UDC 004.8:519.2

O. M. Bespala[1]

THE IDENTIFIABILITY PROBLEM IN MISSING DATA RECOVERY AND CAUSAL EFFECT ESTIMATION

[1] National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”

Анотація

У роботі досліджено вплив ідентифікованості причинної моделі на результати відновлення пропущених даних та оцінювання причинних ефектів. Проведено серію обчислювальних експериментів із використанням класичних методів імпутації та структурних причинних моделей. Розглянуто як ідентифіковані, так і неідентифіковані постановки задач у сенсі теорії причинного виведення . Отримані результати узгоджуються з теоретичними положеннями причинного аналізу та показують, що за відсутності ідентифікованості однакові спостережувані дані можуть відповідати різним повним моделям, що призводить до неоднозначності під час відновлення пропущених значень та інтерпретації причинних зв'язків.

Ключові слова: причинний аналіз, пропущені дані, ідентифікованість, MNAR, причинний ефект, латентний конфаундер, імпутація.

Abstract

This paper investigates the influence of causal model identifiability on missing data recovery and causal effect estimation. A series of computational experiments using classical imputation methods and structural causal models was conducted. Both identifiable and non-identifiable settings were analyzed within the framework of causal inference theory. The results obtained are consistent with established theoretical principles of causal analysis and illustrate that identical observable data may correspond to different complete-data models, leading to ambiguity in missing value reconstruction and causal interpretation [1, 3, 9].

Keywords: causal inference, missing data, identifiability, MNAR, causal effect, latent confounding, imputation.

Introduction

Missing data recovery is a fundamental task in statistical analysis, machine learning, and data science. A wide range of imputation methods are employed for this purpose, including mean substitution, regression imputation, and k-nearest neighbor algorithms [1, 2]. In most applied studies, the primary focus is on algorithm selection and recovery accuracy.

At the same time, within the framework of causal analysis, the property of identifiability plays a critical role. A model or causal parameter is considered identifiable if it can be uniquely determined from observational data given the accepted assumptions [3, 4]. In the absence of identifiability, multiple different models may be equally consistent with the observations yet lead to different conclusions regarding missing values or causal relationships [3, 5].

In particular, the missing data problem can be interpreted as a causal inference task with hidden variables, where the missingness mechanism (MAR/MNAR) determines the possibility of identification [1, 8, 10].

The aim of this work is to experimentally investigate the influence of identifiability on the feasibility of missing data recovery and causal effect estimation.

Research Results

A series of computational experiments was conducted with samples of $N = 2000$ observations and 30 independent replications of each experiment. Situations corresponding to both identifiable and non-identifiable causal models were considered within the framework of structural causal models (SCM) [3, 7].

Figure 1 presents a summary of the obtained results.

The first panel illustrates the MAR (Missing At Random) missingness mechanism within Rubin’s formalism [1]. For this scenario, regression imputation yielded a mean error of $MSE = 0.521 \pm 0.020$, and the k-NN method yielded $MSE = 0.634 \pm 0.026$. The low variance of estimates across independent runs indicates the existence of a practically unique solution under the given identifiability assumptions regarding the missingness mechanism [1, 10].

The second panel illustrates the MNAR (Missing Not At Random) case, which in general is non-identifiable without additional assumptions [1, 8]. Two different complete data distributions were constructed that generated statistically indistinguishable observed samples. The Kolmogorov–Smirnov test detected no

differences between the observable distributions ($p = 1.00$), yet the Jensen–Shannon divergence between the corresponding hidden parts of the distributions was $JSD = 0.388$. This is consistent with the fact that different complete-data models can be empirically equivalent [3, 9].

The third panel addresses the effect of a latent confounder in structural causal models [3, 7]. Two models with different true causal effects ($\alpha = 0.5$ and $\alpha = 0.0$) but the same joint distribution of observed variables were examined. Despite the difference in true effects, the linear regression coefficient estimates were statistically indistinguishable (0.745 ± 0.019 vs. 0.748 ± 0.016, $p = 0.536$), demonstrating the non-identifiability of causal effects in the presence of hidden variables [3, 5].

The right panel provides an example of the causal direction identification problem for linear Gaussian structural models. It is known that such models are often statistically equivalent in both causal directions [7]. The log-likelihood difference between the X→Y and Y→X models was $\Delta LL \approx 6 \cdot 10^{-14}$, which is practically zero and consistent with the theoretical model-equivalence result in the Gaussian case [7].

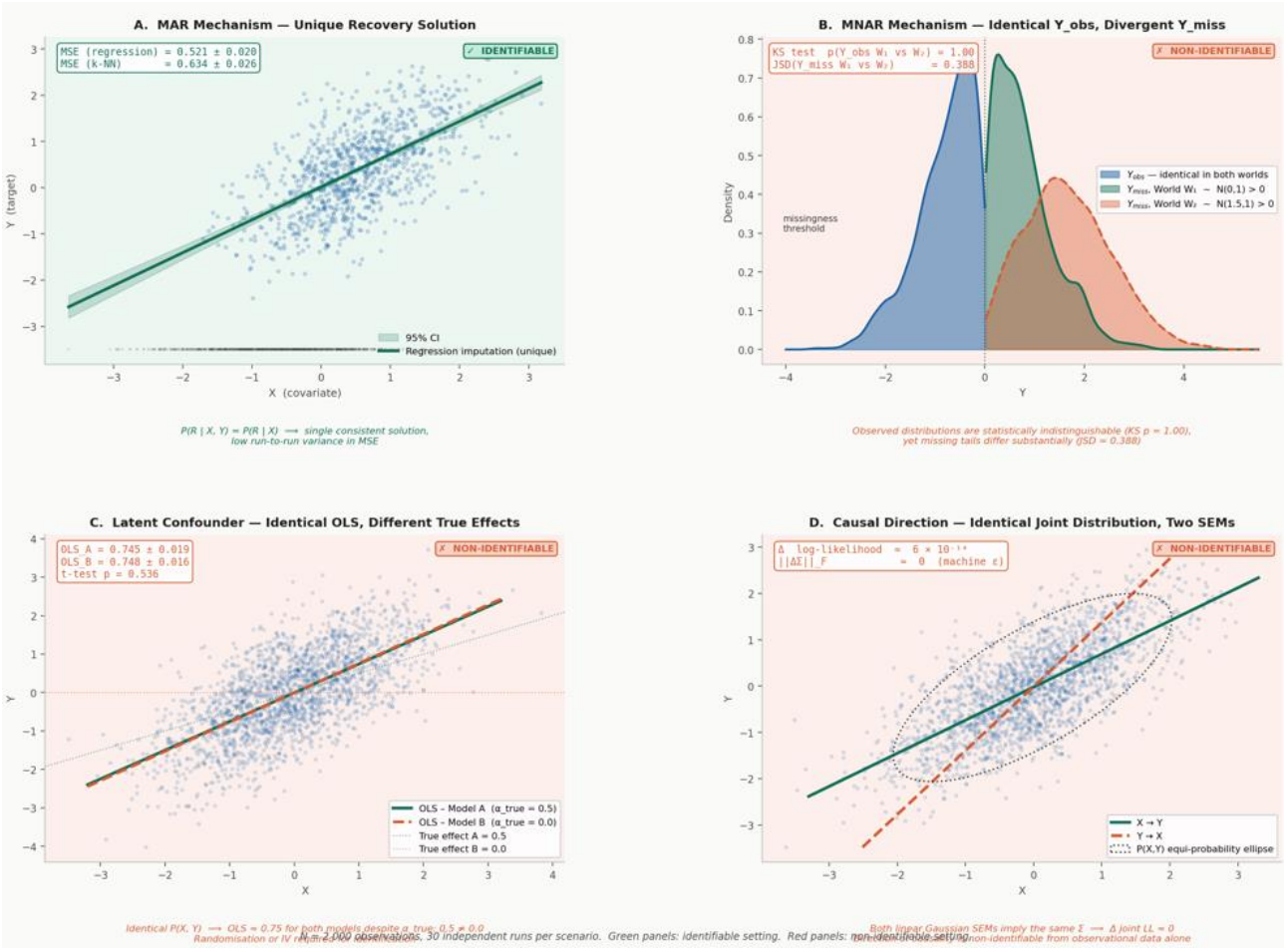

Figure 1 – Examples of identifiable and non-identifiable problems of data recovery and causal analysis

Taken together, these results demonstrate that the complexity of the recovery task is determined not only by the choice of algorithm, but also by the structural properties of the causal model and identifiability conditions [3, 4, 9].

Conclusions

The conducted study demonstrates the consistency of computational experiment results with the theoretical principles of causal analysis regarding the role of identifiability [3, 4].

For identifiable problems, stable missing data recovery results with low inter-run variability are observed [1, 2]. For non-identifiable problems, the existence of multiple models that are equally consistent with the observed data but correspond to different complete distributions or different causal interpretations is shown [3, 9].

The obtained results do not indicate the superiority or inferiority of specific imputation algorithms, but rather illustrate the fundamental role of identifiability conditions in missing data recovery and causal effect

estimation. Therefore, the analysis of model assumptions and verification of identifiability is a critically important step in building causal models [3, 4, 10].

REFERENCES

1. Rubin D. B. Inference and missing data // Biometrika. – 1976. – Vol. 63, no. 3. – P. 581–592.
2. Little R. J. A., Rubin D. B. Statistical analysis with missing data. – Hoboken : Wiley, 2019. – 450 p.
3. Pearl J. Causality: Models, reasoning, and inference. – 2nd ed. – Cambridge : Cambridge University Press, 2009. – 480 p.
4. Pearl J., Glymour M., Jewell N. P. Causal inference in statistics: A primer. – Hoboken : Wiley, 2016. – 150 p.
5. Hernán M. A., Robins J. M. Causal inference: What if. – Boca Raton : Chapman & Hall/CRC, 2020. – 624 p.
6. Shpitser I., Pearl J. Identification of conditional interventional distributions // Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI). – Arlington : AUAI Press, 2006. – P. 437–444.
7. Spirtes P., Glymour C., Scheines R. Causation, prediction, and search. – 2nd ed. – Cambridge (MA) : MIT Press, 2000. – 543 p.
8. Mohan K., Pearl J., Tian J. Graphical models for inference with missing data // Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI). – 2013. – P. 1–10.
9. Ding P., Li F. Causal inference: A missing data perspective // Statistical Science. – 2018. – Vol. 33, no. 2. – P. 214–237.
10. Tian J., Pearl J. A general identification condition for causal effects // Proceedings of the National Conference on Artificial Intelligence (AAAI). – 2002. – P. 567–573.

Беспала Ольга Миколаївна – аспірантка кафедри цифрових технологій в енергетиці Навчально-науковий інститут атомної та теплової енергетики, Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", м. Київ, e-mail: bespala.olha@lll.kpi.ua

Науковий керівник: ***Сліпченко Володимир Георгійович –*** д.т.н., професор, професор кафедри цифрових технологій в енергетиці Навчально-науковий інститут атомної та теплової енергетики, Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", м. Київ

Bespala Olha Mykolaivna – PhD student, Department of Digital Technologies in Energy, Educational and Scientific Institute of Nuclear and Thermal Energy, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, e-mail: bespala.olha@lll.kpi.ua

Scientific supervisor: Slipchenko Volodymyr Heorhiiovych – Doctor of Technical Sciences, Professor, Department of Digital Technologies in Energy, Educational and Scientific Institute of Nuclear and Thermal Energy, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv