Too Good to Conform: The Inverse Quality Paradox in Counterfactual Filtering

Fatima Rabia Yapicioglu PHD.FATIMA.RABIA.YAPICIOGLU@LAMBORGHINI.COM
Automobili Lamborghini, Italy; University of Bologna, DISI, Italy
Abhishek Srinivasan ABHISHEK.SRINIVASAN@SCANIA.COM
Scania CV AB, Sweden; KTH Royal Institute of Technology, Sweden
Juan Carlos Andresen JUAN-CARLOS.ANDRESEN@SCANIA.COM
Scania CV AB, Sweden
Henrik Boström BOSTROMH@KTH.SE
KTH Royal Institute of Technology, Sweden

Editor: Ernst Ahlberg, Ulf Johansson, Henrik Boström, Alberto Carlevaro, Johan Hallberg Szabadváry and Lars Carlsson

1. Introduction

As machine learning models enter high-stakes decision-making, explanations must be interpretable, reliable, and uncertainty-aware (Löfström et al., 2024, 2026). Counterfactuals (CFs) are a prominent explanation form, but their plausibility remains difficult to guarantee (Guidotti, 2024). Conformal prediction (Vovk et al., 2022) addresses this through *generate-then-filter*, which validates candidates post-hoc (Bates et al., 2023; Maalej et al., 2025; Adams et al., 2025), or *filter-then-generate*, which incorporates conformal constraints into generation (Altmeyer et al., 2024; Bilkhoo et al., 2025). We study the former. Although nonconformity-measure (NCM) design is actively studied in other conformal settings (Narteni et al., 2026), its role as a structural failure point in counterfactual filtering remains unexplored. We show that an NCM assigning greater nonconformity to candidates closer to $\mathbf{x}_i$ can systematically reject the most proximal valid CFs — as generator quality (proximity to $\mathbf{x}_i$ among candidates achieving the desired prediction change) improves, nonconformity increases, producing an *inverse quality paradox*. Our contributions are to: (i) formalise this failure for the reciprocal-distance score A_{prox}; (ii) show that it can reject reference-supported candidates while accepting candidates outside the data-supported region; and (iii) demonstrate that a score A_R anchored to a fixed in-distribution reference set removes this dependence on $\mathbf{x}_i$ and retains standard conformal validity under exchangeability.

2. Can NCM choice alone cause degenerate filtering?

In this construction, conformal calibration operates as specified; the degeneracy is induced by the NCM.

Proposition 1 (Inverse quality paradox) *Let* $A_{\text{prox}}(\mathbf{z}) = \frac{1}{d(\mathbf{z},\mathbf{x}_i)+\eta}$ *and let* $p(\mathbf{z}) = \big(1 + |\{\mathbf{c} \in \mathcal{C} : A_{\text{prox}}(\mathbf{c}) \geq A_{\text{prox}}(\mathbf{z})\}|\big)/(|\mathcal{C}|+1)$ *be the conformal p-value of* $\mathbf{z}$ *w.r.t. calibration scores* $\{A_{\text{prox}}(\mathbf{c}) : \mathbf{c} \in \mathcal{C}\}$. *If* $d(\mathbf{z},\mathbf{x}_i) < \min_{\mathbf{c}\in\mathcal{C}} d(\mathbf{c},\mathbf{x}_i)$, *then* $p(\mathbf{z}) = 1/(|\mathcal{C}|+1)$, *so* $\mathbf{z}$ *is rejected for any* $\delta \geq 1/(|\mathcal{C}|+1)$. *Higher generator quality drives* $d(\mathbf{z},\mathbf{x}_i) \to 0$, *and the rejection rate converges to one.*

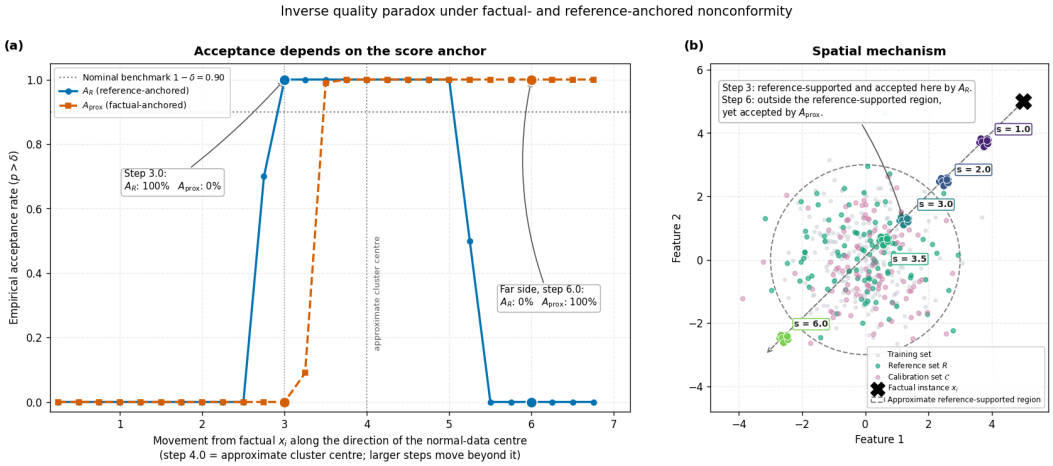

Figure 1: Synthetic experiment (n_{train}=200, $|R|$=$|\mathcal{C}|$=100, k=5, δ=0.10). **(a)** Acceptance rate vs. step s from $\mathbf{x}_i$ toward the normal-data centre. At s=3.0, A_R accepts 100%, A_{prox} 0%; on the far side (s=6.0), A_R rejects while A_{prox} accepts 100% of out-of-distribution candidates. **(b)** Spatial mechanism: step-3 candidates are reference-supported yet closer to $\mathbf{x}_i$ than any calibration point; step-6 candidates lie outside the reference-supported region yet are accepted by A_{prox}.

3. Why does a tighter generator make things worse?

Under A_{prox}, closeness to $\mathbf{x}_i$ *is* the nonconformity signal, while the calibration set $\mathcal{C}$, drawn from the normal distribution, sits far from $\mathbf{x}_i$ ($\bar{d}(\mathcal{C}, \mathbf{x}_i) \approx 7$ in our experiment). Any CF closer than every calibration point collapses to $p = 1/(|\mathcal{C}|+1)$, below any reasonable δ; any candidate *farther* than the calibration points is accepted, plausible or not. A_{prox} can reject proximal candidates supported by the reference distribution while accepting candidates lying well outside that support.

4. Does anchoring to R resolve the paradox?

Figure 1 sweeps candidates from $\mathbf{x}_i$ toward the normal-data centre in steps s (s=4.0 $\approx$ centre; larger steps move beyond it). The measures disagree in both directions. At s=3.0, candidates land inside the reference-supported region: A_R accepts 100%, whereas A_{prox} accepts 0% because the same candidates remain closer to $\mathbf{x}_i$ than every calibration point. A_{prox} begins accepting only from s=3.25 (9%; 99% at s=3.5), precisely once candidates are pushed farther from $\mathbf{x}_i$ than the calibration distances. On the far side ($s \geq 5.5$), candidates exit the data support: A_R drops back to 0%, whereas A_{prox} still accepts 100%. A_R thus defines a *plausibility window*; A_{prox} acts as a monotone distance switch, and validity is preserved under A_R (Proposition 2).

Proposition 2 (Reference anchoring restores validity) *Let $A_R(\mathbf{z})$ be the mean distance from $\mathbf{z}$ to its k nearest neighbours in a reference set R fixed before calibration. Conditional on R, if $\mathbf{z}$ is exchangeable with the calibration points, then*

$$\mathbb{P}(p_R(\mathbf{z}) \leq \delta) \leq \delta.$$

Exchangeability of generated candidates is itself a known concern in counterfactual generation (Bilkhoo et al., 2025); our finding is orthogonal: even granting exchangeability, A_{prox} fails structurally, whereas A_R fails only where exchangeability does.

Acknowledgments

Fatima R. Yapicioglu is a University of Bologna PhD student funded by PNRR and Automobili Lamborghini S.p.A.

References

James Adams, Gesine Reinert, Lukasz Szpruch, Carsten Maple, and Andrew Elliott. Individualised counterfactual examples using conformal prediction intervals. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266, pages 425–444. PMLR, 2025.

Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. Faithful model explanations through energy-constrained conformal counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10829–10837, 2024.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Aman Bilkhoo, Milad Kazemi, Nicola Paoletti, and Mehran Hosseini. CONFEX: Uncertainty-aware counterfactual explanations with conformal guarantees. *arXiv preprint arXiv:2510.19754*, 2025.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

Helena Löfström, Tuwe Löfström, Ulf Johansson, and Cecilia Sönströd. Calibrated explanations: With uncertainty information and counterfactuals. *Expert Systems with Applications*, 246:123154, 2024.

Helena Löfström, Tuwe Löfström, Anders Hjort, and Fatima Rabia Yapicioglu. Concerning uncertainty — a systematic survey of uncertainty-aware XAI. *Preprint*, 2026.

Aicha Maalej, Cecilia Sönströd, and Ulf Johansson. Counterfactual explanations for conformal prediction sets. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266, pages 405–424. PMLR, 2025.

Sara Narteni, Alberto Carlevaro, Fabrizio Dabbene, Marco Muselli, and Maurizio Mongelli. A novel score function for conformal prediction in rule-based binary classification. *Pattern Recognition*, 171:112219, 2026.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, 2 edition, 2022.