond post-hoc diagnostics, and introduce a general framework that *enforces* multivariate calibration into the training of distributional regression models. Within this framework, we further propose a *new* pre-rank function based on principal components analysis (PCA), which assess calibration along the principal directions of the predictive distribution, making it sensitive to dependence-structure misspecifications that existing pre-ranks may fail to detect. Together, these contributions enable enforcement of multivariate calibration during training without compromising predictive performance.

2 BACKGROUND

We consider the problem of multivariate distributional regression, where the target variable $Y \in \mathcal{Y} \subseteq \mathbb{R}^D$ depends on an input variable $X \in \mathcal{X} \subseteq \mathbb{R}^L$. Let

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

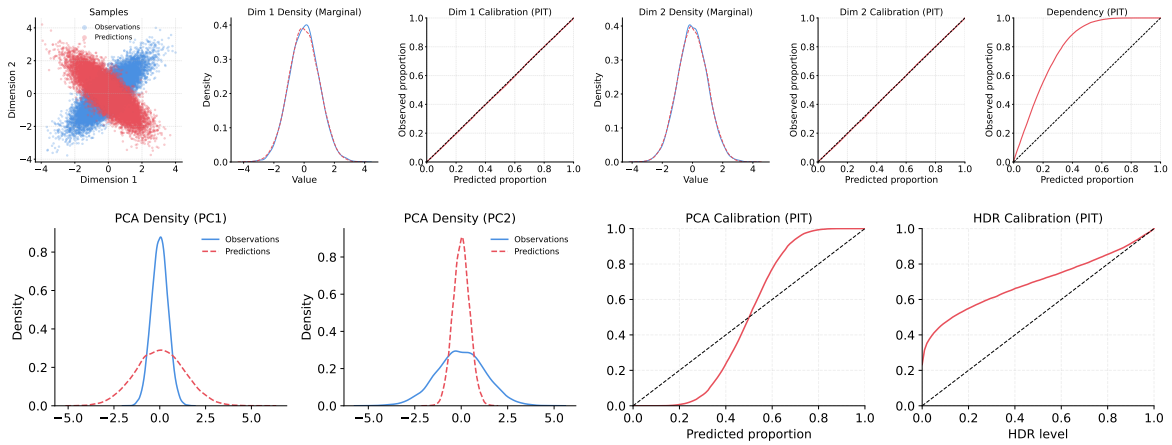


Figure 1: **Marginal calibration is insufficient for multivariate predictive distributions.**(Top) The model is well-calibrated marginally but fails to capture dependence, as shown by the mismatch in sample geometry and the strongly miscalibrated dependency PIT.(Bottom) PCA-based diagnostics reveal miscalibration in projected subspaces, highlighting that calibration assessed only marginally can miss joint errors.

$\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$ be a dataset such that $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} F_X \times F_{Y|X}$, where $F_{Y|X}$ denotes the true conditional distribution of Y given X. The goal is to learn a parametric predictive distribution $\hat{F}_{Y|X}$ that approximates $F_{Y|X}$, and we assume that for any input X, predictive samples can be drawn as $\hat{Y} \mid X \sim \hat{F}_{Y|X}(\cdot \mid X)$.

Univariate calibration. For $D = 1$, **probabilistic calibration** is defined via the probability integral transform (PIT) (Gneiting et al., 2007). Given a predictive distribution $\hat{F}_{Y|X}$, the PIT is

$$Z = \hat{F}_{Y|X}(Y \mid X).$$

The model is calibrated if $Z \sim \mathcal{U}(0,1)$ where $\mathcal{U}(0,1)$ is the uniform distribution. Deviations from uniformity indicate miscalibration (Wilks, 2018).

Probabilistic Calibration Error (PCE). Given PIT values $\{Z_i\}_{i=1}^N$ where $Z_i = \hat{F}_{Y|X=X_i}(Y_i)$, calibration can be assessed by comparing the *empirical* CDF

$$\hat{F}_Z(\alpha) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{Z_i \le \alpha\}, \qquad \alpha \in (0,1), \tag{1}$$

to the uniform CDF. A representative scalar discrepancy is the Probabilistic Calibration Error (PCE), defined on a grid $\{\alpha_j\}_{j=1}^M \subset [0,1]$ as

$$\text{PCE}(\hat{F}_{Y|X}) := \frac{1}{M} \sum_{j=1}^{M} \left| \alpha_j - \hat{F}_Z(\alpha_j) \right|, \tag{2}$$

While effective for post-hoc evaluation, such indicator-based estimators are not directly suitable as training objectives due to non-differentiability and sensitivity to discretization (Kumar et al., 2019; Dheur and Ben Taieb, 2023).

Multivariate calibration. When the target variable is multivariate ($D > 1$), defining and assessing probabilistic calibration becomes substantially more challenging. Calibrating marginals alone is insufficient to ensure reliable joint predictions, as it does not involve the dependence structure of the predictive distribution (Ziegel and Gneiting, 2014), as illustrated in Figure 1..

A general strategy to assess multivariate calibration is to reduce the multivariate problem to a collection of univariate ones by means of scalar projections. Following Allen et al. (2023), a *pre-rank* is a function

$$\rho : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}, \quad (y, x) \mapsto \rho(y, x),$$

that maps a multivariate outcome y, for a given input x, to a scalar summary.

Given ρ, we define $T = \rho(Y, X)$ and $\hat{T} = \rho(\hat{Y}, X)$ with $\hat{Y} \mid X \sim \hat{F}_{Y|X}$. The associated projected PIT is

$$Z_\rho := F_{\hat{T}|X}(T \mid X) \tag{3}$$

The model is **calibrated with respect to a pre-rank function** ρ if the projected PIT Z_ρ is uniformly distributed on $[0,1]$.

3 A NEW MULTIVARIATE CALIBRATION METHOD

We introduce a new regularization-based framework that enforces multivariate calibration during training. Following Wilks (2018) and Wessel et al. (2025), who propose regularization-based calibration methods in the univariate setting, we aim to enforce calibration for

multivariate predictions by augmenting the loss function with a regularization term based on a miscalibration measure.

Let Z_ρ, as defined in (3), denote the projected PIT associated with a pre-rank ρ. We define the regularizer as

$$\mathcal{R}(\hat{F}_{Y|X}; \rho) = \frac{1}{M} \sum_{j=1}^{M} \left| \alpha_j - F_{Z_\rho}(\alpha_j) \right|, \tag{4}$$

where $\{\alpha_j\}_{j=1}^M \subset [0,1]$ is a fixed grid of quantile levels.

In practice, F_{Z_ρ} is unknown and must be estimated from data. A natural empirical estimator is given by the empirical CDF

$$\hat{F}_{Z_\rho}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(Z_{\rho,i} \leq \alpha), \tag{5}$$

where $Z_{\rho,i} = F_{\hat{T}_i|X=X_i}(T_i \mid X = X_i)$, $T_i = \rho(Y_i, X_i)$ and $\hat{T}_i = \rho(\hat{Y}_i, X_i)$. Substituting this estimator into (4) yields an empirical PCE, which is not differentiable.

Differentiable Regularizer. To enable gradient-based optimization, we adopt the kernel-smoothed PCE (PCE-KDE) surrogate of Dheur and Ben Taieb (2023), replacing the indicator function by a smooth approximation using a logistic kernel, yielding the smoothed CDF

$$\Phi_{\text{KDE}}(\alpha_j, \{Z_{\rho,i}\}_{i=1}^N) = \frac{1}{N} \sum_{i=1}^{N} \sigma(\tau(\alpha_j - Z_{\rho,i})), \tag{6}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function and $\tau > 0$ controls the smoothness of the approximation.

The resulting regularizer is

$$\mathcal{R}_{\text{KDE}}(\hat{F}_{Y|X}, \rho) = \frac{1}{M} \sum_{j=1}^{M} \left| \alpha_j - \Phi_{\text{KDE}}(\alpha_j, \{Z_{\rho,i}\}_{i=1}^N) \right|^p, \tag{7}$$

with $p \geq 1$ determines the shape of the penalty.

Training objective. Let S denote a strictly proper scoring rule used to train a multivariate distributional regression model. Given a pre-rank ρ, we define the augmented objective as

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} S\Big(\hat{F}_{Y|X=X_i}, Y_i\Big) + \lambda \, \mathcal{R}_{\text{KDE}}(\hat{F}_{Y|X}, \rho), \tag{8}$$

where $\lambda \geq 0$ controls the strength of calibration enforcement. Setting $\lambda = 0$ recovers standard unregularized training based solely on the scoring rule. Algorithm 1 in Appendix B details the training process with the augmented loss.

Choice of pre-rank. Our framework is agnostic to the choice of pre-rank. We include different pre-ranks proposed in the literature (Scheuerer and Hamill, 2015; Knüppel et al., 2022; Allen et al., 2023), each targeting a specific aspect of multivariate miscalibration. We consider marginal, location, scale, dependence-based, copula-based (Ziegel and Gneiting, 2014), and high-density region (HDR) pre-ranks (Chung et al., 2024). While copula- and HDR-based recalibration are post-hoc methods, we adapt them to our framework by defining their respective pre-rank functions. Pre-rank definitions are in Table 1 in Appendix B.

A new PCA-based pre-rank. We introduce a *new* pre-rank function based on principal component analysis (PCA) that aims to detect miscalibration along the principal directions of variation of the predictive distribution. These directions correspond to projections with maximal predictive variance, concentrating on components where uncertainty is largest and calibration errors are most detectable.

Given (x, y), we draw samples from $\hat{F}_{Y|X=x}$ and estimate the predictive covariance $\hat{\Sigma}_x$. Let $v_k(x)$ denote the k-th principal component of $\hat{\Sigma}_x$. The corresponding PCA pre-rank is

$$\rho_k^{\text{PCA}}(y, x) = v_k(x)^\top y. \tag{9}$$

Algorithm 2 in Appendix C summarizes computation of the k-th PCA pre-rank.

From a ranking perspective, PCA satisfies several desirable invariance and symmetry properties. It is equivariant under centering, linear re-scaling, and orthogonal transformations, and is permutation-equivariant with respect to coordinate relabeling. As a covariance-based method, PCA depends only on second-order structure, which yields a computationally stable diagnostic in high-dimensional settings. Unlike other linear projections, PCA is optimal in maximizing projected variance (Jolliffe, 2002). Consequently, the leading principal components concentrate directions along which covariance misspecification is most pronounced, making them particularly informative for calibration assessment under second-order model errors.

4 EXPERIMENTS

Simulated Data. We revisit the simulation framework of Thorarinsdottir and Schuhen (2018) and Allen et al. (2023) to analyze how different pre-ranks detect different forms of multivariate misspecifications, motivating our proposed PCA pre-rank.

We consider a multivariate Gaussian data generating process and construct predictive distributions with

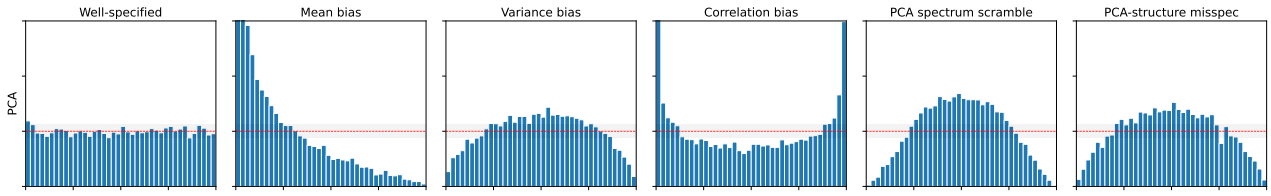
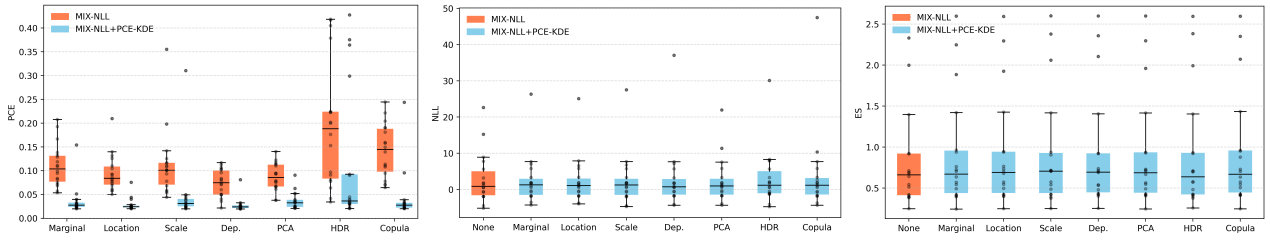


Figure 2: Projected PIT histograms for **PCA pre-rank** under different misspecifications of a Multivariate Gaussian.



(a) PCE (b) NLL (c) Energy Score

Figure 3: Boxplots of PCE, NLL, and ES over 18 datasets comparing the unregularized model (baseline) and our regularized model. In (b) and (c), the "None" box refers to the unregularized model trained without pre-rank.

controlled misspecifications, including mean bias, variance inflation, and covariance distortions that preserve marginals but alter eigenstructure. We repeat the same simulation with a spatial Gaussian random field. More details on the experimental setup and data generation processes are provided in Appendices D.1 and D.2.

Figure 2 shows the projected PIT histograms of our PCA pre-rank for the multivariate Gaussian simulation as a representative example. The corresponding results for all pre-ranks in both simulation settings are presented in Figures 4 and 5 in Appendix D.2. The results show that marginal, location, scale, and HDR pre-ranks primarily detect misspecifications in the marginals but fail to capture distortions in dependence or eigenstructure. The dependency pre-rank detects misspecifications in dependence structure but does not capture biases in mean and variance. When the marginals are calibrated but the covariance eigenstructure is distorted, most pre-ranks appear approximately uniform. In contrast, as shown in Figure 2, our PCA pre-rank reveals clear deviations from uniformity in this setting. This indicates sensitivity to misspecifications in the principal directions of variation.

Real-world Data. We evaluate the impact of our pre-rank based regularization framework on 18 multi-output regression datasets (Camehl et al., 2025; Dheur and Taieb, 2025), with output dimension D between 2 and 16. The unregularized model is a mixture of multivariate Gaussians trained using NLL, while the regularized model augments this objective with our pre-rank regularizer defined in (8). Calibration is assessed via PCE defined in (2), and predictive performance is evaluated via NLL and the Energy Score (ES). More details are provided in Appendix E. Figure 3 compares the unregularized baseline with our regularized model. Across datasets and pre-ranks, our regularization consistently reduces PCE while maintaining comparable NLL and ES to the unregularized model, demonstrating that improved calibration does not come at the expense of predictive performance. Tables 5–8 in Appendix E.1 give more detailed numerical results.

Calibration across Pre-Ranks. To assess whether the effect of regularization generalizes beyond the pre-rank used during training, we evaluate calibration under multiple evaluation pre-ranks while training with a single pre-rank regularizer. Figures 6 and 7 in Appendix E.1 show that our pre-rank based regularization improves multivariate calibration not only for the pre-rank used during training, but also across several other evaluation pre-ranks. Thus, the benefits are not confined to a single calibration aspect, but extend across multiple facets of multivariate calibration.

5 Conclusion

We introduce a new multivariate calibration framework for distributional regression by augmenting the learning objective with a differentiable pre-rank regularizer. We further propose a novel PCA-based pre-rank that assesses calibration along principal components of the predictive distribution, providing sensitivity to dependence-structure misspecifications that may remain undetected by existing pre-ranks. Across 18 datasets, our framework improves multivariate calibration without compromising predictive performance.

References

Allen, S., Ziegel, J. F., and Ginsbourger, D. (2023). Assessing the calibration of multivariate probabilistic forecasts. *arXiv preprint arXiv:2307.05846*.

Buchweitz, E., Romano, J. V., and Tibshirani, R. J. (2025). Asymmetric penalties underlie proper loss functions in probabilistic forecasting. *arXiv preprint arXiv:2505.00937*.

Camehl, A., Fok, D., and Gruber, K. (2025). On superlevel sets of conditional densities and multivariate quantile regression. *Journal of Econometrics*, 249:105807.

Chung, Y., Char, I., and Schneider, J. (2024). Sampling-based multidimensional recalibration. In *Proceedings of the International Conference on Machine Learning*.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A*, 147(2):278–292.

Dheur, V. and Ben Taieb, S. (2023). A large-scale study of probabilistic calibration in neural network regression. *arXiv preprint arXiv:2306.02738*.

Dheur, V. and Taieb, S. B. (2025). Multivariate latent recalibration for conditional normalizing flows. *arXiv [cs.LG]*.

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

Feldman, S., Bates, S., and Romano, Y. (2022). Calibrated multiple-output quantile regression with representation learning.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2018). Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12):4323–4338.

Gneiting, T. (2014). *Calibration of medium-range weather forecasts*. European Centre for Medium-Range Weather Forecasts Reading, UK.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities. *TEST*, 17(2):211–235.

Guan, L. (2021). Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*. PMLR.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M. C., and Roelofs, B. (2021). Soft calibration objectives for neural networks.

Klein, N. (2024). Distributional regression for data analysis. *Annual Review of Statistics and Its Application*, 11(2024):321–346.

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

Kumar, A., Sarawagi, S., and Jain, U. (2019). Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Scheuerer, M. and Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*.

Thorarinsdottir, T. L. and Schuhen, N. (2018). Verification: Assessment of calibration and accuracy. In *Statistical postprocessing of ensemble forecasts*, pages 155–186. Elsevier.

Wang, Z., Gao, R., Yin, M., Zhou, M., and Blei, D. M. (2022). Probabilistic conformal prediction using conditional random samples.

Wessel, J. B., Schillinger, M., Kwasniok, F., and Allen, S. (2025). Enforcing tail calibration when training probabilistic forecast models. *arXiv preprint arXiv:2506.13687*.

Wilks, D. S. (2018). Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society*, 144(710):76–84.

Winkler, R. L., Muñoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M., and Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *TEST*, 5:1–60.

Ziegel, J. F. and Gneiting, T. (2014). Copula calibration. *Bernoulli*, 20(2):934–965.

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. Yes

(b) Complete proofs of all theoretical results. Yes

(c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes. We will release the code after acceptance.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not applicable

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. Yes

(b) The license information of the assets, if applicable. Yes

(c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable

(d) Information about consent from data providers/curators. We use data from open source benchmarks.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Materials

A Related Work

Calibration in probabilistic regression has been studied extensively in the univariate setting (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007). While strictly proper scoring rules such as the negative log-likelihood (NLL) or the continuous ranked probability score (CRPS) are widely used for training probabilistic models (Winkler et al., 1996; Gneiting and Raftery, 2007), they do not guarantee calibrated predictions. This has motivated methods that explicitly encourage calibration during training. Wilks (2018) introduces calibration penalties for ensemble post-processing, and more recently, Wessel et al. (2025) propose training-time regularization schemes that enforce calibration with respect to specific aspects of the predictive distribution, such as tail behavior. Related work has also explored regularization and loss-based approaches to improve calibration in univariate regression (Gebetsberger et al., 2018; Buchweitz et al., 2025). These methods, however, are fundamentally designed for scalar targets and do not address calibration of joint predictive distributions for multivariate outcomes.

Extending calibration notions to multivariate regression is considerably more challenging due to the need to account for dependence structures across target dimensions. Early work by Gneiting et al. (2008) introduces multivariate rank histograms as diagnostic tools for assessing probabilistic calibration of multivariate forecasts. Copula-based notions of calibration further formalize joint calibration by evaluating the uniformity of the copula probability integral transform (Ziegel and Gneiting, 2014). More recently, Allen et al. (2023) propose a unifying diagnostic framework based on pre-rank functions. Their work demonstrates that different pre-ranks probe complementary aspects of multivariate miscalibration, such as location, scale, or dependence, but focuses exclusively on evaluation rather than enforcement. Futher approaches aim to improve multivariate calibration via post-hoc recalibration. Copula-based recalibration methods adjust joint distributions after training, but typically require access to the full joint CDF and are not easily integrated into gradient-based learning pipelines. Sampling-based methods such as HDR recalibration (Chung et al., 2024) leverage density-level projections to recalibrate predictive samples with respect to highest density regions, yielding joint calibration guarantees in a post-processing step. While effective in some settings, these approaches depend critically on the quality of the estimated predictive density and are not designed to influence the training dynamics of the underlying model.

B Background

B.1 Pre-rank functions

Practical note on Copula-based Pre-Rank. When using copula-based pre-ranks, one requires access to the joint CDF $\hat{F}_{Y|X}(y)$, i.e., the probability that all components of Y are less than or equal to y given X. However, many models provide only the conditional density $\hat{f}_{Y|X}$.

We approximate the joint CDF via Monte Carlo sampling. Given input X_i and target Y_i, we draw S samples $\hat{Y}_{i,1}, \ldots, \hat{Y}_{i,S} \sim \hat{f}_{Y|X=X_i}$ and estimate:

$$\hat{F}_{Y|X=X_i}(y) \;\approx\; \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\left\{\hat{Y}_{i,s} \leq y\right\}, \tag{10}$$

where the indicator $\mathbf{1}\left\{\hat{Y}_{i,s} \leq y\right\}$ is true if and only if $\hat{Y}_{i,s}^{(d)} \leq y^{(d)}$ for all components $d = 1, \ldots, D$.

Since the indicator function is not differentiable, we replace it with a smooth approximate using the sigmoid function $\sigma(z) = 1/(1+e^{-z})$ and a temperature parameter $\tau > 0$. This gives:

$$\hat{F}_{Y|X=X_i}(y) \;\approx\; \frac{1}{S}\sum_{s=1}^{S}\prod_{d=1}^{D} \sigma\left(\tau\left(y^{(d)} - \hat{Y}_{i,s}^{(d)}\right)\right), \tag{11}$$

where $y^{(d)}$ and $\hat{Y}_{i,s}^{(d)}$ denote the d-th components of the vectors y and $\hat{Y}_{i,s}$, respectively. The product over dimensions enforces that all components of $\hat{Y}_{i,s}$ fall below the threshold y_i, mimicking the joint indicator condition.

This smooth approximation is fully differentiable with respect to the model parameters (via the samples $\hat{Y}_{i,s}$), and thus compatible with gradient-based optimization routines such as backpropagation.

Computation of Energy Score We use Energy Score (ES) as a scoring rule metric to evaluate our model performance. ES generalizes Continuous Ranked Probability Score (CRPS) to multivariate settings and is computed empirically as:

$$\mathrm{ES}(\hat{F}, y) = \frac{1}{G}\sum_{i=1}^{G}\|\hat{Y}_i - y\| - \frac{1}{2G^2}\sum_{i=1}^{G}\sum_{j=1}^{G}\|\hat{Y}_i - \hat{Y}_j\| \tag{12}$$

where $\{\hat{Y}_i\}_{i=1}^{G} \sim \hat{F}_{Y|X}$ are G samples drawn from the predictive distribution. We set $G = 100$ in all experiments.

Equivalence of Pre-Rank Calibration.

Proposition B.1 For every fixed $x \in \mathbb{R}^L$, the function $y \mapsto \rho_2(y, x)$ must be a strictly monotonic bijective transformation of $y \mapsto \rho_1(y, x)$. That is, there exists a strictly increasing or decreasing bijection h_x such that for all $y \in \mathbb{R}^D$,

$$\rho_2(y, x) = h_x(\rho_1(y, x)).$$

Proof. Fix $x \in \mathbb{R}^L$ and define $T_1 = \rho_1(Y, x)$ and $T_2 = \rho_2(Y, x) = h_x(T_1)$, where h_x is a strictly monotonic bijection. Let $\hat{F}_{T_1|X=x}$ and $\hat{F}_{T_2|X=x}$ denote the empirical conditional CDFs of T_1 and T_2, respectively, estimated using the same sample of predicted values $\{\hat{Y}_i\}_{i=1}^{N'}$ drawn from the predictive distribution $\hat{F}_{Y|X=x}$.

As explained in the background section, we estimate these conditional CDFs using the empirical estimator. Since this construction is used solely for evaluation, differentiability of the CDF is not required. Then for any $t \in \mathbb{R}$,

$$\begin{aligned}
\hat{F}_{T_2|X=x}(t) &= \frac{1}{N'}\sum_{i=1}^{N'}\mathbf{1}_\tau(\rho_2(\hat{Y}_i, x) \le t) \\
&= \frac{1}{N'}\sum_{i=1}^{N'}\mathbf{1}_\tau(\rho_1(\hat{Y}_i, x) \le h_x^{-1}(t)) \\
&= \hat{F}_{T_1|X=x}(h_x^{-1}(t)).
\end{aligned}$$

Since $T_2 = h_x(T_1)$ and

$$\hat{F}_{T_2|X=x}(t) = \hat{F}_{T_1|X=x}(h_x^{-1}(t)),$$

we have:

$$\hat{F}_{T_2|X=x}(T_2) = \hat{F}_{T_1|X=x}(h_x^{-1}(T_2)) = \hat{F}_{T_1|X=x}(T_1),$$

where we used the fact that $T_1 = h_x^{-1}(T_2)$ by construction. It follows that the PIT value computed under ρ_2 coincides with the one computed under ρ_1:

$$U_2 := \hat{F}_{Z_2|X=x}(Z_2) = \hat{F}_{Z_1|X=x}(Z_1) =: U_1.$$

It follows that U_1 and U_2 have the same distribution. In particular,

$$U_2 \sim \mathcal{U}[0,1] \quad \Longleftrightarrow \quad U_1 \sim \mathcal{U}[0,1],$$

which establishes the equivalence of the two calibration criteria under the assumed transformation.

Flexibility and Complementarity of Pre-ranks. As emphasized by Allen et al. 2023, pre-rank functions offer flexibility for probing different aspects of multivariate calibration. Each pre-rank targets a specific property of the predictive distribution, such as location, scale, or dependence, allowing practitioners to focus on the calibration aspects most relevant to their application.

Pre-rank	Formula
Marginal (d)	$\rho^d_{\text{marg}}(y, x) = y_d$
Location	$\rho_{\text{loc}}(y, x) = \frac{1}{D}\sum_{d=1}^{D} y_d$
Scale	$\rho_{\text{scale}}(y, x) = \frac{1}{D}\sum_{d=1}^{D}(y_d - \bar{y})^2$
Dependency (h)	$\rho_{\text{dep}}(y, x; h) = -\gamma_y(h)/s_y^2$
HDR	$\rho_{\text{hdr}}(y, x) = \hat{f}_{Y\|X=x}(y)$
Copula	$\rho_{\text{cop}}(y, x) = \hat{F}_{Y\|X=x}(y)$

Table 1: Pre-rank functions. Here $y = (y_1, \ldots, y_D) \in \mathbb{R}^D$ denotes the multivariate target, $\bar{y} = \frac{1}{D}\sum_{d=1}^{D} y_d$, $d \in \{1, \ldots, D\}$ indexes the output dimensions, $\gamma_y(h) = \frac{1}{2(D-h)}\sum_{d=1}^{D-h} |y_d - y_{d+h}|^2$ is a variogram-based dependency measure with lag $h \in \{1, \ldots, D-1\}$, and s_y^2 denotes the empirical variance across coordinates.

To illustrate what each pre-rank can and cannot detect under different forms of distributional misspecification, we conduct a simulation study. The resulting PIT histograms, reported in D.2, are straightforward to interpret and make explicit which systematic deficiencies a given pre-rank is sensitive to. Deviations from uniformity directly reveal specific forms of miscalibration, such as biases in location, dispersion, or dependence structure, thereby facilitating a clear and interpretable assessment of predictive quality.

C Algorithms

Algorithm 1 Training with Pre-rank Regularization

Input: data $\{(x_i, y_i)\}_{i=1}^N$, model $\hat{F}^{\theta}_{Y|X}$, scoring rule S, pre-rank ρ, $\lambda \geq 0$, M, $\tau > 0$, $p \geq 1$, learning rate η, epochs T
for $t = 1, \ldots, T$ **do**
  for minibatch $\mathcal{B} = \{(y_i, x_i)\}_{i=1}^B$ **do**
    $T_i \leftarrow \rho(x_i, y_i), \quad \forall i \in [B]$
    $\hat{Y}_i^{(m)} \sim \hat{F}^{\theta_t}_{Y|X=x_i}, \quad \forall i \in [B], m = 1, \ldots, M$
    $\hat{T}_i^{(m)} \leftarrow \rho(\hat{Y}_i^{(m)}, x_i) \quad \forall i \in [B], m = 1, \ldots, M$
    $Z_{\rho,i} \leftarrow \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{\hat{T}_i^{(m)} \leq T_i\}, \quad \forall i \in [B]$
    for $j = 1, \ldots, M$ **do**
      $\Phi_{\text{KDE}}(\alpha_j) \leftarrow \frac{1}{B}\sum_{i=1}^{B} \sigma\big(\tau(\alpha_j - Z_{\rho,i})\big)$
    end for
    $\mathcal{R}_{\text{PCE-KDE}} \leftarrow \frac{1}{M}\sum_{j=1}^{M} \big|\alpha_j - \Phi_{\text{KDE}}(\alpha_j)\big|^p$
    $\mathcal{L} \leftarrow \frac{1}{B}\sum_{i=1}^{B} S(\hat{F}^{\theta_t}_{Y|X=x_i}, y_i) + \lambda\, \mathcal{R}_{\text{PCE-KDE}}$
    $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} \mathcal{L}$
  end for
end for
Return: $\hat{F}^{\theta_T}_{Y|X}$

Algorithm 2 PCA Pre-Rank $\rho_k^{\text{PCA}}(x, y)$

Input: Observation $y \in \mathbb{R}^D$, samples $\{\hat{Y}^{(m)}\}_{m=1}^M$ with $\hat{Y}^{(m)} \sim \hat{F}_{Y|X=x}$, index $k \leq \min(D, M-1)$
Compute the covariance of $\{\hat{Y}^{(m)}\}_{m=1}^M$
Compute the k-th principal component u_k
$T_k \leftarrow \langle y, u_k \rangle$
Output: T_k

D Simulation Studies

D.1 Misspecifications

Let $\Sigma \in \mathbb{R}^{D\times D}$ be a symmetric positive definite covariance matrix.

PCA Spectrum Scramble. Let the eigendecomposition of Σ be : $\Sigma = U\Lambda U^\top, \Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_D), \quad \lambda_1 \geq \cdots \geq \lambda_D > 0$. Define the reversed spectrum $\Lambda^{\mathrm{rev}} := \mathrm{diag}(\lambda_D,\dots,\lambda_1)$, and for a scrambling parameter $\gamma \in [0,1]$, define $\tilde{\Lambda}(\gamma) := (1-\gamma)\Lambda + \gamma\Lambda^{\mathrm{rev}}$. To preserve total variance, renormalize : $\tilde{\Lambda}(\gamma) \leftarrow \tilde{\Lambda}(\gamma)\cdot \frac{\mathrm{tr}(\Lambda)}{\mathrm{tr}(\tilde{\Lambda}(\gamma))}$. The resulting *PCA spectrum–scrambled covariance* is $\Sigma_{\mathrm{scr}}(\gamma) := U\,\tilde{\Lambda}(\gamma)\,U^\top$.

PCA-Structure Misspecification. Let $e := \frac{1}{\sqrt{D}}(1,\dots,1)^\top$, denote the mean direction, and let $V = [e, v_2, \dots, v_D] \in \mathbb{R}^{D\times D}$, be any orthonormal basis with first column e.

Express Σ in this basis: $S := V^\top \Sigma V = \begin{pmatrix} s_{11} & s_{12}^\top \\ s_{12} & S_\perp \end{pmatrix}, \qquad S_\perp \in \mathbb{R}^{(D-1)\times(D-1)}$. Let the eigendecomposition of $S_\perp$ be $S_\perp = W\,\mathrm{diag}(\mu_1,\dots,\mu_{D-1})\,W^\top, \qquad \mu_1 \geq \cdots \geq \mu_{D-1}$. For parameters $c > 1$ and $k \leq D-1$, define distorted eigenvalues $\tilde{\mu}_i = \begin{cases} c\,\mu_i, & i = 1,\dots,k, \\ \mu_i/c, & i = D-k,\dots,D-1, \\ \mu_i, & \text{otherwise.} \end{cases}$

Define : $\tilde{S}_\perp := W\,\mathrm{diag}(\tilde{\mu}_1,\dots,\tilde{\mu}_{D-1})\,W^\top, \qquad \tilde{S} := \begin{pmatrix} s_{11} & s_{12}^\top \\ s_{12} & \tilde{S}_\perp \end{pmatrix}.$

The resulting *PCA-structure-misspecified covariance* is $\Sigma_{\mathrm{pca}} := V\,\tilde{S}\,V^\top$. By construction, $e^\top\Sigma_{\mathrm{pca}}e = e^\top\Sigma e$, while the covariance structure in the subspace orthogonal to e is distorted.

Isotropy Misspecification. Let $\{s_i = (x_i, y_i)\}_{i=1}^D \subset \mathbb{R}^2$ denote spatial grid locations and let

$$\Sigma_{ij} = \sigma^2 \exp\left(-\frac{\|s_i - s_j\|_2}{\tau}\right)$$

be the isotropic exponential covariance. To introduce an example of a geometric isotropy misspecification, we rescale one spatial axis before computing distances. For a scaling factor $\alpha > 1$, define the transformed coordinates

$$\tilde{s}_i := (x_i,\ \alpha y_i),$$

and the corresponding covariance

$$\Sigma_{ij}^{\mathrm{aniso}} := \sigma^2 \exp\left(-\frac{\|\tilde{s}_i - \tilde{s}_j\|_2}{\tau}\right).$$

This construction preserves marginal variances while inducing direction-dependent correlation decay. Such misspecification is invisible to purely marginal diagnostics but alters the spatial dependence structure and principal directions of variation.

PC Anisotropy Flip. Let $\Sigma \in \mathbb{R}^{D\times D}$ be symmetric positive definite with eigendecomposition

$$\Sigma = U\Lambda U^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_D), \quad \lambda_1 \geq \cdots \geq \lambda_D > 0.$$

We define an example of covariance misspecification that assigns strong variance to the *wrong* principal directions while preserving total variance.

First, reverse the eigenvalue ordering, $\Lambda^{\mathrm{rev}} := \mathrm{diag}(\lambda_D,\dots,\lambda_1)$,

and introduce anisotropy by amplifying the leading and shrinking the trailing components. For an axis ratio $a > 1$ and integers $k \leq D/2$, define $\tilde{\lambda}_i = \begin{cases} a\,\lambda_{D-i+1}, & i = 1,\dots,k, \\ \lambda_{D-i+1}, & k < i < D-k, \\ \lambda_{D-i+1}/a, & i = D-k+1,\dots,D. \end{cases}$

To preserve total variance, renormalize $\tilde{\Lambda} \leftarrow \tilde{\Lambda} \cdot \frac{\mathrm{tr}(\Lambda)}{\mathrm{tr}(\tilde{\Lambda})}$

Finally, to further misalign principal directions, we apply an orthogonal rotation R acting in the $(1,2)$ principal subspace (e.g. a $\pi/2$ rotation), and define $\Sigma_{\text{flip}} := (UR)\,\tilde{\Lambda}\,(UR)^\top$. This *PC anisotropy flip* preserves marginal variances and overall energy but assigns large variance to orthogonal directions relative to the true dominant modes. It is therefore detectable by PCA-based preranks, while potentially evading mean- or marginal-based calibration diagnostics.

D.2 Simulation studies

For a well-calibrated predictive distribution, the PIT is uniformly distributed on $[0,1]$, yielding a flat histogram. Deviations from uniformity indicate specific forms of miscalibration: U-shaped histograms (underdispersion), inverted-U shapes (overdispersion), and skewness (bias).

Simulation 1 (Multivariate Gaussian). We first study a synthetic $d = 10$ dimensional multivariate Gaussian setting in which the true data-generating distribution is $y \sim \mathcal{N}(0, \Sigma_{\text{true}})$ with an exponential covariance on a one-dimensional index set, $(\Sigma_{\text{true}})_{ij} = \exp(-|i-j|/\tau)$ (unit marginal variance and $\tau = 1$). We draw $N = 10{,}000$ forecast cases and, for each case, generate an ensemble of size $M = 20$ from a misspecified predictive distribution $F_{\text{pred}} = \mathcal{N}(\mu_{\text{pred}}, \Sigma_{\text{pred}})$. We consider standard distributional errors that target low-order moments—constant mean shift ($\mu_{\text{pred}} = 0.5$), variance inflation (Σ_{pred} scaled by 1.75), and correlation-range misspecification (shorter range, $\tau = 0.3$), as well as two PCA-focused covariance perturbations: a *PCA spectrum scramble* that alters the eigenvalue profile of Σ_{true} (while preserving its eigenvectors), and a *PCA-structure* misspecification that preserves marginal variances but rotates principal directions. We evaluate prerank-based projected PIT histograms across these cases, illustrating how different preranks selectively detect distinct types of multivariate misspecification beyond marginal calibration.

Table 2: Mis-specifications considered in Simulation 1 (Multivariate Gaussian). λ^π denotes a trace-preserving interpolation between the eigenvalue spectrum of Σ and its reversed ordering, and Q is an orthogonal matrix that preserves variance along the mean direction.

Mis-specification	μ	σ^2	τ
Mean bias	± 0.5	1	1
Variance bias	0	± 0.75	1
Correlation bias	0	1	± 0.7
PCA-spectrum	0	λ^π	1
PCA-structure	0	$Q\Sigma Q^\top$	1

Simulation 2 (Gaussian random fields). We consider a spatially structured $d = 25$ dimensional setting by sampling $N = 10{,}000$ i.i.d. realizations of a Gaussian random field on a 5×5 grid, $y \sim \mathcal{N}(0, \Sigma_{\text{true}})$, where Σ_{true} is an isotropic exponential covariance built from Euclidean distances on the grid (unit marginal variance and range parameter $\tau = 1$). To probe which aspects of misspecification are detected by different preranks, we evaluate a suite of predictive models $F_{\text{pred}} = \mathcal{N}(\mu_{\text{pred}}, \Sigma_{\text{pred}})$: (i) well-specified, (ii) constant mean shift (0.5), (iii) variance inflation ($\times 1.75$), (iv) correlation-range misspecification ($\tau = 0.3$), (v) geometric isotropy misspecification (rescaling one spatial axis by a factor 5), and two PCA-targeted covariance perturbations, PC anisotropy flip and a *PCA-structure* misspecification that preserves marginal variances while altering principal directions. For each case we report prerank-based projected PIT histograms (for Location, Marginal, PCA, Dependency, Scale and HDR), which highlights the selective sensitivity of each prerank to the corresponding type of misspecification in a structured high-dimensional field.

Location pre-rank respond to most misspecifications that affect the mean of the predictive distribution, but again fail to detect the PCA-structure misspecification. Marginal pre-rank detect biases in the mean and variance, but remain insensitive to correlation, anisotropy, and PCA-based misspecifications. PCA pre-rank consistently detect all misspecifications considered, including errors in spatial dependence, anisotropy, and principal component structure. HDR capture most deviations from the true distribution, but exhibit reduced sensitivity to isotropy misspecification. Dependency pre-rank again display the same previous behavior, but reliably detect errors in spatial correlation, anisotropy, and PCA misspecifications. Finally, scale pre-rank is insensitive to mean bias

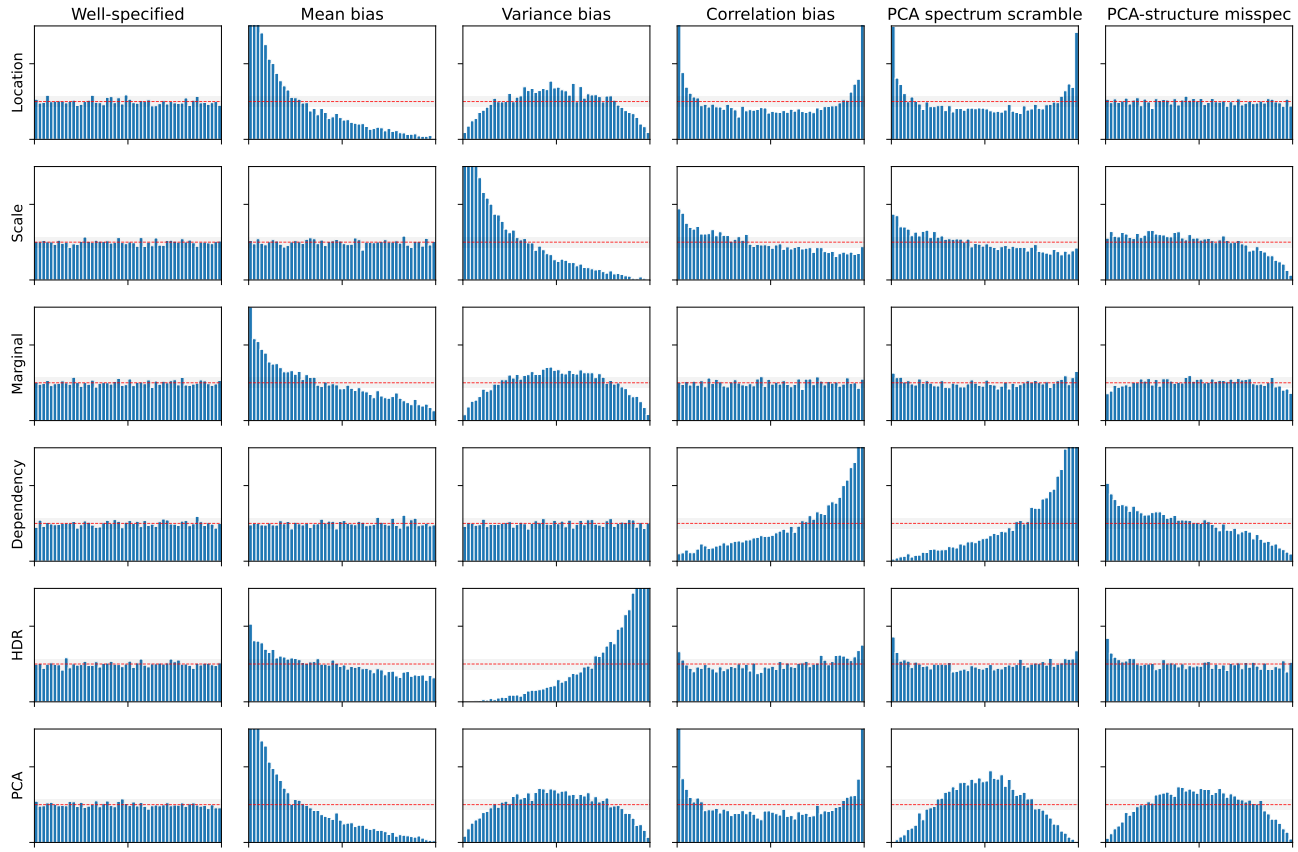


Figure 4: Projected PIT histograms for **Simulation 1** for different pre-rank functions under controlled misspecifications of a **Multivariate Gaussian**. Each column corresponds to a misspecification scenario. Each row corresponds to a pre-rank function. While most pre-ranks detect marginal and variance-related misspecifications, only the **PCA pre-rank** consistently reveals miscalibration in scenarios where the covariance structure is altered without affecting marginal distributions.

and, while trivially sensitive to variance misspecification, also respond to spatial correlation, anisotropy, and PCA biases.

Role of the PCA Pre-Rank. The simulation studies indicate that distortions of the covariance eigenstructure may remain undetected by existing pre-ranks. The PCA pre-rank addresses this limitation by assessing calibration along the principal directions of predictive distribution. Because these directions capture the dominant modes of the joint variability, deviations from uniformity in the corresponding projected PIT values reflect misspecification of the dependence structure.

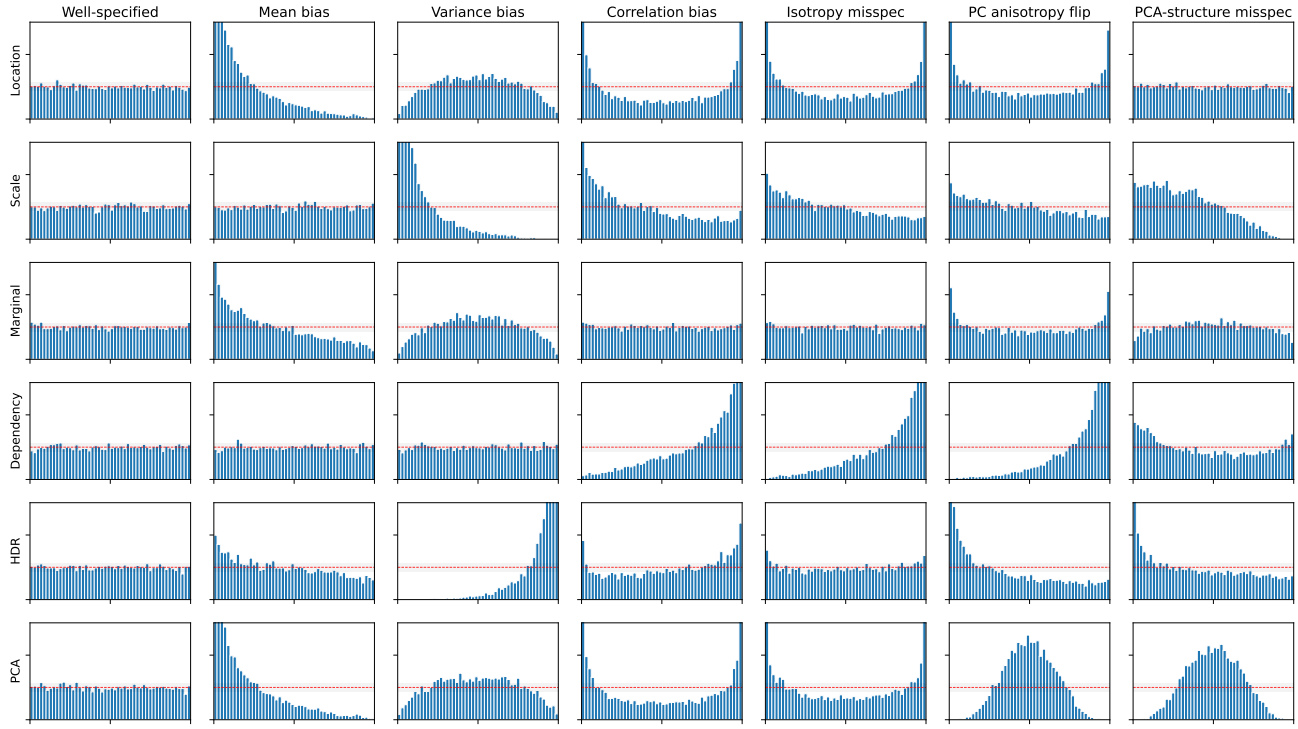


Figure 5: Projected PIT histograms for **Simulation 2** for different pre-rank functions under controlled misspecifications of an example of **Gaussian random fields**. Each column corresponds to a misspecification scenario. Each row corresponds to a pre-rank function. While most pre-ranks detect marginal and variance-related misspecifications, only the **PCA pre-rank** consistently reveals miscalibration in scenarios where the covariance structure is altered without affecting marginal distributions.

E Experiments

Experiments on Real-world Data. We evaluate our method on 18 benchmark datasets from prior work (Guan, 2021; Feldman et al., 2022; Wang et al., 2022; Dheur and Ben Taieb, 2023; Camehl et al., 2025; Dheur and Taieb, 2025), retaining those with at least 400 training instances and following the preprocessing and splits of Dheur and Taieb (2025). The datasets range from 424 to 406,440 training samples, with $L \in [1, 279]$ input features and $D \in [2, 16]$ outputs. Our base model is a mixture of K multivariate Gaussian components parameterized by a hypernetwork and trained using negative log-likelihood (NLL) augmented with our regularizer. Calibration is assessed via PCE defined in(2), while predictive performance is measured by NLL and energy score (ES). Projected PIT values are estimated through predictive sampling, and the regularization weight λ is tuned on a validation set (Table 3). Results are averaged over five random seeds.

Neural probabilistic regression model. Our base probabilistic predictor models a conditional predictive distribution as a mixture of K multivariate Gaussian components, where all parameters are generated by a hypernetwork. For each input $x \in \mathcal{X}$ and each mixture component $k \in [K]$, the network predicts the mixture weight $\pi_k(x)$, the mean vector $\mu_k(x) \in \mathbb{R}^D$, and the lower triangular Cholesky factor $L_k(x)$. The covariance matrix is then computed as $\Sigma_k(x) = L_k(x)L_k(x)^\top$, ensuring positive semi-definiteness by construction. The resulting conditional density takes the form: $\hat{f}_{Y|X=x} = \sum_{k=1}^{K} \pi_k(x)\mathcal{N}(\cdot \mid \mu_k(x), \Sigma_k(x)))$ where $\pi_k(x) \geq 0$ and $\sum_{k=1}^{K} \pi_k(x) = 1$. We train this model using the NLL scoring rule.

Hyperparameters. For the MIX-NLL baseline, we use a mixture of $K = 5$ multivariate Gaussian components. The neural network consists of three fully connected layers with 100 hidden units each, ReLU activations, and is trained using the Adam optimizer with a learning rate of 10^{-4}. To compute the PCE-KDE regularizer, we estimate the projected PITs $\hat{F}_{T|X}(T)$ using $S = 100$ samples drawn from the predictive distribution with the parameters set to $p = 1$ and $M = 100$. The temperature parameter τ in the smoothed indicator function is set

to 100, following prior work in Dheur and Ben Taieb (2023).

The regularization strength λ in (8) controls the degree of calibration enforcement with respect to the chosen pre-rank. As observed in prior work (Karandikar et al., 2021; Wessel et al., 2025), increasing λ typically improves calibration (lower PCE) but may degrade predictive performance (higher NLL or ES). Following the tuning strategy used in Karandikar et al. (2021) and Dheur and Ben Taieb (2023), we select λ to minimize PCE while ensuring that ES does not increase by more than 10% relative to the best ES obtained when $\lambda = 0$. This strategy allows us to improve calibration without sacrificing predictive accuracy. We performed this tuning separately for each dataset and pre-rank pair. We select λ on validation set from $\{0, 0.01, 0.1, 1, 5, 10\}$. Table 3 shows the selected λ for each dataset-pre-rank pair. Notably, the majority of selected values are large, often $\lambda = 10$, suggesting that future work could explore larger values or employ more sophisticated tuning strategies such as Bayesian Optimization. In our experiments, we used $\lambda = 10$.

Datasets	Marginal	Loc.	Scale	Dep.	PCA	HDR	Copula
households	10.0	10.0	10.0	5.0	10.0	10.0	5.0
air	10.0	10.0	10.0	10.0	10.0	10.0	5.0
births1	10.0	10.0	10.0	10.0	10.0	5.0	10.0
births2	10.0	10.0	5.0	5.0	10.0	0.01	10.0
wage	10.0	10.0	5.0	10.0	5.0	1.0	10.0
scm20d	10.0	10.0	10.0	10.0	5.0	10.0	0.01
scm1d	10.0	10.0	10.0	10.0	10.0	0.1	10.0
wq	5.0	10.0	10.0	5.0	5.0	5.0	10.0
scpf	5.0	10.0	10.0	0.0	5.0	1.0	10.0
meps21	5.0	5.0	5.0	10.0	10.0	10.0	5.0
meps19	5.0	10.0	1.0	10.0	10.0	10.0	10.0
meps20	5.0	10.0	1.0	1.0	10.0	10.0	10.0
house	5.0	10.0	5.0	10.0	5.0	5.0	10.0
bio	5.0	10.0	10.0	10.0	5.0	10.0	5.0
blog data	10.0	10.0	10.0	10.0	10.0	10.0	10.0
calcofi	10.0	5.0	10.0	10.0	10.0	10.0	5.0
ansur2	10.0	5.0	5.0	10.0	10.0	5.0	10.0
taxi	10.0	10.0	10.0	5.0	10.0	5.0	10.0

Table 3: Values of λ after hyperparameter tuning with each regularization and each pre-rank. The baseline used is MIX-NLL (Model trained without regularization).

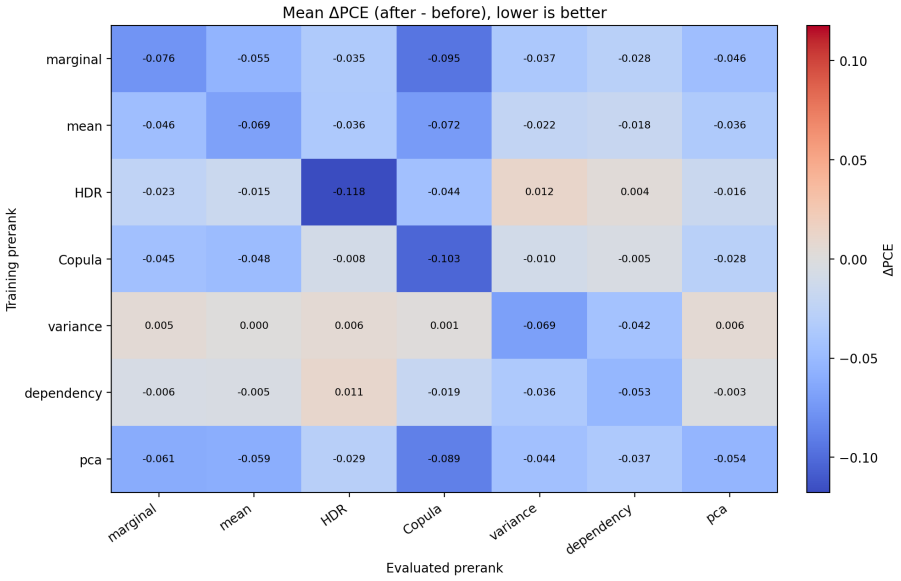


Figure 6: ΔPCE across 18 multi-output regression datasets when training with a single pre-rank regularizer (rows) and evaluating with different pre-ranks (columns). Each cell reports the change in PCE relative to the unregularized model (negative values indicate improved calibration). Regularization improves calibration not only for the targeted pre-rank (diagonal), but also across other evaluation pre-ranks. In particular, training with the PCA pre-rank yields broadly distributed improvements across multiple calibration aspects.

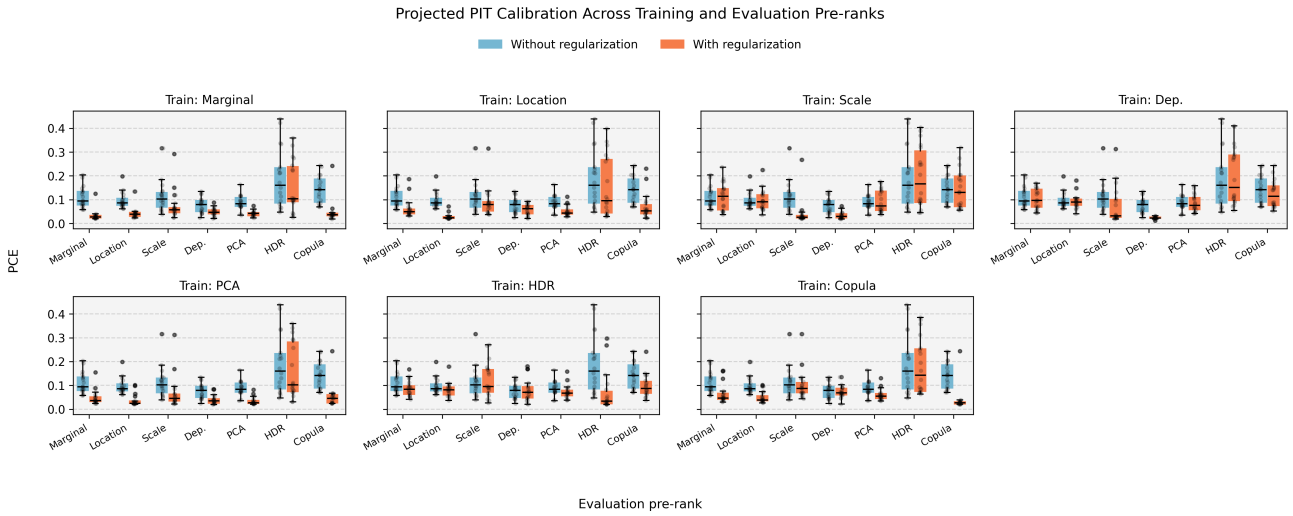


Figure 7: Grouped boxplots of PCE across 18 multi-output regression datasets when training with a single pre-rank. Each panel corresponds to a training pre-rank, and within each panel the evaluation pre-ranks are shown on the x-axis. Blue boxes denote PCE before regularization (baseline), and orange boxes denote PCE after regularization. Across training choices, regularization reduces PCE not only for the targeted evaluation pre-rank but also across multiple other pre-ranks, confirming that calibration improvements generalize beyond the objective used during training.

Distribution of the Test Statistic. To assess the statistical significance of observed calibration errors, we characterize the finite-sample distribution of the average Probability Calibration Error (PCE) under the null hypothesis of perfect calibration. For a fixed dataset and pre-rank, the test statistic is defined as the average PCE over the test set, further averaged across five independent training runs.

We fix a dataset d with test set size n_d and a given pre-rank r. Let $U^{(k)}_{d,r,1}, \ldots, U^{(k)}_{d,r,n_d}$ denote the Probability Integral Transform (PIT) values obtained from real test observations under pre-rank r in run $k \in \{1, \ldots, 5\}$. The empirical test statistic is defined as

$$\widehat{T}_{d,r} \;=\; \frac{1}{5}\sum_{k=1}^{5}\left(\frac{1}{n_d}\sum_{i=1}^{n_d}\mathrm{PCE}\Big(U^{(k)}_{d,r,i}\Big)\right),$$

and corresponds to the red dashed vertical line shown in Figures 8–14.

To characterize the distribution of this statistic under the null hypothesis of perfect calibration, we consider i.i.d. random variables $U^{(b)}_1, \ldots, U^{(b)}_{n_d} \;\sim\; \mathrm{Unif}(0,1), \qquad b = 1, \ldots, B$, with $B = 5 \times 10^4$. For each Monte Carlo replicate b, we compute

$$T^{(b)}_{d,r} \;=\; \frac{1}{n_d}\sum_{i=1}^{n_d}\mathrm{PCE}\Big(U^{(b)}_i\Big),$$

yielding an empirical approximation of the null distribution $\mathcal{L}_0(T_{d,r}) \;\approx\; \{T^{(b)}_{d,r}\}_{b=1}^{B}$.

In Figures 8–14, the blue histograms show these null distributions for each dataset and pre-rank, while the red dashed vertical line indicates the observed average PCE computed from real data using the MIX NLL model. The position of the red line relative to the blue distribution provides a direct visual assessment of calibration: values lying in the right tail indicate statistically significant deviations from perfect calibration, while overlap with the bulk of the null distribution indicates limited power to detect miscalibration.

Empirical Calculation of Energy Score. We use Energy Score (ES) as a scoring rule metric to evaluate our model performance. ES generalizes Continuous Ranked Probability Score (CRPS) to multivariate settings

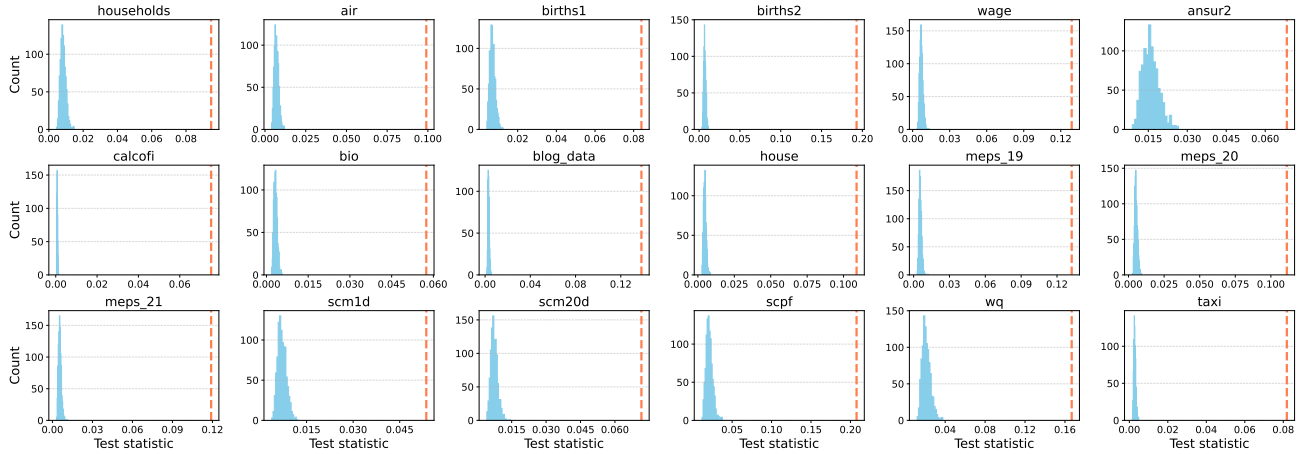


Figure 8: Null distribution (blue histogram) of the average PCE under perfect calibration (i.i.d. PIT values drawn from Unif$(0, 1)$ with sample size matching the test set), for each dataset using the **marginal** pre-rank. The red dashed vertical line marks the observed average PCE on real data (MIX NLL), averaged over five runs.

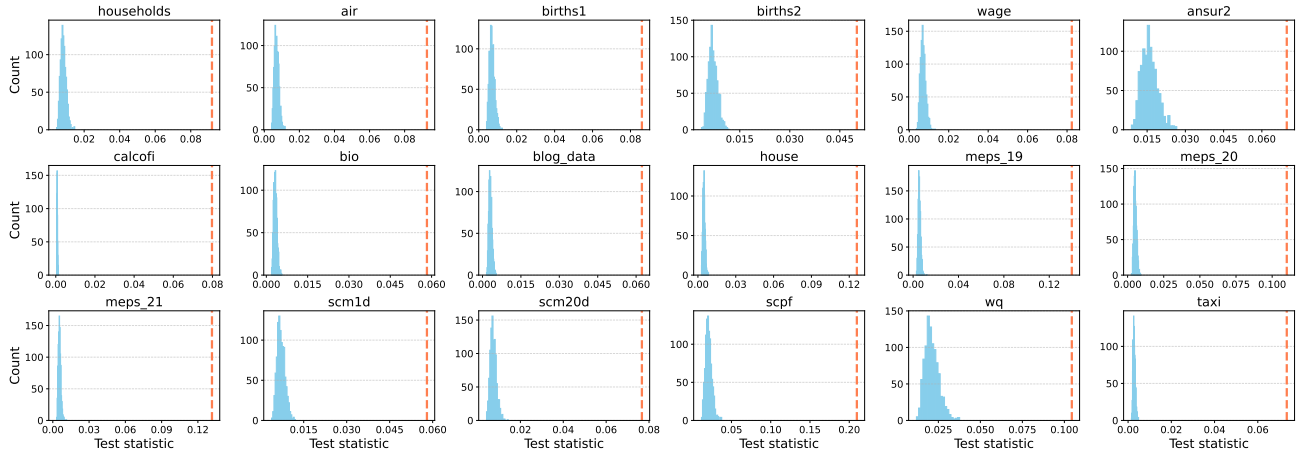


Figure 9: Same as Figure 8, but using the **mean** pre-rank.

and is computed empirically as:

$$\mathrm{ES}(\hat{F}, y) = \frac{1}{G}\sum_{i=1}^{G} \|\hat{Y}_i - y\| - \frac{1}{2G^2}\sum_{i=1}^{G}\sum_{j=1}^{G} \|\hat{Y}_i - \hat{Y}_j\| \tag{13}$$

where $\{\hat{Y}_i\}_{i=1}^{G} \sim \hat{F}_{Y|X}$ are G samples drawn from the predictive distribution. We set $G = 100$ in all experiments.

E.1 Detailed results

Reliability plots. Figures 16-20 show reliability plots obtained by evaluating calibration under different pre-rank functions. Each column evaluates calibration using a different pre-rank (marginal, location, scale, dependence, PCA-based, HDR-based, and copula-based). The dashed diagonal indicates perfect calibration. Deviations from the diagonal highlight specific forms of miscalibration, illustrating how different pre-ranks emphasize distinct aspects of the joint predictive distribution and reveal complementary calibration behavior.

Pre-rank Calibration. Table 6 reports the exact PCE values averaged over five runs for each pre-rank on which the `MIX-NLL+PCE-KDE` model was trained using the optimal λ. For comparison, we also include the PCE values computed with respect to each pre-rank for the baseline `MIX-NLL` model trained without regularization

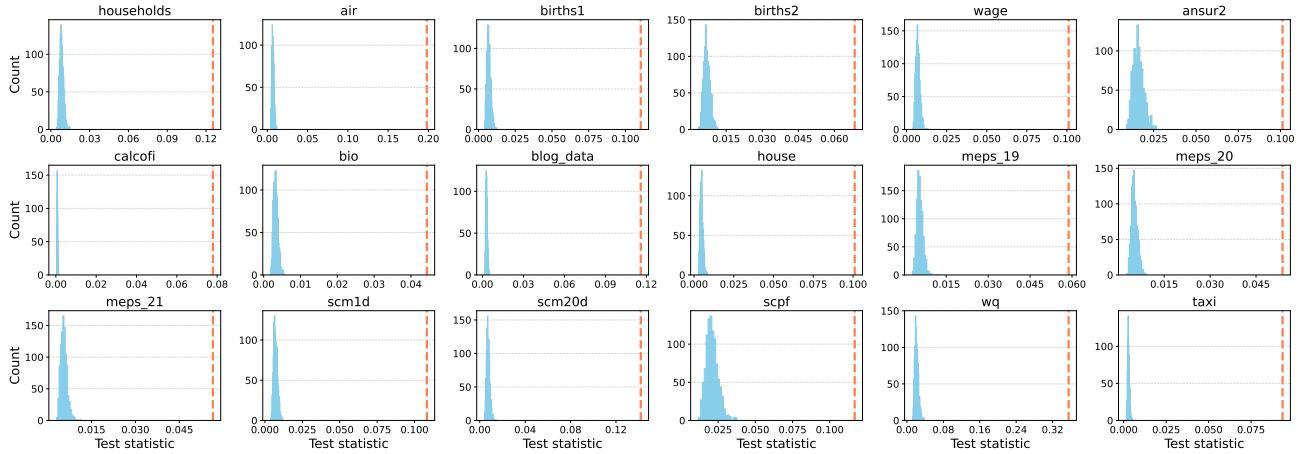


Figure 10: Same as Figure 8, but using the **variance** pre-rank.

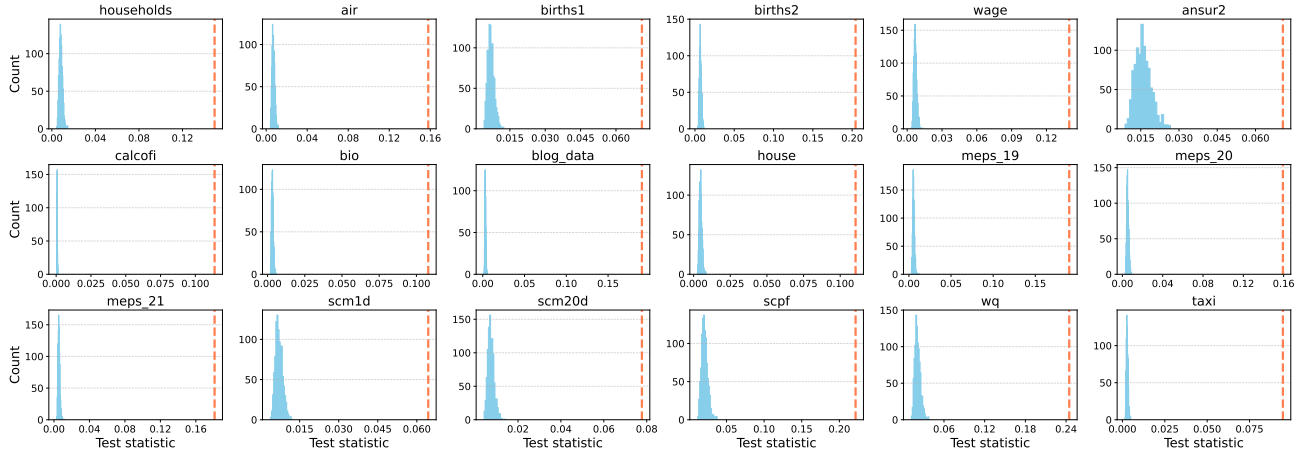


Figure 11: Same as Figure 8, but using the **CDF** pre-rank.

(see Table 5). Note that although the baseline model was not trained with respect to any specific pre-rank, we still evaluate its performance on each pre-rank to highlight the benefit of regularization.

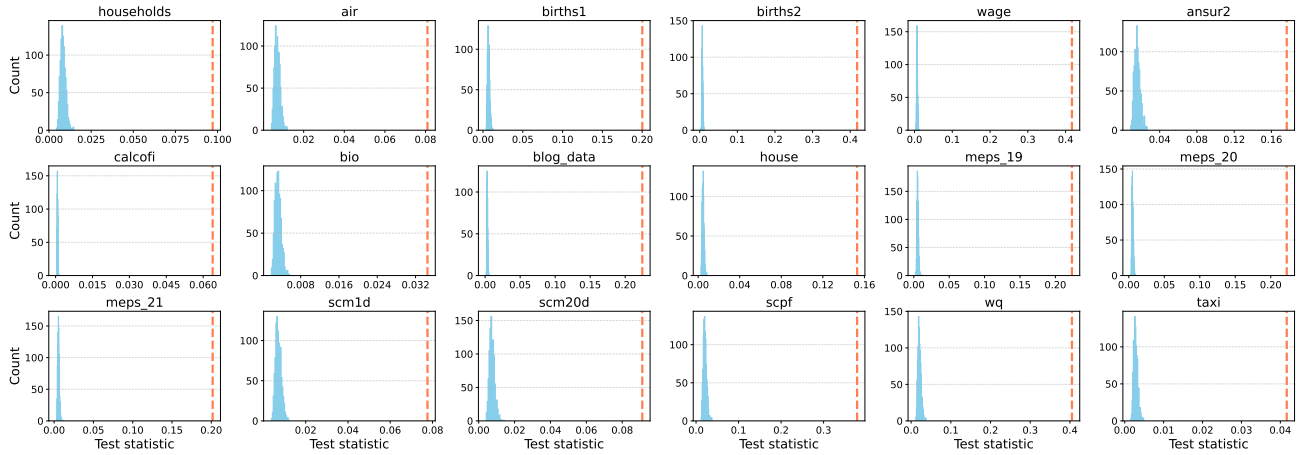


Figure 12: Same as Figure 8, but using the **density** pre-rank.

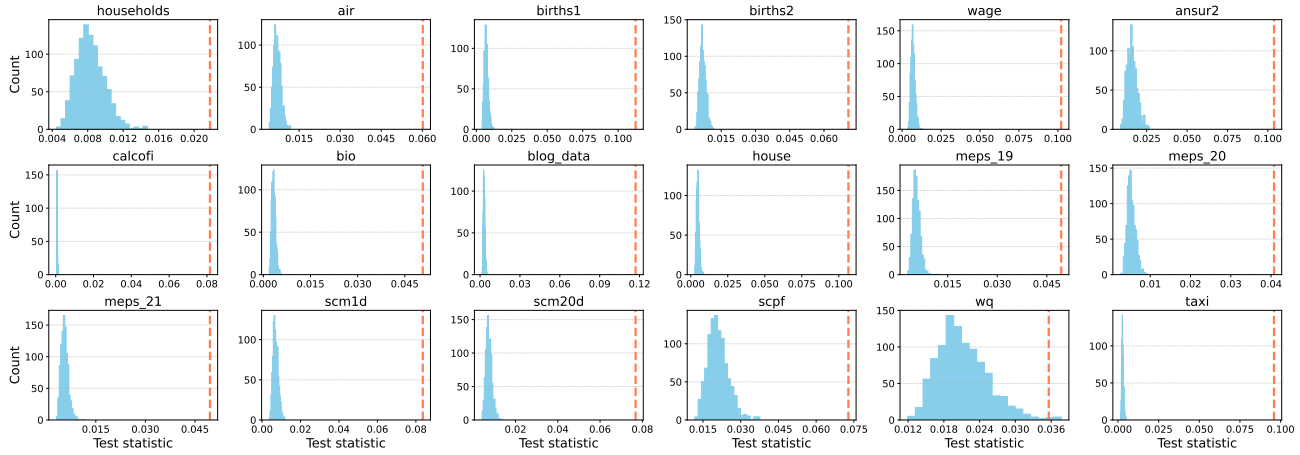


Figure 13: Same as Figure 8, but using the **dependency** pre-rank.

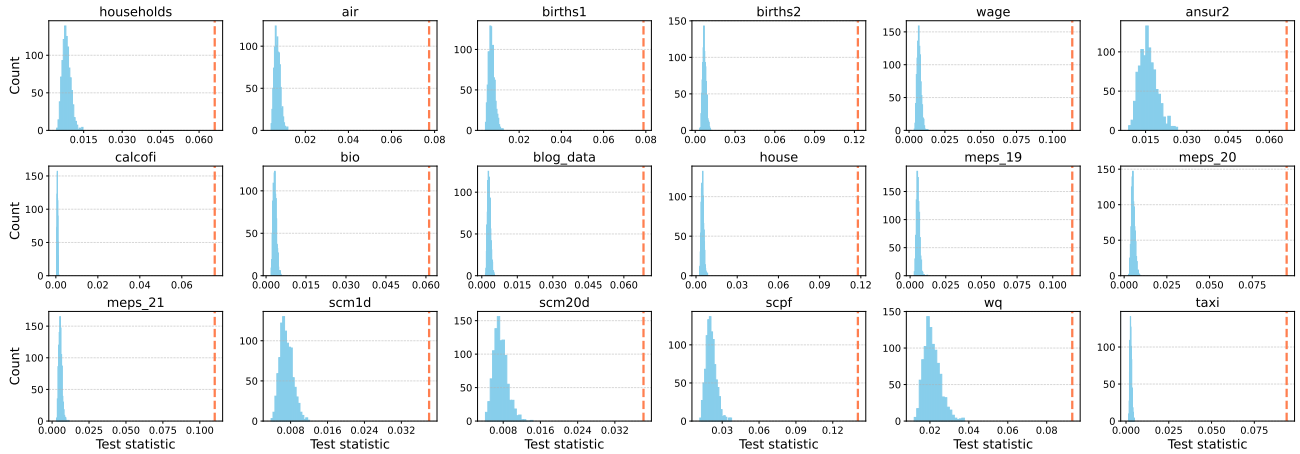


Figure 14: Same as Figure 8, but using the **PCA** pre-rank.

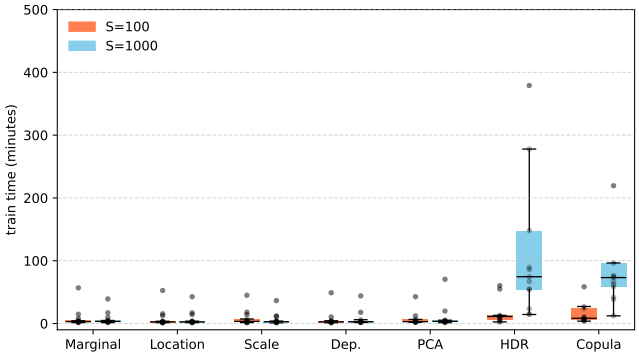


Figure 15: Training time (in minutes) across pre-rank functions for $s = 100$ and $s = 1000$, aggregated over all datasets. HDR and Copula pre-ranks incur higher computational cost, with PCA remaining substantially more efficient than HDR and Copula pre-rank.

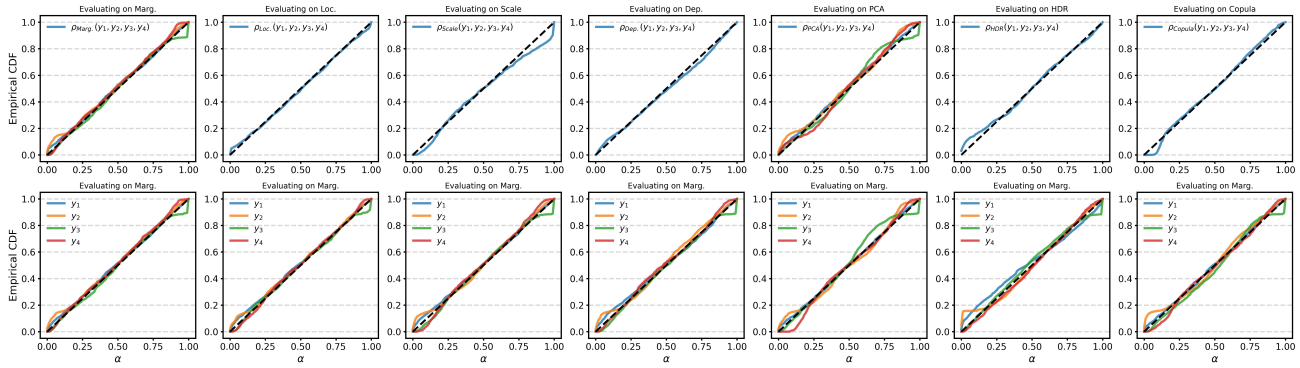


Figure 16: Reliability plots for **births2** dataset

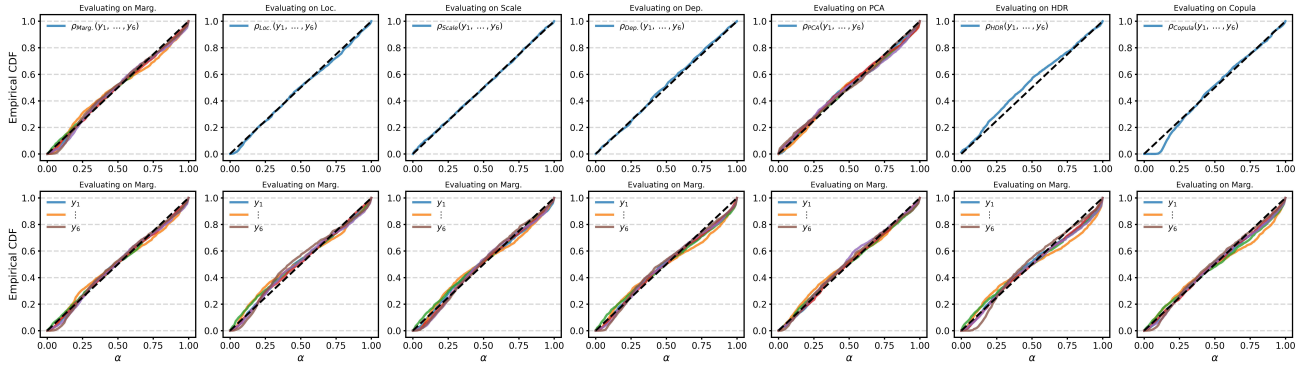


Figure 17: Reliability plots for **air** dataset

Train pre-rank	Mean PCE (before)	Mean PCE (after)	Mean ΔPCE	Improved (%)
Copula	0.1890	0.0714	-0.1176	94.4
CDF	0.1426	0.0391	-0.1034	94.4
Marginal	0.1096	0.0333	-0.0763	100.0
Scale	0.1126	0.0434	-0.0692	100.0
Mean	0.0970	0.0285	-0.0685	100.0
PCA	0.0886	0.0349	-0.0537	100.0
Dependency	0.0765	0.0239	-0.0526	100.0

Table 4: **Effect of training-time regularization by training pre-rank.** For each training pre-rank objective, we report the mean PCE across the 18 datasets before and after regularization, the mean change ΔPCE (after−before; lower is better), and the percentage of datasets for which PCE decreases.

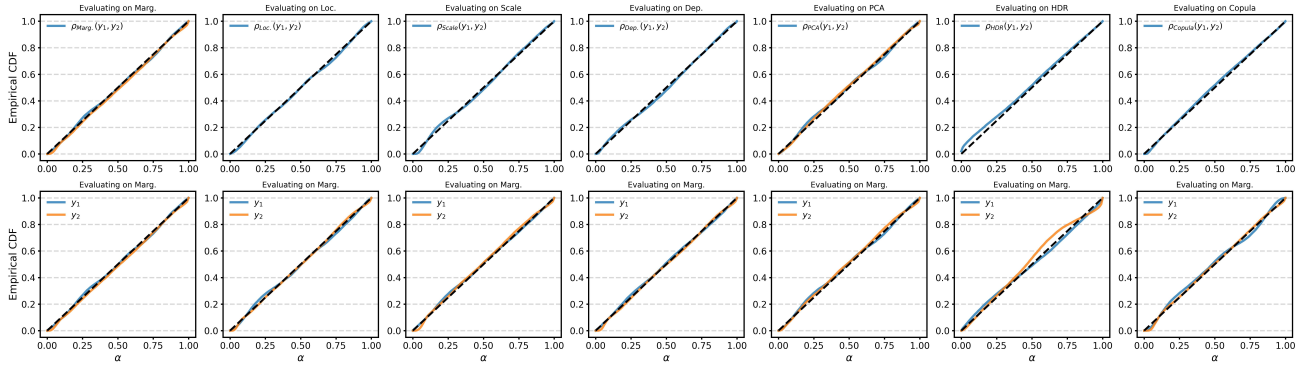


Figure 18: Reliability plots for **blogdata** dataset

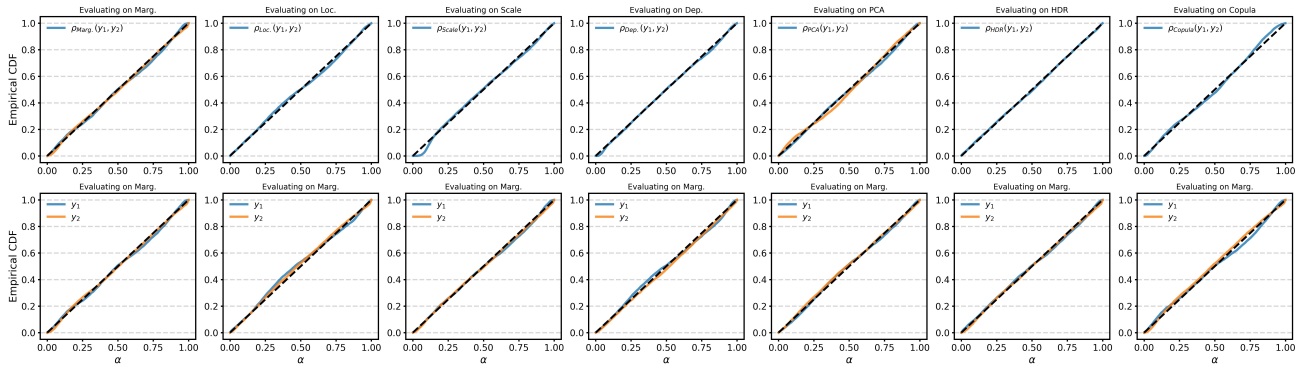


Figure 19: Reliability plots for **house** dataset

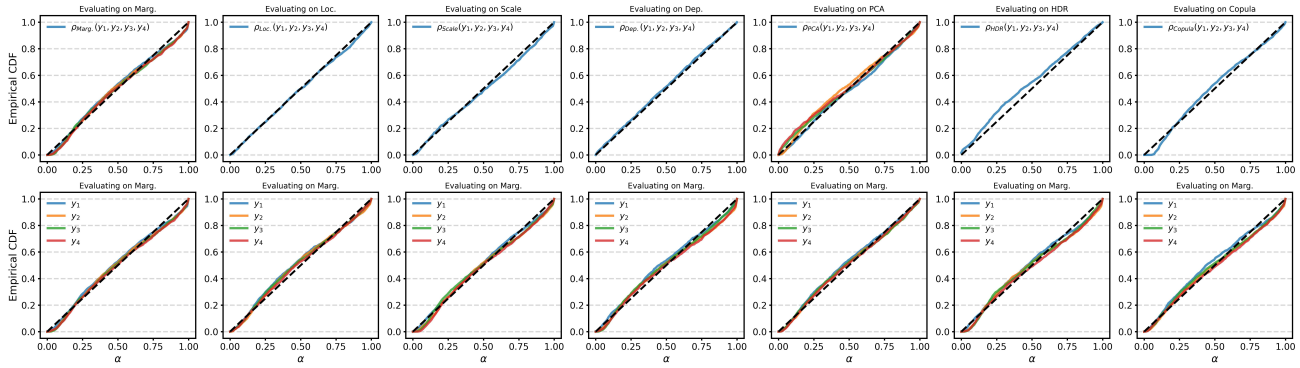


Figure 20: Reliability plots for **households** dataset

Dataset	Marg.	Loc.	Scale	Dep.	PCA	HDR	Copula
households	0.095 (0.004)	0.092 (0.002)	0.125 (0.006)	0.022 (0.002)	0.066 (0.002)	0.097 (0.007)	0.149 (0.011)
air	0.099 (0.002)	0.093 (0.003)	0.198 (0.005)	0.060 (0.003)	0.077 (0.001)	0.081 (0.007)	0.158 (0.004)
births1	0.084 (0.003)	0.086 (0.003)	0.110 (0.006)	0.112 (0.006)	0.079 (0.002)	0.200 (0.011)	0.071 (0.003)
births2	0.193 (0.008)	0.050 (0.004)	0.068 (0.002)	0.071 (0.002)	0.123 (0.006)	0.418 (0.018)	0.204 (0.031)
wage	0.129 (0.001)	0.082 (0.002)	0.101 (0.004)	0.102 (0.004)	0.114 (0.003)	0.417 (0.005)	0.140 (0.003)
scm20d	0.072 (0.001)	0.077 (0.002)	0.142 (0.004)	0.077 (0.007)	0.038 (0.001)	0.091 (0.003)	0.078 (0.001)
scm1d	0.054 (0.003)	0.058 (0.005)	0.108 (0.002)	0.084 (0.009)	0.038 (0.001)	0.078 (0.004)	0.064 (0.004)
wq	0.167 (0.005)	0.104 (0.003)	0.355 (0.008)	0.036 (0.001)	0.093 (0.001)	0.405 (0.007)	0.245 (0.001)
scpf	0.208 (0.009)	0.210 (0.013)	0.117 (0.003)	0.073 (0.006)	0.140 (0.005)	0.379 (0.015)	0.222 (0.019)
meps21	0.119 (0.002)	0.132 (0.002)	0.056 (0.003)	0.050 (0.004)	0.110 (0.005)	0.202 (0.006)	0.181 (0.006)
meps19	0.132 (0.004)	0.140 (0.005)	0.059 (0.002)	0.049 (0.002)	0.114 (0.004)	0.223 (0.009)	0.190 (0.006)
meps20	0.111 (0.004)	0.110 (0.007)	0.054 (0.001)	0.041 (0.001)	0.095 (0.004)	0.223 (0.007)	0.159 (0.005)
house	0.109 (0.002)	0.126 (0.002)	0.101 (0.001)	0.107 (0.001)	0.118 (0.002)	0.153 (0.003)	0.110 (0.002)
bio	0.057 (0.002)	0.058 (0.005)	0.044 (0.002)	0.051 (0.002)	0.061 (0.003)	0.034 (0.002)	0.108 (0.005)
blogdata	0.138 (0.002)	0.062 (0.002)	0.116 (0.004)	0.117 (0.004)	0.068 (0.001)	0.224 (0.003)	0.191 (0.006)
calcofi	0.075 (0.000)	0.080 (0.001)	0.078 (0.000)	0.082 (0.000)	0.076 (0.000)	0.064 (0.001)	0.114 (0.001)
ansur2	0.068 (0.004)	0.070 (0.006)	0.101 (0.006)	0.104 (0.006)	0.066 (0.004)	0.176 (0.011)	0.071 (0.005)
taxi	0.082 (0.001)	0.073 (0.001)	0.094 (0.003)	0.096 (0.003)	0.094 (0.002)	0.042 (0.001)	0.095 (0.002)

Table 5: **PCE** results of real-world experiments using the **MIX-NLL** model. PCE values are computed using seven pre-rank functions across 18 real datasets and averaged over five runs. Standard errors are shown in parentheses.

Dataset	Marg.	Loc.	Scale	Dep.	PCA	HDR	Copula
households	0.030 (0.001)	0.024 (0.002)	0.021 (0.002)	0.021 (0.003)	0.026 (0.001)	0.039 (0.003)	0.031 (0.001)
air	0.040 (0.001)	0.026 (0.002)	0.027 (0.001)	0.027 (0.002)	0.039 (0.002)	0.045 (0.005)	0.029 (0.001)
births1	0.028 (0.001)	0.027 (0.002)	0.031 (0.002)	0.027 (0.002)	0.034 (0.000)	0.034 (0.003)	0.027 (0.001)
births2	0.031 (0.002)	0.029 (0.002)	0.045 (0.002)	0.032 (0.001)	0.052 (0.006)	0.428 (0.002)	0.033 (0.002)
wage	0.051 (0.020)	0.044 (0.013)	0.025 (0.001)	0.024 (0.002)	0.052 (0.008)	0.364 (0.012)	0.095 (0.015)
scm20d	0.033 (0.002)	0.025 (0.001)	0.038 (0.003)	0.022 (0.002)	0.033 (0.001)	0.093 (0.003)	0.029 (0.001)
scm1d	0.025 (0.001)	0.024 (0.002)	0.036 (0.003)	0.020 (0.002)	0.035 (0.001)	0.090 (0.015)	0.036 (0.007)
wq	0.154 (0.007)	0.076 (0.010)	0.311 (0.018)	0.028 (0.003)	0.091 (0.004)	0.376 (0.026)	0.244 (0.001)
scpf	0.039 (0.002)	0.026 (0.003)	0.041 (0.005)	0.081 (0.005)	0.063 (0.004)	0.299 (0.010)	0.032 (0.006)
meps21	0.026 (0.001)	0.025 (0.001)	0.031 (0.001)	0.024 (0.001)	0.025 (0.001)	0.032 (0.001)	0.023 (0.001)
meps19	0.026 (0.002)	0.023 (0.001)	0.050 (0.002)	0.025 (0.002)	0.024 (0.001)	0.031 (0.001)	0.022 (0.001)
meps20	0.024 (0.001)	0.023 (0.001)	0.049 (0.001)	0.033 (0.001)	0.024 (0.000)	0.034 (0.005)	0.024 (0.001)
house	0.027 (0.003)	0.020 (0.001)	0.025 (0.003)	0.020 (0.001)	0.033 (0.004)	0.028 (0.002)	0.021 (0.000)
bio	0.021 (0.001)	0.020 (0.001)	0.021 (0.001)	0.021 (0.001)	0.021 (0.001)	0.021 (0.001)	0.021 (0.001)
blogdata	0.023 (0.001)	0.024 (0.000)	0.023 (0.001)	0.023 (0.001)	0.025 (0.000)	0.030 (0.001)	0.024 (0.001)
calcofi	0.020 (0.000)	0.021 (0.000)	0.020 (0.000)	0.020 (0.000)	0.021 (0.000)	0.020 (0.000)	0.020 (0.000)
ansur2	0.032 (0.004)	0.040 (0.009)	0.031 (0.004)	0.025 (0.003)	0.038 (0.003)	0.040 (0.015)	0.039 (0.009)
taxi	0.021 (0.001)	0.022 (0.001)	0.025 (0.000)	0.023 (0.001)	0.022 (0.001)	0.022 (0.000)	0.022 (0.001)

Table 6: **PCE** results after applying **PCE-KDE regularization with MIX-NLL** using the optimal λ. Results are reported for seven pre-rank functions across 18 real datasets, averaged over five runs. Standard errors are shown in parentheses.

Dataset	Marg.	Loc.	Scale	Dep.	PCA	HDR	Copula
households	3.33 (0.08)	3.40 (0.07)	3.34 (0.08)	3.23 (0.11)	3.33 (0.15)	3.29 (0.09)	3.35 (0.11)
air	5.80 (0.19)	5.81 (0.19)	5.75 (0.20)	5.71 (0.18)	5.79 (0.19)	5.67 (0.16)	5.80 (0.21)
births1	1.82 (0.39)	1.37 (0.19)	1.74 (0.44)	1.63 (0.27)	1.36 (0.21)	2.12 (0.09)	1.39 (0.17)
births2	−4.28 (0.36)	−3.72 (1.76)	−4.71 (0.39)	−4.33 (0.45)	−3.91 (0.76)	−4.10 (0.84)	−4.35 (0.72)
wage	1.63 (0.96)	1.20 (0.96)	0.77 (0.32)	0.81 (0.15)	1.01 (0.76)	0.61 (0.12)	2.55 (0.71)
scm20d	7.70 (0.32)	7.89 (0.18)	7.72 (0.30)	7.67 (0.22)	7.56 (0.21)	7.89 (0.18)	7.69 (0.13)
scm1d	7.02 (1.35)	6.58 (0.50)	6.77 (1.31)	7.17 (0.70)	11.35 (5.23)	8.04 (1.63)	10.34 (5.82)
wq	26.33 (7.30)	25.08 (6.20)	27.51 (8.82)	37.05 (34.21)	21.94 (1.62)	30.10 (12.71)	47.46 (38.00)
scpf	−3.52 (2.14)	−3.97 (1.32)	−4.48 (0.88)	−4.14 (1.27)	−4.35 (0.93)	−4.74 (1.40)	−3.69 (2.48)
meps21	−1.79 (0.18)	−1.79 (0.06)	−1.74 (0.10)	−1.72 (0.11)	−1.68 (0.05)	−1.06 (0.21)	−1.83 (0.09)
meps19	−1.84 (0.09)	−1.81 (0.06)	−2.02 (0.13)	−1.71 (0.13)	−1.74 (0.10)	−1.23 (0.11)	−1.77 (0.13)
meps20	−1.75 (0.15)	−1.71 (0.13)	−1.85 (0.17)	−1.79 (0.25)	−1.67 (0.11)	−1.01 (0.21)	−1.62 (0.20)
house	0.97 (0.46)	0.99 (0.37)	1.86 (1.70)	0.66 (0.07)	0.95 (0.36)	8.31 (10.85)	0.98 (0.33)
bio	−0.69 (0.06)	−0.64 (0.12)	−0.66 (0.08)	−0.60 (0.17)	−0.67 (0.08)	−0.48 (0.14)	−0.67 (0.11)
blogdata	−0.37 (0.07)	−0.45 (0.04)	−0.45 (0.08)	−0.51 (0.12)	−0.35 (0.12)	−1.30 (0.52)	−0.62 (0.20)
calcofi	0.59 (0.01)	0.59 (0.00)	0.60 (0.01)	0.60 (0.01)	0.60 (0.01)	0.60 (0.01)	0.59 (0.00)
ansur2	1.87 (0.04)	1.86 (0.05)	1.82 (0.04)	1.83 (0.05)	1.87 (0.05)	1.96 (0.10)	1.87 (0.04)
taxi	1.79 (0.03)	1.79 (0.03)	1.76 (0.03)	1.76 (0.04)	1.81 (0.02)	1.74 (0.02)	1.81 (0.02)

Table 7: **NLL** results after applying **PCE-KDE regularization with MIX-NLL** using the optimal λ. Results are reported for seven pre-rank functions across 18 real datasets, averaged over five runs. Standard errors are shown in parentheses.

Dataset	Marg.	Loc.	Scale	Dep.	PCA	HDR	Copula
households	0.939 (0.034)	0.944 (0.024)	0.937 (0.028)	0.921 (0.031)	0.937 (0.036)	0.933 (0.037)	0.952 (0.040)
air	1.421 (0.041)	1.427 (0.030)	1.418 (0.049)	1.406 (0.041)	1.417 (0.043)	1.405 (0.031)	1.435 (0.042)
births1	0.724 (0.013)	0.734 (0.017)	0.718 (0.012)	0.714 (0.014)	0.730 (0.010)	0.712 (0.007)	0.729 (0.009)
births2	0.964 (0.067)	0.940 (0.060)	0.905 (0.061)	0.924 (0.055)	0.930 (0.064)	0.922 (0.060)	0.960 (0.052)
wage	0.762 (0.082)	0.766 (0.083)	0.714 (0.020)	0.738 (0.018)	0.719 (0.013)	0.704 (0.010)	0.878 (0.079)
scm20d	2.247 (0.049)	2.295 (0.042)	2.379 (0.033)	2.358 (0.046)	2.297 (0.066)	2.384 (0.017)	2.350 (0.063)
scm1d	1.886 (0.078)	1.925 (0.084)	2.060 (0.097)	2.105 (0.114)	1.960 (0.066)	1.993 (0.152)	2.072 (0.086)
wq	2.597 (0.095)	2.592 (0.098)	2.601 (0.093)	2.599 (0.099)	2.599 (0.090)	2.595 (0.096)	2.596 (0.078)
scpf	0.493 (0.283)	0.497 (0.271)	0.487 (0.259)	0.480 (0.261)	0.492 (0.262)	0.482 (0.268)	0.507 (0.283)
meps21	0.410 (0.017)	0.402 (0.004)	0.423 (0.013)	0.437 (0.018)	0.415 (0.011)	0.413 (0.023)	0.413 (0.028)
meps19	0.392 (0.022)	0.401 (0.027)	0.376 (0.019)	0.423 (0.023)	0.405 (0.017)	0.395 (0.018)	0.412 (0.018)
meps20	0.405 (0.033)	0.421 (0.028)	0.390 (0.027)	0.402 (0.038)	0.428 (0.032)	0.425 (0.024)	0.426 (0.029)
house	0.571 (0.067)	0.574 (0.054)	0.576 (0.068)	0.535 (0.015)	0.565 (0.049)	0.574 (0.037)	0.576 (0.056)
bio	0.245 (0.016)	0.248 (0.016)	0.250 (0.015)	0.253 (0.019)	0.246 (0.014)	0.258 (0.017)	0.247 (0.014)
blogdata	0.637 (0.012)	0.680 (0.010)	0.707 (0.020)	0.690 (0.015)	0.668 (0.015)	0.376 (0.131)	0.635 (0.004)
calcofi	0.421 (0.001)	0.421 (0.001)	0.421 (0.001)	0.421 (0.001)	0.421 (0.001)	0.421 (0.001)	0.422 (0.001)
ansur2	0.544 (0.011)	0.546 (0.017)	0.541 (0.012)	0.540 (0.014)	0.549 (0.018)	0.577 (0.048)	0.546 (0.012)
taxi	0.706 (0.011)	0.703 (0.008)	0.707 (0.007)	0.701 (0.007)	0.707 (0.007)	0.698 (0.010)	0.703 (0.010)

Table 8: **Energy score (ES)** results after applying **PCE-KDE regularization with MIX-NLL** using the optimal λ. Results are reported for seven pre-rank functions across 18 real datasets, averaged over five runs. Standard errors are shown in parentheses.

Dataset	Copula	density	Dependency	Marginal	Mean	PCA	Scale
air	0.143 (0.002)	0.074 (0.002)	**0.050** (**0.001**)	0.091 (0.002)	0.085 (0.003)	0.071 (0.002)	0.191 (0.003)
ansur2	0.078 (0.003)	0.160 (0.002)	0.094 (0.003)	0.067 (0.002)	0.071 (0.002)	**0.066** (**0.002**)	0.095 (0.003)
bio	0.041 (0.002)	0.032 (0.001)	0.036 (0.001)	0.038 (0.001)	**0.021** (**0.001**)	0.031 (0.001)	0.030 (0.001)
births1	0.063 (0.002)	0.202 (0.003)	0.094 (0.002)	0.070 (0.002)	0.071 (0.002)	**0.071** (**0.001**)	0.093 (0.002)
births2	0.105 (0.002)	0.397 (0.003)	0.071 (0.002)	0.174 (0.002)	**0.041** (**0.002**)	0.108 (0.003)	0.058 (0.002)
blog_data	0.037 (0.002)	0.202 (0.003)	0.116 (0.003)	0.077 (0.002)	0.061 (0.002)	**0.057** (**0.002**)	0.115 (0.003)
calcofi	0.047 (0.002)	**0.028** (**0.001**)	0.033 (0.001)	0.038 (0.001)	0.028 (0.001)	0.037 (0.001)	0.031 (0.001)
house	0.091 (0.002)	0.119 (0.003)	0.091 (0.002)	0.093 (0.002)	0.113 (0.002)	0.102 (0.002)	**0.087** (**0.002**)
households	0.137 (0.002)	0.088 (0.002)	**0.020** (**0.001**)	0.089 (0.002)	0.089 (0.002)	0.066 (0.002)	0.123 (0.002)
meps_19	0.058 (0.002)	0.217 (0.003)	**0.040** (**0.002**)	0.060 (0.002)	0.053 (0.002)	0.042 (0.002)	0.055 (0.002)
meps_20	0.044 (0.002)	0.207 (0.003)	**0.033** (**0.002**)	0.054 (0.002)	0.037 (0.002)	0.034 (0.002)	0.051 (0.002)
meps_21	0.054 (0.002)	0.208 (0.003)	**0.038** (**0.002**)	0.056 (0.002)	0.045 (0.002)	0.035 (0.002)	0.054 (0.003)
scm1d	0.062 (0.002)	0.077 (0.002)	0.075 (0.002)	0.050 (0.002)	0.053 (0.002)	**0.038** (**0.002**)	0.104 (0.003)
scm20d	0.072 (0.002)	0.088 (0.002)	0.068 (0.002)	0.068 (0.002)	0.074 (0.002)	**0.037** (**0.001**)	0.130 (0.003)
scpf	0.089 (0.003)	0.343 (0.003)	**0.071** (**0.003**)	0.125 (0.003)	0.100 (0.003)	0.106 (0.003)	0.119 (0.003)
taxi	0.083 (0.002)	0.034 (0.001)	0.081 (0.002)	0.073 (0.002)	0.067 (0.002)	0.078 (0.002)	**0.079** (**0.002**)
wage	0.135 (0.003)	0.412 (0.003)	0.093 (0.003)	0.127 (0.003)	**0.078** (**0.003**)	0.113 (0.003)	0.092 (0.003)
wq	0.245 (0.003)	0.399 (0.003)	**0.026** (**0.001**)	0.167 (0.003)	0.111 (0.003)	0.093 (0.003)	0.350 (0.003)

Table 9: PCEs when the model is trained with **Marginal+ALL** and evaluated on each simple pre-rank. Each entry shows mean PCE (std) over five runs. Bold indicates the best (lowest) PCE within each dataset row.

Dataset	Copula	density	Dependency	Marginal	Mean	PCA	Scale
air	0.144 (0.002)	0.081 (0.003)	**0.048** (**0.001**)	0.093 (0.002)	0.088 (0.002)	0.071 (0.001)	0.190 (0.003)
ansur2	0.070 (0.003)	0.152 (0.002)	0.092 (0.003)	0.064 (0.002)	0.066 (0.002)	**0.062** (**0.002**)	0.092 (0.003)
bio	0.041 (0.002)	0.038 (0.001)	0.040 (0.001)	0.041 (0.001)	**0.022** (**0.001**)	0.033 (0.001)	0.033 (0.001)
births1	0.060 (0.002)	0.184 (0.003)	0.092 (0.002)	0.070 (0.002)	0.070 (0.002)	**0.070** (**0.001**)	0.091 (0.002)
births2	0.114 (0.003)	0.386 (0.003)	0.065 (0.002)	0.174 (0.003)	**0.038** (**0.002**)	0.106 (0.003)	0.057 (0.002)
blog_data	0.045 (0.002)	0.199 (0.003)	0.108 (0.003)	0.079 (0.002)	0.058 (0.002)	**0.054** (**0.002**)	0.107 (0.003)
calcofi	0.050 (0.002)	**0.032** (**0.001**)	0.032 (0.001)	0.041 (0.001)	0.033 (0.001)	0.037 (0.001)	0.031 (0.001)
house	0.100 (0.002)	0.117 (0.003)	0.098 (0.002)	0.101 (0.002)	0.120 (0.002)	0.110 (0.002)	**0.093** (**0.002**)
households	0.146 (0.003)	0.090 (0.002)	**0.022** (**0.001**)	0.093 (0.002)	0.093 (0.002)	0.065 (0.002)	0.117 (0.003)
meps_19	0.050 (0.002)	0.226 (0.003)	0.038 (0.002)	0.064 (0.002)	0.044 (0.002)	**0.036** (**0.002**)	0.055 (0.003)
meps_20	0.045 (0.002)	0.234 (0.003)	0.036 (0.002)	0.066 (0.002)	0.040 (0.002)	**0.036** (**0.002**)	0.052 (0.003)
meps_21	0.052 (0.002)	0.207 (0.003)	0.036 (0.002)	0.062 (0.002)	0.043 (0.002)	**0.039** (**0.002**)	0.052 (0.003)
scm1d	0.066 (0.002)	0.090 (0.002)	0.088 (0.002)	0.057 (0.002)	0.067 (0.002)	**0.043** (**0.002**)	0.099 (0.003)
scm20d	0.075 (0.002)	0.086 (0.002)	0.067 (0.002)	0.070 (0.002)	0.074 (0.002)	**0.036** (**0.001**)	0.142 (0.003)
scpf	0.084 (0.003)	0.347 (0.003)	**0.068** (**0.003**)	0.123 (0.003)	0.098 (0.003)	0.106 (0.003)	0.113 (0.003)
taxi	0.081 (0.002)	**0.034** (**0.001**)	0.081 (0.002)	0.074 (0.002)	0.065 (0.002)	0.080 (0.002)	0.079 (0.002)
wage	0.137 (0.003)	0.414 (0.003)	0.097 (0.003)	0.130 (0.003)	**0.083** (**0.003**)	0.118 (0.003)	0.096 (0.003)
wq	0.246 (0.003)	0.394 (0.003)	**0.026** (**0.001**)	0.163 (0.003)	0.113 (0.003)	0.088 (0.003)	0.340 (0.003)

Table 10: PCEs when the model is trained with **PCA+ALL** and evaluated on each simple pre-rank. Each entry shows mean PCE (std) over five runs. Bold indicates the best (lowest) PCE within each dataset row.