研究計画書

(Research Plan)

記入日 (Date): 2025/04/26

(Name in English) : Dang Van Cong Hoang

学後の研究計画 (Research Plan)

Extending the Linear Mixed Causal Model to Handle Multiclass Discrete Variables

Introduction & Background

Causal discovery is the process of figuring out causal relationships from observational data. It is an important area in data science with applications in economics, biology, and social sciences. Traditional methods, like the Linear Non-Gaussian Acyclic Model (LiNGAM), assume that variables are continuous and have linear relationships. This limits their use when dealing with mixed data, which includes both continuous and discrete variables [5]. Mixed data is prevalent in real-world datasets, such as social surveys containing continuous variables (e.g., income) and discrete variables (e.g., occupation types).

The Linear Mixed causal model (LiM), proposed by Zeng et al. [1], addresses mixed data by providing identifiability conditions for continuous and binary discrete variables, using a two-step hybrid approach for causal structure discovery. However, LiM is currently limited to binary discrete variables, leaving a significant gap in handling multiclass discrete variables (e.g., nominal categorical variables with multiple categories). Extending LiM to include multiclass variables is important for wider use, especially in fields like social sciences where multiclass responses are common.

This research aims to modify the LiM model to work with multiclass discrete variables while keeping its identifiability. This will allow for accurate causal discovery in complex mixed datasets. Conducted as part of a Master's by Research program, this study will help address a key limitation of LiM and improve its practical usefulness.

Research Questions

1. How can the LiM model be adapted to include multiclass discrete variables using multinomial logistic regression while ensuring identifiability?

2. What are the identifiability conditions for the extended LiM model in bivariate and potentially multivariate cases?

3. How does the extended LiM model perform compared to existing methods (e.g., PC, standard LiNGAM, L-LiNGAM) in terms of accuracy and robustness on synthetic and real-world mixed datasets?

Hypotheses

1. Hypothesis 1: The extended Linear Mixed Causal Model (LiM) achieves identifiability for multiclass discrete variables in bivariate cases, consisting of one continuous and one multiclass variable (3-5 categories), under assumptions of non-Gaussian noise and causal sufficiency.

2. Hypothesis 2: The extended LiM model outperforms existing methods (PC, standard LiNGAM, L-LiNGAM) in accuracy, measured by FDR, F1 score, and Structural Hamming Distance on synthetic and real-world mixed datasets.

Notes:

Hypothesis 1 builds on the LiM framework (Zeng et al., 2022), assuming non-Gaussian noise (e.g., Laplace, Exponential) and causal sufficiency to ensure identifiability, consistent with the proof-by-contradiction approach leveraging asymmetric properties of multinomial logistic regression (see Methodology, Theoretical Development). The focus on bivariate cases ensures feasibility within the one-year Master's timeline, with identifiability tested through theoretical proofs and supported by preliminary simulations.

Hypothesis 2 evaluates performance across diverse experimental conditions, including sample sizes (500, 1000, 5000), non-Gaussian noise types (Laplace, Exponential), and multiclass imbalance (e.g., class ratio 1:2:3), to assess accuracy and robustness. Preliminary simulations suggest the extended LiM model achieves higher F1 scores (e.g., 0.94 vs. PC's 0.79), supporting the claim of improved accuracy.

Literature Review

Causal discovery for mixed data has been an active research area, with several methods addressing the challenge of combining continuous and discrete variables:

- **LiNGAM and Its Limitations:** Shimizu et al. [5] introduced LiNGAM, which assumes linear relationships and non-Gaussian noise for continuous variables, enabling unique causal structure identification. However, LiNGAM struggles with discrete variables, as their non-linear relationships (e.g., sigmoid curves) violate the linearity assumption.

- **Latent LiNGAM (L-LiNGAM):** Yamayoshi et al. [2] proposed L-LiNGAM, which assumes discrete variables arise from continuous latent variables via link functions, handling binary and limited multiclass data. While effective in some cases, L-LiNGAM lacks formal identifiability guarantees for general multiclass variables and relies on link function assumptions.

- **Linear Mixed Causal Model (LiM):** Zeng et al. [1] developed LiM, which supports continuous and binary discrete variables with identifiability guarantees under causal sufficiency. Experiments show LiM outperforms PC, logistic regression, and standard LiNGAM in binary mixed data settings. However, its limitation to binary variables restricts its applicability to multiclass scenarios.

- **Constraint-Based and Score-Based methods:** Methods like the PC algorithm use conditional independence tests (e.g., likelihood-ratio tests) to handle mixed data [4]. While these methods are flexible, they often return Markov equivalence classes rather than unique causal directions and are sensitive to sample size. Score-based methods like Reproducing Kernel Hilbert Space (RKHS) [3] allow for mixed data types and multidimensional variables but require high computational costs, making them less suitable for multiclass mixed data compared to LiM's framework.

Research Gap: The literature shows a gap in extending LiM to multiclass discrete variables. While L-LiNGAM is promising, its dependence on latent variables and the lack of identifiability for multiclass cases limit its effectiveness. This proposal aims to fill this gap by extending LiM which allows for robust causal discovery in mixed datasets. This is particularly important in fields like social sciences, where multiclass variables (such as education levels and occupations) are common, supporting applications in policy analysis and economic research.

Methodology

The research will follow a structured approach to extend the LiM model for multiclass discrete variables, involving theoretical development, algorithm implementation, and experimental evaluation.

Theoretical Development

Redefine LiM's structural equations to incorporate multiclass discrete variables using multinomial logistic regression:

$$P(x_i = m \mid \mathrm{pa}(i)) = \frac{\exp(\beta_m^T \mathrm{pa}(i) + \gamma_m)}{\sum_{l=1}^{k} \exp(\beta_l^T \mathrm{pa}(i) + \gamma_l)},$$

, where x_i is a multiclass variable with k categories, and $pa(i)$ are its parents.

Prove identifiability conditions, primarily for bivariate cases (one continuous, one multiclass variable), using a proof-by-contradiction approach that leverages the asymmetric properties of non-Gaussian noise and the multinomial logistic structure, building on the framework of Zeng et al. (2022) [1]. The proof will use distributional differences to identify causal directions, similar to how LiM works with binary variables. If challenges come up, I will explore alternative methods.

If time allows, I will extend the proof to multivariate cases, making sure it aligns with LiM's assumptions (such as no unobserved confounders and acyclicity).

Focus on bivariate cases first to ensure the work is feasible within the Master's timeline, treating multivariate extensions as a secondary goal.

Algorithm Implementation

Extend LiM's two-step hybrid algorithm (which includes score optimization followed by constraint refinement) to work with multiclass variables.

Implement this in Python using libraries like lingam for causal discovery and statsmodels for multinomial logistic regression.

Experimental Evaluation

Generate synthetic datasets with known causal structures, including continuous and multiclass discrete variables (3-5 categories), using Python.

Compare the extended LiM model to PC, standard LiNGAM, and L-LiNGAM using FDR, F1 score, and Structural Hamming Distance.

Test on a real-world dataset (e.g., UCI Adult dataset with multiclass variables like occupation) to assess practical applicability.

Analyze robustness to varying sample sizes (e.g., 500, 1000, 5000) and noise levels.

Data Source

Synthetic Data: Generated using Python to create mixed datasets with known causal structures. Variables will include continuous (e.g., non-Gaussian noise) and multiclass discrete (3-5 categories, modeled via multinomial distributions). Ground truth causal graphs will enable precise evaluation of accuracy.

Real-World Data: The UCI Adult dataset (UCI Machine Learning Repository), which includes continuous variables (e.g., income, hours worked) and multiclass discrete variables (e.g., occupation), will be used for real-world validation. Additional survey datasets (e.g., from Kaggle) with multiclass responses may be explored if time permits.

Limitations and Risks

Theoretical Complexity: Proving identifiability for multiclass variables may be challenging, especially for multivariate cases. **Mitigation**: Focus on bivariate cases initially, using LiM's proofs as a template. Consult with advisors for complex proofs.

Computational Constraints: Multinomial logistic regression for multiclass variables increases computational demands, potentially slowing experiments. **Mitigation**: Optimize code using efficient libraries and limit dataset sizes (e.g., 5-10 variables). Use cloud computing resources if needed.

Data Availability: Real-world datasets with suitable multiclass variables may be limited or require preprocessing. **Mitigation**: Rely primarily on synthetic data for controlled experiments and use well-documented datasets like UCI Adult for validation.

Scope Creep: Extending to multivariate cases or integrating constraint-based methods (e.g., PC) could exceed the one-year timeline. Mitigation: Strictly scope the project to bivariate cases and treat additional extensions as secondary objectives.

Timeline

Months 1-2: Literature review (LiM, multinomial regression, causal discovery for mixed data).

Months 3-7: Theoretical development (structural equations, identifiability proofs).

Months 8-9: Algorithm implementation and optimization.

Months 10-11: Experimental evaluation (synthetic and real-world data).

Month 12: Write-up and submission of thesis, targeting a conference paper.

Expected Outcomes

A generalized LiM model that supports multiclass discrete variables, with proven identifiability for bivariate cases.

An open-source Python implementation of the extended LiM algorithm.

Experimental results showing improved accuracy compared to PC, standard LiNGAM, and L-LiNGAM on mixed datasets.

A Master's thesis and a potential conference paper that contribute to causal discovery for mixed data.

References

[1] Y. Zeng, S. Shimizu, H. Matsui, and F. Sun, "Causal discovery for linear mixed data," in Proceedings of the First Conference on Causal Learning and Reasoning, B. Sch¨olkopf, C. Uhler, and K. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 177, PMLR, Nov. 2022, pp. 994–1009.

[2] M. Yamayoshi, J. Tsuchida, and H. Yadohisa, "An estimation of causal structure based on latent lingam for mixed data," Behaviormetrika, vol. 47, pp. 105–121, 2020.

[3] B. Huang, K. Zhang, Y. Lin, B. Sch¨olkopf, and C. Glymour, "Generalized score functions for causal discovery," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ser. KDD '18, London, United Kingdom: Association for Computing Machinery, 2018, pp. 1551–1560, isbn: 9781450355520.

[4] M. Tsagris, G. Borboudakis, V. Lagani, and I. Tsamardinos, "Constraintbased causal discovery with mixed data," Int J Data Sci Anal, vol. 6, no. 1, pp. 19–30, 2018, Epub 2018 Feb 2.

[5] S. Shimizu, P. O. Hoyer, A. Hyv228;rinen, and A. Kerminen, “A linear non-gaussian acyclic model for causal discovery,” Journal of Machine Learning Research, vol. 7, no. 72, pp. 2003–2030, 2006.