Super-relevant synthetic data with individualized Shapley values

Fernando Delbianco - Fernando Tohmé

Departamento de Economía, Universidad Nacional del Sur

Instituto de Matemática de Bahía Blanca, CONICET

Abstract

Individualized inference (or prediction) is an approach to data analysis that provides tailored analytical insights for specific queries. It is increasingly relevant thanks to the availability of large datasets. This paper presents an algorithm that identifies relevant observations through similarity metrics and further refines this selection by weighting with Shapley values. The probability distribution over this selection allows for generating synthetic controls, which in turn can be used to generate a robust inference (or prediction). Data collected from repeating this procedure for different queries provides a deeper understanding of the general process that generates the data.

Keywords: Individualized inference, Relevance selection, Relevance classification, Synthetic controls.

1 Introduction

The increasing availability of large datasets enables new approaches to individualized inference, which facilitates personalized prediction and query-specific data analysis. Traditional global models often fail to capture the specificities present in different segments of the data relevant to individual queries. Thus, the ability to detect and use local, highly relevant information from the dataset becomes crucial for robust individualized inference.

In this context, we propose an algorithm that begins by identifying a neighborhood of relevant observations based on a fuzzy similarity metric tailored to each query. This initial subset is then refined using Shapley values, which measure the marginal contribution of individual observations to predictions, identifying the most influential or "super-relevant" data points. By constructing a probability distribution over these refined observations, the algorithm generates synthetic controls that improve the robustness of the inference. This technique not only improves prediction accuracy but also provides deeper insights into the data-generating mechanism by focusing on query-specific relevant information.

The plan of the paper is as follows. Section 2 presents a brief literature review, providing the setting for our contribution. Section 3 describes the methodological approach, and Section 4 describes in detail its algorithmic implementation. Section 5 presents the results of applying this methodology to real-world data.

2 Literature Review

Individualized inference has gained increasing attention as the volume and dimensionality of data grow, motivating methods that move beyond aggregate analyses to tailored predictions (Meng (2018) Cai et al. (2021) Liang et al. (2022)). Previous work (Delbianco et al. (2021), Delbianco and Tohmé (2024), Delbianco and Tohmé (2025a), Delbianco and Tohmé (2025b)) introduced methodologies for identifying relevant subsets of data to answer specific individual queries, emphasizing the importance of local relevance in predictive modeling.

Shapley values, originally developed in cooperative game theory, have been increasingly adopted in machine learning and statistics as a principled technique for attributing the contribution of individual features or observations to a model's output (Lundberg and Lee (2017) Molnar (2020) Rozemberczki et al. (2022)). Their application in data selection and weighting offers a rigorous way to refine relevant samples by quantifying each observation's importance to predictions.

Furthermore, recent advances in conformal prediction (Angelopoulos and Bates (2023), Zhou et al. (2025)) provide frameworks to construct predictive models with valid coverage guarantees, enhancing the reliability of individualized prediction intervals (see Jin et al. (2023)). Our approach leverages this literature by combining relevance-based selection, Shapley value refinement, and synthetic data generation via evolutionary distribution algorithms to deliver robust and interpretable individualized inference.

This study contributes by integrating relevance detection with Shapley value-guided refinement and synthetic controls, generating a novel algorithm that both improves individualized prediction accuracy and offers interpretability into the data subsets most influencing the results.

3 Methodology

Assume that $\mathcal{O}$ consists of entries $\{o_i = \langle x_i, y_i \rangle\}_{i=1,\ldots,n}$, where each o_i is a vector with two components: x_i, a vector of p variables called the *tail* of o_i; and y_i, the *head* of o_i. A query q consists of a *tail*, x_0, with no *head*, and with $x_o \neq x_i$, for $i = 1, \ldots, n$. A distribution P_{θ^q} over $\mathcal{O}$ has as its support the set of *relevant* observations $\mathcal{O}_q \subseteq \mathcal{O}$. P_{θ^q} is defined by the *similarity* between x_0 and the x_i's. A *robust* inference $y = f(x)$ yields a distribution $\overline{P}_{\theta^q}$. Its support consists of pairs $\langle x, y \rangle$ similar to those in $\mathcal{O}_q$, yielding a set $\overline{\mathcal{O}}_q$ such that Delbianco et al. (2021):

Theorem: *The distribution $\overline{P}_{\theta^q}$ with support $\overline{\mathcal{O}}_q$ minimizes $E(\overline{P}_{\theta^q}) - E(P_{\theta^q})$, yielding a class $\bar{Y}_q = \{y : \text{there exists } \langle x, y \rangle \in \overline{\mathcal{O}}_q \text{ with } \overline{P}_{\theta^q}(\langle x, y \rangle) > 0\}$ of potential answers to the query .*

A further refinement can be obtained by considering the model $y = f(x)$. The contribution of an individual $o_k = \langle x_k, y_k \rangle$ to the prediction $f(x_0)$ for the query q is its *Shapley value*

$$\phi_k(f, x_0) = \sum_{S \subseteq \mathcal{O}_q \setminus \{o_k\}} \frac{|S|!(|\mathcal{O}_q| - |S| - 1)!}{|\mathcal{O}_q|!} (f(x_0|S \cup \{x_k\}) - f(x_0|S))$$

where S is a subset of $\mathcal{O}_q$ *without* o_k. Given a tail x_0, we can normalize the Shapley values as follows:

$$\phi_0(f, \emptyset) + \sum_{k=1}^{n} \phi_k(f, x_0) = f(x_0)$$

Consider a binary function $L(o_k)$, for $o_k \in \mathcal{O}_q$, which is 1 if $|(y_k - f(x_k))| \leq \gamma$, for a given $\gamma \geq 0$ while $L(o_k) = 0$, otherwise. We normalize the Shapley values, i.e., for each k:

$$\bar{\phi}'_k := \frac{\bar{\phi}_k}{f(x_0) - E_{x_k \in \mathcal{O}_q)} f(x_k)}$$

Then, we run a regression

$$L(o_k) = \beta_0 + \sum_{o_j \in \mathcal{O}_q} \beta_j \bar{\phi}'_j + \epsilon$$

Only the observations in $\mathcal{O}_q$ with non-zero coefficients in the regression are kept. This yields a restricted set of relevant observations, $\mathbf{O}_q \subseteq \mathcal{O}_q$. Its elements are called *super-relevant* observations.

Finally, the distribution $\bar{P}_{\theta^q}$ over $\mathbf{O}_q$, inferred in the last step, is used to generate synthetic observations, yielding the set of *super-relevant synthetic* data, denoted $\bar{\mathcal{O}}_q$.

The whole process is depicted in Figures 1 and 2 (Delbianco and Tohmé, 2025b).

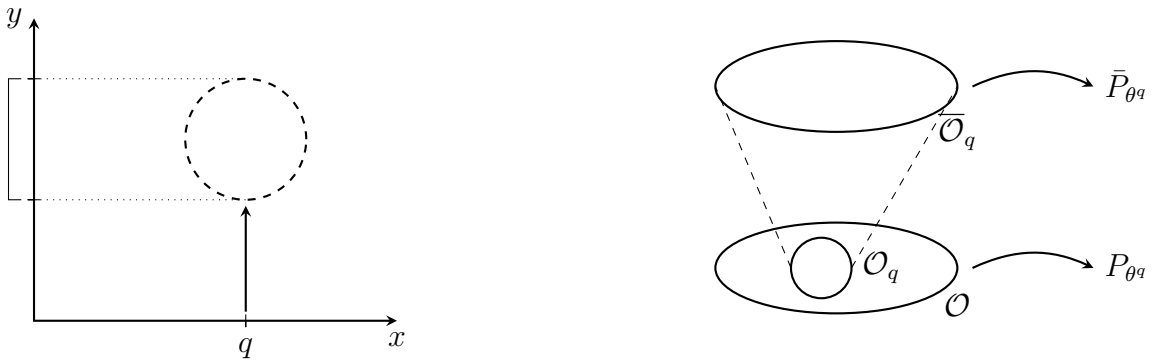

Figure 1: Individualized synthetic controls

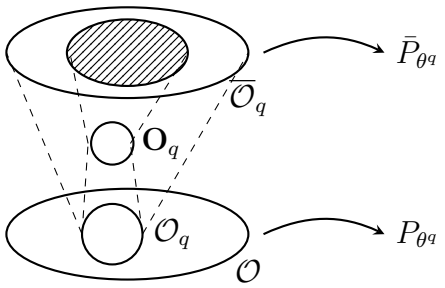

Figure 2: Using Shapley values to refine relevant selection

4 Algorithm

We will now describe how the previous procedures are integrated into an algorithm to compute the answer to a query q. The sequence is as follows:

1. *Query*: The first step involves defining the exact specification of the query q. It is constituted by a *tail* x_0, which can be conceived as a vector with as many dimensions as explanatory variables of an (unknown) head y_0 that might also be multi-dimensional.

2. *Neighborhood*: The computational cost of determining a sample of *relevant* observations requires the application of a simple procedure to detect appropriate candidates. A possibility is to find a neighborhood of radius α. This defines $\mathcal{O}_q$.

3. $\mathcal{O}_q$ is used to find a model $y = f(x)$ and compute $f(x_0)$.

4. *Shapley value*: Compute the class of Shapley values $\{\phi_k(f, x_0)\}_{o_k \in \mathcal{O}_q}$.

5. *Regression*: The loss function $L(\cdot)$ is regressed on the Shapley values to find their corresponding weights. The observations with weights that are not significant are discarded. The remaining observations constitute a set $\mathbf{O}_q$. We identify $\{\phi_j(f, x_0)\}_{o_j \in \mathbf{O}_k}$ with the *latent variable* θ_q.

6. *Estimation*: Find the distribution over $\mathbf{O}_q$.

7. *Simulation*: Draw new observations from this distribution over $\mathbf{O}_q$, generating a class of observations $\bar{\mathbf{O}}_q$.

8. *Inference*: A model $y = \mathbf{f}_q(x)$ is inferred on the basis of the data in $\bar{\mathbf{O}}_q$. Then, the answer to q is $\mathbf{f}(x_0)$.

Algorithm 1 Individualized Prediction with Relevant Synthetic Controls

Input: $\mathcal{O} = \{\langle x, y\rangle, x \in X, y \in Y\}$, $X_0 \subseteq X$, Regression method $\mathcal{A}$, Similarity intensity α
Output: $\hat{y}_0$ for each element $x_0 \in X_0$ of q, $\mathcal{O}_q$, $\mathbf{O}_q$, $\bar{\mathcal{O}}_q$
$\alpha \in [0, 1]$
$\mathcal{A}$ ▷ Regression method (e.g. OLS)
$\mathcal{L}$ ▷ Loss function (e.g. $L(o_k)$)
for each $x \in X_0$ **do**
1 Predict y_0 with $\mathcal{A}$ and X
2 Use α to find $\mathcal{O}_q$
3 Use $\mathcal{O}_q \rightarrow$ *Shapley values* in $\mathcal{O}_q$
4 Regress $\mathcal{L}$ on the *Shapley values* $\rightarrow \mathbf{O}_q$ that reject the null hypothesis H_0
5 Use $\mathbf{O}_q \rightarrow \bar{P}_{\theta^q}$
6 Generate synthetic data based on $\bar{P}_{\theta^q} \rightarrow \bar{\mathcal{O}}_q$
7 Predict y_0 with $\bar{\mathcal{O}}_q \rightarrow \hat{y}_0$
end for each $x_0 \in X_0$
Return: Y_0

5 Illustration

To illustrate how the proposed algorithm works, we'll predict the price of cell phones based on their characteristics. To do this, we use a database available on Kaggle, which includes various characteristics of each cell phone and price variability (in Indian rupees), using seven features (Ratings, Ram, ROM, Mobile Size, Primary Cam, Selfi-Cam and Battery Power) of $n = 807$ observations.

The variability of the entries in this dataset is large enough, providing a good dataset for testing our approach. Table 1 presents the descriptive statistics.

	count	mean	std	min	25%	50%	75%	max
Price	807	14269.2	23092.7	479	984	1699	18994.5	153000
Ratings	807	4.11264	0.36783	2.8	3.9	4.1	4.4	4.8
RAM	807	5.94176	2.05636	0	6	6	6	12
ROM	807	64.3903	53.8686	2	32	32	64	256
Mobile_Size	807	5.62066	3.95741	2	4.5	4.77	6.3	44
Primary_Cam	807	47.8216	11.1559	5	48	48	48	64
Selfie_Cam	807	8.86865	4.54725	0	8	8	12	23
Battery_Power	807	3278.86	937.629	1020	3000	3000	3800	6000

Table 1: Descriptive Statistics of the Cellphone dataset

If we restrict the process of predicting the answer to q to $\mathbf{O}_q$, Figures 3 and 4 show that the prediction is tighter, both using OLS and LASSO, according to the procedure of *Conformal Prediction* (Angelopoulos and Bates, 2023).

What can be observed in the figures is that while the median decreases with the rel-

evant and simulated values, the dispersion increases. Specifically, there are more outliers where the method has high uncertainty in its prediction. With the methods simulated using significant Shapley values, this dispersion is reduced, while maintaining a lower median of the conformal interval.

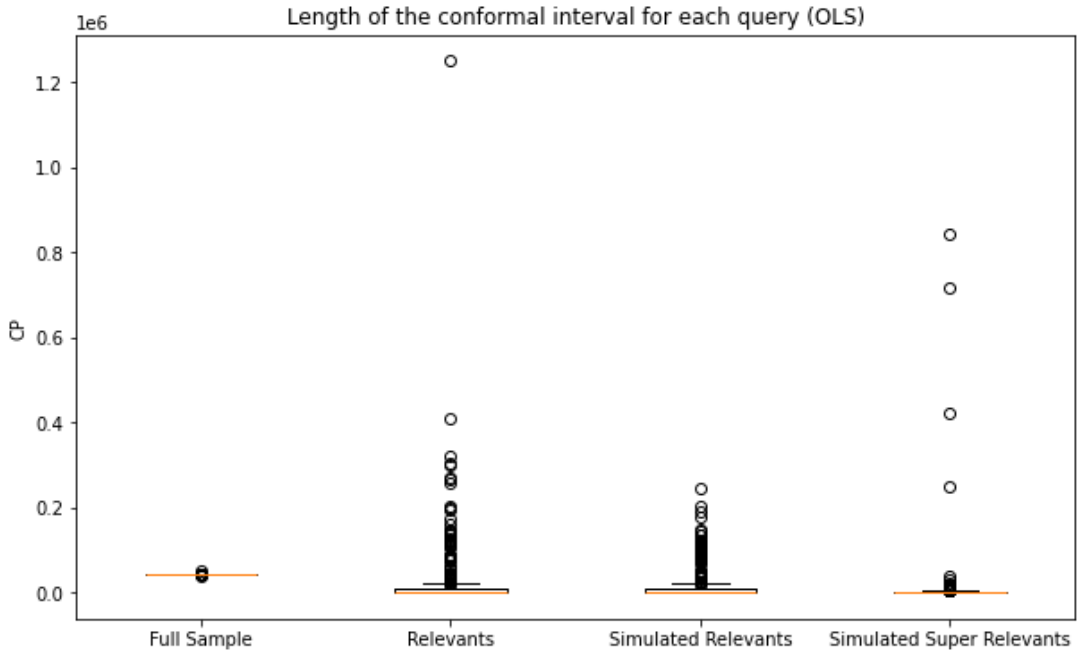

Figure 3: Predicting Cellphone prices in INR(indian Rupees) using Ratings, Ram, ROM, Mobile Size, Primary Cam, Selfi Cam and Battery Power ($n = 807$). **Source**: *Kaggle.com*.

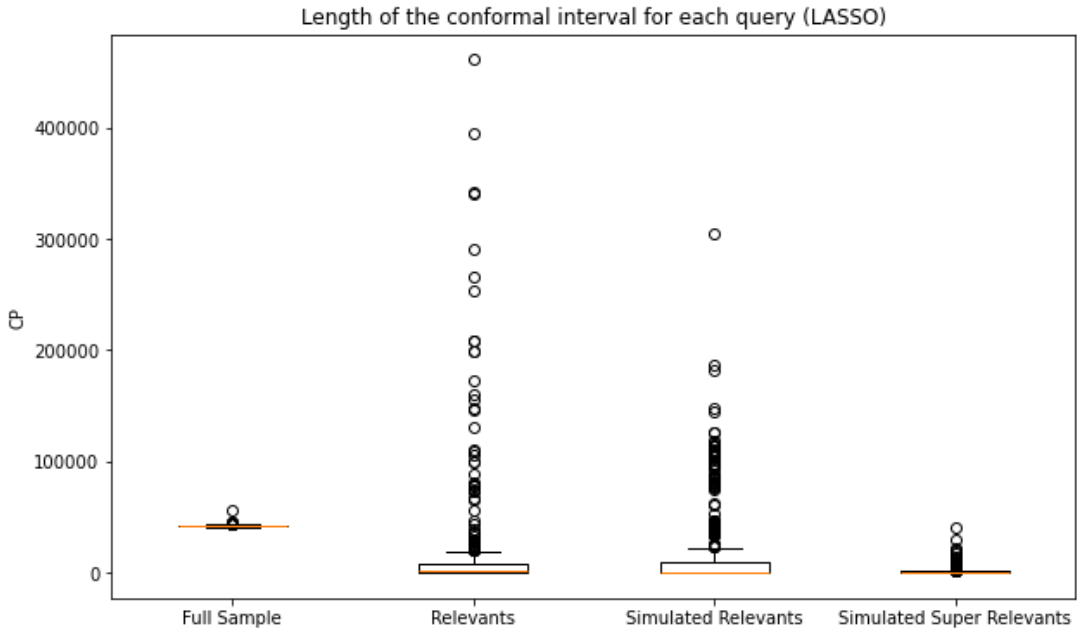

Figure 4: Predicting Cellphone prices in INR(indian Rupees) using Ratings, Ram, ROM, Mobile Size, Primary Cam, Selfi Cam and Battery Power ($n = 807$). **Source**: *Kaggle.com*.

6 Conclusions

This paper presents a novel algorithm for individualized inference using Shapley values to identify super-relevant observations for each query. By constructing probability distributions over these super-relevant data points and generating synthetic controls, the method improves the robustness and accuracy of individualized predictions. The approach not only enhances the quality of predictions but also facilitates a deeper understanding of the data-generating process by analyzing the answers to multiple queries.

The results presented here, obtained by focusing on a refined subset of relevant observations, yield more reliable predictions than those obtained using the entire dataset. The integration of Shapley values to select observations ensures a sound assessment of the contribution of individual data points. Furthermore, the use of methods of conformal prediction provides tools for quantifying the uncertainty of the answers to queries.

Future work may involve further theoretical analysis on the convergence properties and statistical guarantees of the method to enhance its applicability in broader contexts.

References

Angelopoulos, A. and Bates, S. (2023). Conformal prediction: a gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.

Cai, C., Chen, R., and Xie, M.-g. (2021). Individualized group learning. *Journal of the American Statistical Association*, pages 1–17.

Delbianco, F., Fioriti, A., and Tohmé, F. (2021). A methodology to answer to individual queries: finding relevant and robust controls. *Behaviormetrika*, 48(2):259–282.

Delbianco, F. and Tohmé, F. (2025a). Identifying highly relevant entries in datasets: A relevance-based classification. *Journal of Classification*, pages 1–21.

Delbianco, F. and Tohmé, F. (2025b). Individualized conformal prediction: using synthetic data as relevant controls. *Behaviormetrika*, pages 1–27.

Delbianco, F. and Tohmé, F. A. (2024). What is a relevant control?: An algorithmic proposal. In *Simposio Argentino de Inteligencia Artificial y Ciencias de Datos (ASAID 2024)-JAIIO 53 (Universidad Nacional del Sur, 12 al 16 de agosto de 2024)*.

Jin, Y., Ren, Z., and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120.

Liang, M., Choi, Y.-G., Ning, Y., Smith, M. A., and Zhao, Y.-Q. (2022). Estimation and inference on high-dimensional individualized treatment rule in observational data using split-and-pooled de-correlated score. *Journal of Machine Learning Research*, 23(262):1–65.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. (2022). The shapley value in machine learning. In *The 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence*, pages 5572–5579. International Joint Conferences on Artificial Intelligence Organization.

Zhou, X., Chen, B., Gui, Y., and Cheng, L. (2025). Conformal prediction: A data perspective. *ACM Computing Surveys*.